Essays in Sports Economics

Stefanie Pohlkamp

Dissertation Thesis

Submitted for the Degree of Doctor Rerum Politicarum

Department of Economics, University of Hamburg

Submitted by
Stefanie Pohlkamp
born in Siegen, Germany

October 29th, 2014

Thesis Committee

Chairman: Prof. Dr. Dr. Lydia Mechtenberg

First Examiner: Prof. Dr. Gerd Mühlheußer

Second Examiner: Prof. Thomas Siedler (PhD)

The disputation was held on April 1, 2015

List of individual work

I. The impact of referees on match outcomes in professional sports: Evidence from the German Football Bundesliga

II. Are football referees really neutral or do they have prejudices?

Acknowledgments

First and foremost, I wish reveal my deepest gratitude to my supervisor Professor Dr. Gerd Mühlheußer for his superb guidance and support to my dissertation writing.

I would also like to thank Professor Dr. Thomas Siedler, Dr. Berno Büchel, Nina Claus, Sandra Hentschel, Kathrin Thiemann and Niklas Wallmeier for their helpful comments and assistance.

Lastly, my special thanks go to my parents, Gitta and Adalbert Pohlkamp for their unlimited encouragement and to my friends for their solace.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

This dissertation presents a collection of research papers in Sports Economics. The research area Sports Economics is a relatively young discipline. Its origin lies in the U.S. sports market and the revolutionary work by Rottenberg (1956) on "The Baseball Players' Labor Market". Here, he points out characteristics from baseball and the special dealing with labor contracts in baseball such as dealing with free agents, the reserve rule and the draft or selection rule.[1] These peculiarities, particularly in contract design, lead to follow-up research which discusses further special properties of the sports market. Neale (1964) emphasized special features like "The Inverted Joint Product", referring to the fact that the product *sport* can only be produced if there are at least two competitors (e.g. "Louis-Schmeling Paradox"). Recent work includes, for example, research by Kahn (2000), who uses "The Sports Business as a Labor Market Laboratory" for labor market research. He underlines the sport sector's advantages of being the only market where we often have publicly available data on name, life history, performance statistics and wages.

Yet high data quality is not the only reason that makes Sports Economics an interesting research field. The sports sector furthermore is of great economic relevance. The accountancy firm Deloitte collected information on revenues from the most famous sports leagues and prepared a ranking in 2012. According to this list, the National Football League (NFL) has total sales of 6.7 billions euros per season, making it the most successful league, followed by Major League Baseball (MLB) with 5.0 billions euros and the National Basketball Association (NBA) with 3.1 billions euros. The German Football Bundesliga ranks sixth with total sales of about 1.7 billions euros per season.[2]

Further evidence of the economic relevance of sports is provided in a study by the Germany Federal Ministry for Economic Affairs and Energy from 2012. They find that active sport consumption amounts to 80 billions euros per year, one in three German enterprises gives financial support to sports and last but not least the advertising expenditure of manufacturers of sports equipments amounts to one billion euros.[3]

[1] The *reserve rule* or *reserve clause* is a very popular rule in professional U.S. sports. Almost every major sports league (MLB, NHL, NBA) introduced this rule. The *reserve clause* holds that no player with a valid contract from a club may negotiate with other professional clubs. Thus, all player rights are assigned to the club. In the meantime, the reserve clause has been abolished in most U.S. sports and has since been replaced by free agency. *Free agents* are players in professional U.S. sports without a valid contract. This implies that *free agents* can change to another professional club of their own choice.

[2] Cf. `www.handelsblatt.com/sport/fussball/nachrichten/internationale-rangliste-bundesliga-erringt-top-platzierung-beim-umsatz/6113310.html`, last access: June 25, 2014.

[3] Cf. `www.bmwi.de/Dateien/BMWi/PDF/Monatsbericht/Auszuege/02-2012-I-4`, last access: June 25, 2014.

While there are numerous broad research areas (e.g. contest design, competitive balance in sports leagues, cartels) in Sports Economics, I only focus on three subject areas namely referees and player performance in football as well as doping in sports.[4]

The first paper "The impact of referees on match outcomes in professional sports: Evidence from the German Football Bundesliga" is an empirical work that has been presented at the Nachwuchsworkshop of the Arbeitskreis Sportökonomie in Magglingen in May 2012, at the PhD Seminar in Economics in Hamburg in May 2012, at the 5th ESEA conference in Esbjerg in September 2013 and at a poster session at the 3rd Potsdam PhD Workshop in Empirical Economics in March 2014. Here, I use an empirical approach from Bertrand and Schoar (2003) to estimate fixed effects from referees on match outcome or referee decisions. I am interested in whether referees have individual effects on the final result of a football match even though they are assumed to behave as impartial observers. Afterwards, I use specific referee characteristics (age, experience, regional association, profession) to explain these referee fixed effects to find reasons for deviant referee behavior. This phenomenon is called *referee bias* in the Sports Economics literature and implies that referees behave differently from what is to be expected even though referees, particularly in Germany, receives extensive training and monitoring as well as financial compensation. Within this paper, I not only present new empirical evidence of referee bias but I also survey the existing research on referee bias and present an overview of football referees in Germany. In fact, although it is expected that referees should have no effects on match outcome and even referee decisions, the analyzes of my data shows that football referees do have significant individual effects on match outcome for home teams, though not for away teams. Further, I find significant referee effects for most referee decisions like awarded yellow and red cards as well as awarded goals and penalties. Again, these results often only hold for home teams. Analyzing the estimated referee fixed effects again reveals differences among referees especially between home and away teams. Using the referee fixed effect for match result as an example, I also find evidence that these referee fixed effects have different consequences for different teams in the German Bundesliga. This implies that successful teams like Bayern München or Borussia Dortmund always benefit from a significant referee effect while other teams (e.g. SC Freiburg, 1. FC Kaiserslautern) are systematically disadvantaged. These estimated referee fixed effects are also used to assess whether referees exhibit so-called "referee styles". Thus, I find significant evidence that referees with a significant effect on match outcome also affect further referee decisions like awarded yellow cards or goals. Despite finding these significant referee effects, referee characteristics like age, profession or regional association are not adequate to find a satisfactory explanation for my earlier results. This leads me to my second paper where special information on teams and football matches is used to find evidence and reasons for referee bias, too.

The second paper "Are football referees really neutral or do they have prejudices?" is a theoretical as well as an empirical paper on referee bias in football. While recent research (cf. Dohmen, 2003; Rickman and Witt, 2008) often use principal-agent-theory to find theoretical

[4] For example, in German professional football the media rights are centrally marketed by the "Deutsche Fußball Liga" (DFL), suggesting a potential offense against the ban on cartels.

explanations for referee bias, I design a game-theoretic model to predict referee behavior. Afterwards these predictions are empirically tested by a non-parametric test (Pearson χ^2 test).[5] While the first paper presents explanations for referee bias from a referee's view, this second paper discerns differences in referee behavior with different teams or special match circumstances. To that purpose, I build different subgroups like home and away teams, rich and poor teams, matches in stadiums with and without a racetrack around the pitch as well as local and non-local derbies to find evidence for referee bias in football. The equilibrium of my model predicts that the fraction of wrong referee decisions among all referee decisions ("fail rate") must be equal across all subgroups if a referee is unbiased (non-prejudiced). Further, the descriptive analyzes of the data show that differences across these subgroups exist for all referee decisions. Therefore, I use the Pearson χ^2 test to verify if these differences depend on special match characteristics (e.g. number of spectators, league positions) or on the referees' subjective prejudices. I find significant referee bias with respect to referee decisions on awarded goals as well as disputable awarded and not awarded goals. Further evidence of referee bias is found with respect to not awarded penalties. As expected, a significant referee effect is most often detected among the subgroup for home and away teams.

The third paper "The impact of Intermediate Information on Effort Provision in Tournaments with Heterogenous Contestants" is joint work with Prof. Dr. Christian Deutscher and Sandra Hentschel from Bielefeld University. This paper has been presented at the 89th Annual Conference of the Western Economic Association International in Denver, Colorado in June 2014 and at the PhD Seminar in Sports Economics in Paderborn in August 2014. Here, we are interested in individual effort and whether it is affected by intermediate information. Although much work has been done concerning tournaments (e.g. Lazear and Rosen, 1981; Frick et al., 2008; Bach et al., 2009; Backes-Gellner and Pull, 2013), the focus often lies on designing these contests or how heterogeneous contestants would behave in tournaments. The effect of intermediate information is not often examined yet. Part of the reason may be the difficulty of obtaining reliable data. That is why only few studies exist that use experimental data (e.g. Gürtler and Harbring, 2010; Ludwig and Lünser, 2012). We test the effect from intermediate information on effort with data from the German Football "Bundesliga" that include the three seasons 2010/11 to 2012/13. Since player's effort is not directly observable, we use data on running distance and number of runs and sprints. But these variables are only available on match-level and we are not able to investigate the effect of intermediate information (score) for the starting players. Hence we are only interested in substitutes because here we have information on the score at the time when they are substituted. Information on the score is important for our analyzes because it represents our proxy for intermediate information. More precisely, we test whether the goal difference (score) has significant effects on player's efforts in a match. Among other control variables (e.g. team and player fixed effects, number of sending offs), we test for pre-match heterogeneity using betting odds. We find that effort is highest when the team is leading by one goal. Further, we find that effort is higher in matches with a tied score compared to when the team is behind by one goal. This result is in line with prospect theory, which predicts that

[5] The origin of this model and empirical test lies in research on racial-discrimination of police officers (Knowles et al., 2001).

individuals value potential losses higher than gains. Lastly, if a match is decided intermediate information decreases individual effort.

The fourth paper "Nobody's Innocent - The Role of Customers in the Doping Dilemma" is joint work with Prof. Dr. Eike Emrich from Saarland University and Dr. Berno Büchel from the University of Hamburg. This paper has been accepted for publication in the Journal of Sports Economics. This study follows a theoretical approach to find a solution to the doping problem in sports. Here, we build on existing work that uses game theory, notably inspection games, to solve the doping problem. We introduce a new stage to the existing inspection game and consider the "power" of customers to prevent doping in sports. One would assume that reacting on (doping) scandals with a withdrawal of support would stop fraudulent behavior. Instead our model shows that the effect goes into the opposite direction and a withdrawal induces cheating. In our equilibrium, we find that doping is prevalent because customers undervalue the incentives of organizations (e.g. World Anti Doping Agency) to detect doped athletes. In a next step, we are interested in inducing a doping-free behavior and our model predicts that transparency would help to avoid doping. Consequently one of our recommendations is that customers should be informed about the aggregate number of doping tests and not only about convicted dopers.

This dissertation contributes to different interesting research fields in Sports Economics. First, I provide two new empirical methods to assess whether referees in football make biased decisions even though they are expected to always decide neutrally. Further, I develop a new theoretical approach to modelling the referees' decision making while the existing research uses principal-agent-theory to find explanations for referee favoritism. The findings from my third paper also have implications for kinds of contests, not only in sports but also for example for firms that hold tournaments among employees. The fourth paper not only shows how customers influence doping behavior, our results also hold for every type of detected fraudulent activity within organizations. Examples in the economic sector include child labor in textiles or food scandals.[6]

[6] Famous scandals are for example horsemeat found in oven-ready lasagne in 2013, dioxin found in eggs in 2010 or detected child labor from a supplier of the sports article manufacturer Nike in 2011.

Chapter 2

The impact of referees on match outcomes in professional sports: Evidence from the German Football Bundesliga

Abstract

Using data from German 1^{st} Bundesliga seasons 1993/94 through 2007/08, we examine whether referees in football have a significant individual influence on match results or other outcomes such as goals, penalties, yellow and red cards. We show that not only does a significant individual referee effect exist, but referees moreover have different "referee styles" and there are differences in their behavior for home and away teams. Further, we use referee characteristics like age, experience and a referee's profession to explain these individual effects.

Keywords fixed effects, referee bias, decision making, football
JEL Classification D70, J00, L83, M50

2.1 Introduction

This paper investigates whether there is a significant individual influence of football referees on match outcomes in the German 1^{st} Bundesliga and it contributes research on referee bias in football. Though, we would expect that referees decide neutrally and impartially, we find evidence that the opposite is true. 1^{st} Bundesliga referees frequently find themselves at the center of press coverage after a matchday. Articles in sports magazines are full of discussions on erroneously awarded or not awarded goals, decisions on offside as well as eligible or ineligible penalties. Even coaches or managers of professional football clubs often mention that their teams are systematically discriminated against by referees. In the season 2011/12, the chairman Karl-Heinz Rummenigge and the president Uli Hoeneß of FC Bayern München maintained that referees always decide against their team.[1] A similar assertion was made by Fredi Bobic in the season 2010/11. Here, the head of sports of VFB Stuttgart suspected systematic disadvantages against his team. He even argued that his team lost six points due to wrong referee decisions.[2] A further example of football teams believing that referees are biased against them is a letter of protest from Hertha BSC Berlin in the season 2009/10 in which the management claim that the club is disadvantaged by referees and even referee appointments.[3] One highlight of the referee discussion is the decision of the German Football Association (DFB) not to appoint referee Wolfgang Stark for further matches by Borussia Dortmund in the 2012/13 season. The reason for this decision was the suspicion that this referee is biased, based on his numerous dubious decisions to the detriment of Borussia Dortmund.[4]

In light of this public discussion, the aim of this paper is to analyze whether these statements are only evidence of referees, who are after all merely human and make mistakes, having a "bad day" or whether referees and their decisions systematically influence the result of a football match. Thus, we ask whether referees deliberately treat home and away teams in different ways, even though some would assume that German referees have no incentives to make biased decisions. Before a referee is nominated for 1. and 2. Bundesliga matches he must have gained practical experiences in lower divisions matches and has been promoted by the German Football Association on the basis of good monitoring results as a referee. Good reports are also necessary to remain in service or even to be promoted to FIFA referee. In total, 10 German referees and 10 referee-assistants are listed on the FIFA list.[5] Germany has the most referees on the

[1] Cf. www.spiegel.de/sport/fussball/bayern-boss-rummenigge-aetzt-gegen-schiedsrichter-a-815498.html and www.spiegel.de/sport/fussball/bayern-praesident-hoeness-schiedsrichter-im-zweifelsfall-gegen-uns-a-815673.html, last access: May 27th, 2013.

[2] Cf. www.sport1.de/de/fussball/fussball_bundesliga/artikel_315916.html, last access: May 27th, 2013.

[3] Cf. www.stern.de/sport/fussball/fussball-bundesliga-hertha-protestiert-gegen-schiedsrichter-1557820.html, last access: May 27th, 2013.

[4] Cf. www.welt.de/sport/fussball/bundesliga/borussia-dortmund/article111939939/Stark-pfeift-diese-Saison-keine-BVB-Spiele-mehr.html, last access: May 27th, 2013.

[5] Cf. de.fifa.com/aboutfifa/footballdevelopment/technicalsupport/refereeing/men.html, last access: March 4th, 2013.

international FIFA list, which is further evidence of the quality of German referees.[6]

Another reason why German referees have no incentives to be partial is financial compensation, even though referees do not work as professionals in Germany.[7] To avoid grave errors in judgment as well as discussions on referee performance, it is always attempted to improve the decision making of referees. Recent examples include the goal-line referees in Champions and Europa League matches and the introduction of the goal-line technology during the 2014 FIFA World Championships.

Previous studies, as discussed in detail below, find evidence of a home advantage and later this advantage is explained with reference to biased decisions from officials in favor of home teams.[8] The extra time at the end of the second half and referee decisions like goals, penalties or bookings are used to indicate whether referees systematically favor home teams. One explanation for this bias is often the social pressure which is caused by the crowd in the stadium. Another strand of literature follows the argument that referees work in a principal-agent relationship as the agents of the football governing body as their principal. Thus, it is assumed that referees act impartially to satisfy their principal but simultaneously they want to pacify the crowd in the stadium (e.g. Sutter and Kocher, 2004; Dohmen, 2003).

In this paper, we examine whether referees have individual effects on match outcomes (result, goal difference), possible match-winning decisions like goals and penalties and also non match-winning decisions like bookings. To that purpose we generate a data set covering seasons 1993/94 through 2007/08 and containing 4,590 matches and 70 referees. Following the innovative approach by Bertrand and Schoar (2003), a new method to estimate referee influence is applied within this study. First, dependent variables like result, goal difference, goals, penalties and bookings are estimated by regression analyzes.[9] The right hand side of this regression equation is independent of referees and only describes the two teams of a match (e.g. performance, budget). Afterwards this regression equation is enhanced by referee fixed effects. A significant individual influence is found if the explanatory power of this regression model increases after we include referee fixed effects, and if the F-test for joint significance of these referee dummy variables is significant. Following this, the estimated referee fixed effects from this first stage are used to describe different "referee styles" as the correlation between the significant fixed effects is examined. In a second stage regression, we use observable referee properties (e.g. age, experience, FIFA status, professional job) instead of the referee fixed effects to

[6] The FIFA annually decide about the maximum number of referees from each association. Among other factors, the following three points are taken into account: First, the level of refereeing of each association. Second, the level of competitions of each association and third, the professional level of each association's competitions (cf. `www.fifa.com/mm/document/affederation/administration/01/98/73/21/circularno.1334-2013listsofinternationalreferees-assistantreferees-futsalrefereesandbeachsoccerreferees.pdf`, last access: May 28th, 2013).

[7] Further information on training, monitoring and rewards is given in section 2.3.

[8] A detailed overview on home advantage was done by Courneya and Carron (1992). A so called referee bias is found in different major European football leagues like in England, Spain and Germany.

[9] As in the point classification in football leagues, we define the result as equal to three points if a team win the match and zero if they lose. A draw yields one point for each team.

explain which referee characteristics are responsible for the significant individual referee effects.[10]

One central conclusion is that referees have a significant individual influence on match outcome. Moreover, the significant individual effects vary between home and away teams. It follows that referees treat home and away teams in different ways. Especially for our most interesting variable match result, we only find significant referee effects for home teams. Furthermore, we estimate "referee styles" and find that referees with significant individual effects on the match result also have positive and significant effects on other referee fixed effects like awarded yellow and red cards, and goals in home matches. In away matches, referees with a significant effect on goal difference have different and specifically smaller impacts on the other referee effects.

Last but not least, we use observable referee properties to explain the significant individual effects. Here, we suppose that referees may follow career concerns and that they are interested in a good reputation at the beginning of their career. That would be in line with the behavior of judges (e.g. Levy, 2005) or managers (e.g. Frank and Goyal, 2007) where empirical evidence is found that a good reputation improves performance. In our case, reputation is described as the opportunity to be a FIFA referee and the chance to referee matches in international competitions (e.g. Champions League). Referring to the career concerns model from Holmström (1999) we assume that a referees' experience initially has a decreasing effect on the individual effects but this only holds up to a certain point of inflection. Furthermore, we aim to explain these significant individual effects with reference to the referees' professional jobs, their regional football association and physical height as well as variables pertaining to their prior performance.

However, if we use these referee characteristics instead of the referee fixed effects, we do not find strong empirical support for our assumptions regarding the individual effects. In particular, we find significant effects neither for the FIFA dummy-variable nor for home and away teams. Moreover, there is only limited empirical evidence that referees follow career concerns if we only control for away teams. In fact, the coefficients for experience and age are always insignificant if we control for home teams. Concerning the referees' profession, we find more empirical power for away teams than for home teams but both teams have in common that we find significant effects if referees work as an engineer. Altogether we have to note that observable referee characteristics are not adequate to explain the significant individual referee effects on match outcome and further referee decisions.

The remainder of this paper is organized as follows: Section 2.2 provides an overview of the literature concerning referee bias. Section 2.3 describes the referee system in the German Bundesliga. In section 2.4 the data set is described and summary statistics are presented. Section 2.5 presents the empirical results and section 2.6 concludes.

[10] The effects of individual characteristics are also used to explain the impact of CEO decisions on firm performance (cf. Malmendier and Tate, 2005, 2008).

2.2 Literature on Referee Bias in Football

Numerous studies aim to explain the behavior of football referees and how they favor home teams (e.g. Nevill et al., 2002; Dohmen, 2003; Sutter and Kocher, 2004; Garicano et al., 2005). One main result about favoritism is that referees are influenced by spectators. Moreover, increasing years of experience reduces favoritism but referees also follow career concerns which result in an increasing referee bias again as referees grow older. The extra time at the end of the second half is often used as an indicator of referee bias. The number of "awarded goals" and "penalties" or "not awarded penalties" and bookings are also used to test for favoritism. But how a referee affects the result of a match is not yet often examined. Most studies on referee bias rely on public data on football matches were analyzed. One notable exception is Nevill et al. (2002). They follow the hypothesis of a relation between the experience of a referee and referee bias. To that end, the authors asked a group of 40 referees - some of whom had only just begun their careers while others had up to 43 years of experience - to judge 47 incidents in a game between Liverpool and Leicester City which they watched on video tape. One group of these referees watched the video with crowd noise and the other group without. The authors' empirical findings from the experiment show a significant influence of the crowd noise as well as a significant non-linear relationship between a referee's experience and the number of fouls caused by the home team which she recognized.

Dohmen (2003) studies the behavior of referees in German professional soccer and assumes a principal agent relationship between the governing body and the referees.[11] He finds that a special social atmosphere, as in a football arena, can persuade agents to make decisions that are neither in the interest of their principal nor in the agent's own interest. He uses information about injury time at the end of a half, awarded and not awarded penalty kicks, the power of teams and other decisive data like yellow or red cards.[12] Dohmen (2003) finds that the length of injury time in the second half is much longer the closer the match. He supposes that fans influence a referee's decision and that there are differences between a stadium with and without a racetrack. Thus, a second result of this study is that a high attendance - to - capacity ratio reduces the home bias in stadiums without a racetrack around the field and the bias increases in stadiums with a track. In a further study, Dohmen (2008) confirms his previous results, while he also uses data on wrong and right referee decisions on goals and penalties, as well as yellow and red cards.

Another study on referee bias shows the existence of a "home team bias" in referee decisions in Italian (Serie A) and United States professional football leagues. Lucey and Power (2004) analyze whether referee decisions are indeed systematically biased, and second, they investigate whether the amount of extra time granted is influenced by the social environment that is created by the crowd in the stadium.[13] Moreover, the authors examine two different reward systems. On the one hand, the incentives of Italian referees are similar to those of their German

[11] In Germany the governing body is the DFB.
[12] Injury time refers to extra time at the end of a half that the referee grants to balance a time loss that is caused by injuries during the preceding half.
[13] The decision which is examined is the amount of injury time after the end of the second period in a game.

colleagues.[14] By contrast, in the US system, the referees' rewards per game are based on their past performance. Finally, the authors identify two factors which influence referee decisions. First, referees are interested in being impartial because their aim is to be reappointed and/or promoted, and secondly, referees endeavor to satisfy the crowd in the stadium.

Boyko et al. (2007) also examine, amongst other research objectives, if referees are biased and whether this bias has an influence on the outcome of a match. Second, they investigate whether the influence from the crowd really drives the behavior of referees or whether it are the players who are influenced by crowd noise. They conclude that favoritism of referees is an individual characteristic and that more experienced referees cause less home advantage.

Buraimo et al. (2007), using a different approach, underline the importance of within-game dynamics. They use the minute within the game as the unit of observation and the probability of a yellow card or a dismissal within each minute. The study uses data from two leagues (Premier League and 1^{st} Bundesliga) over the same time period and compares the results. Again, a distinction is made between stadiums with and without a racetrack around the football field. The matches are differentiated by teams which are classified as "favorite" or "underdog". The results do not differ from other studies and they find a home referee bias, too. Yet the study's methodology has further implications on favoritism. First, the authors establish that home teams which play in stadiums without a racetrack have a lower probability of receiving yellow and red cards. Further, the authors find that home teams which are not very successful and are considered the underdog of the match, will benefit from the referee bias.

Rickman and Witt (2008) analyze whether financial incentives help to control favoritism in hierarchical principal agent settings and consider that principals can use a combination of financial rewards and imperfect monitoring to incentivize their agents. Using football data, the authors examine whether a governing body (the "higher" principal) can influence favoritism displayed by referees (the principal) towards players/teams (the agents). The empirical approach is similar to other studies mentioned above. The authors divide their data into a Premier League with amateur referees (the pre-professional English Premier League) and a league with professional referees with greater financial incentives (post-professional Premier League). Their results show that favoritism disappears after professionalization of the English referees and that financial incentives for referees have a decreasing effect on the amount of extra time.

Scoppa (2008) tests empirically whether soccer referees decide neutrally as they should, or whether these decisions are biased by social pressure or corruptions. He examines the Italian "Serie A" seasons 2003/04 and 2004/05. In his analyzes, he not only controls for crowd pressure but also finds empirical evidence that teams which were later involved in the "Serie A Scandal" in 2006 are favored. Examining only games that were close at the end of second half, Scoppa (2008) uses injury time as an indicator of favoritism and finds significant effects whenever home teams lag behind by one goal. As an alternative measure, Scoppa (2008) uses the ratio of awarded penalties and goals to estimate whether home teams receive significantly more penalties than visiting teams, regardless of whether home teams tend to attack more often. On the basis of a *t-test* for the difference between these two averages, he rejects the hypothesis that the

[14] The referees receive a fixed salary for every match they judge.

probabilities of an awarded penalty are the same across these two groups of teams. Finally, the author concludes that biased referee decisions have no consequences on the final rankings in the league at the end of the season.

Using a random effects as well as a fixed effects approach, Page and Page (2010) find that the home advantage effect significantly differs across referees. They conclude that home advantage effects are increased by social pressure from the spectators and that the variability in these effects is evidence of significant individual differences between the referees.

In yet another study, Dawson (2012) examines whether a referee's experience influences his performance when social pressure and other factors are controlled for. He uses data from European competitions like the UEFA Champions League and the UEFA Cup for the years 2002/03 through 2006/07.[15] Within this data, the author controls for the quality of the two teams, the period of the competition, the number of spectators and the presence of a racetrack. The dependent variable is constituted by a weighted average of yellow and red cards. Dawson (2012) finds a negative correlation between referee experience and the number of sanctions. In a further step, he controls for an interaction between experience and crowd size. Again, he finds negative effects from more experienced referees on the number of sanctions.

Reilly and Witt (2013) use player/match-level data instead of match-level data from five seasons of the English Premier League and treat players as the unit of observation to determine whether a referee bias exists. The authors use red and yellow cards as the dependent variables for referee bias instead of injury time or penalties. Using a logit fixed effects model and a non-panel logit model, the authors conclude that referees penalize visiting players more than players from home teams but they find no statistically significant evidence that these differences in referee behavior are caused by social pressure.

Like Page and Page (2010) and Boyko et al. (2007), we are interested in the individual differences in referee bias, especially whether there are any effects on match outcome. For the purposes of this study the main focus lies on the number of points that are won (0,1,3) within a match. We also use regression analyzes but focus on referee fixed effects. In the studies discussed above, referee fixed effects are only used to control for specific referees. Now, we try to describe these fixed effects and establish whether they have a significant influence on match outcome, even though presumably there should be no significant referee effects.

[15] Since 2009 the UEFA Cup has been known as the Europa League.

2.3 German Referees

The following section provides information on the referee system in Germany. We describe how people interested in refereeing become a referee and how they are educated, and we depict the referee reward and monitoring system as well as the possibilities for promotion and relegation within the DFB.

Training of German Referees As Ebersberger et al. (1989) describe, a successful referee has to comply with the following expectations: A German referee is believed to referee a game impartially, to be fair, not to make any differences in the judgment of similar processes of the match or events, to treat every player equally, to make clear decisions, to be as close to match events as possible, to show players understanding (they must not react sensitively to criticism), to be consistent and unbiased. The aim of a referee is to avoid any disturbance on the football ground by players and spectators. To meet these expectations it is very important that a referee is physically fit. A good referee must also display mental flexibility. Furthermore, for top class referees it is really essential that they arrange their working time in a flexible way because the professional career must be subordinated to the activities of a referee (Strigel, 1999).

The organization of German referees mirrors the structure of the DFB (Ebersberger and Pohler, 1997), comprising different regional associations. Referees are allocated to the different associations on a geographic basis.[16] These associations are managed by an elected committee. The activities and the functions of the committee and its chairman are defined by the referee regulation of the regional associations. Their most important functions are education, monitoring, further training and qualification of the referees. The nomination of the referees for the matches is also a task of these referee groups.

After passing the exam, the referees only know the basics and are therefore expected to participate in further education.[17] Moreover, they are required maintain the necessary physical shape because only referees with a relative strong physical fitness are able to handle tough matches without any problems. Thus, ambitious referees must devote a lot of time to private study but must also participate in further training (Teipel et al., 1999).

At the beginning their careers, referees start in lower division matches (county level). There is an annual opportunity for promotion to higher leagues, successful referees having to achieve the admission into the "promotion squad" of the DFB.[18] Referees who are promoted to official DFB referees are subordinated to the DFB referee board, whose functions include classifying, assisting and educating the referees. A final and important step in the career of a successful referee is the nomination as FIFA referee with the chance to direct matches in international

[16] Cf. section 2.5.4.

[17] Further training is organized by the local district chairman for two evenings a month. Here the focus lies on individual rules for refereeing in soccer and the behavior of the referees on the pitch.

[18] For example, an eighteen year-old referee has the possibility to be promoted to the German Bundesliga within six years with the help of a special program of the DFB.

competitions (Artium and Rimkus, 2001).[19]

During their career it will become more difficult to fulfill these hard performance requirements with increasing age. Therefore age limits are defined on the national and international level. The DFB's age limit is 47 years, FIFA referees must not be older than 45 years.

As we have shown, referees receive specific education and training, and if a referee is promoted to 1^{st} Bundesliga he has to perform consistently well. This implies that the referees' career path in German football is highly selective. Since referees can be relegated too, they have strong incentives to be neutral and to perform well. Therefore, we would not expect to find any significant referee effects on match outcomes.

Remuneration of German Referees In Germany, referees are remunerated according to the league in which they referee. Matches in the third league pay €750, in the second league referees receive €1,800 per match and in the first division they are paid €3,800 (DFB, 2009). By this measure, 1^{st} Bundesliga referees rank second among the European leagues. At €6,000 per match, the Primera División is the only league where referees earn more than in Germany. In England, professional referees earn €1,170 per match but receive a base salary of €38,500 per season.[20] Despite such high allowances for referees, there are no plans to introduce the professional referee in Germany. Reasons include unresolved issues about consequences after a relegation or injury. However, since 2012/13 the DFB has honored their professional work and paid additional compensation. For example the five longest serving referees FIFA gain an additional €40,000 in that season. The two referees who were promoted from 2^{nd} to 1^{st} Bundesliga even gained €20,000.[21]

Monitoring German Referees Promotion or relegation of a referee depends on the one hand on performance tests which referees should pass four times in a year, and on the other hand on their monitoring results. Four criteria are relevant to evaluate the performance of referees: (1) physical fitness, (2) involvement outside the soccer field, (3) personality and (4) results of referee monitoring. The referee committee evaluates the referees' performance. In higher leagues this is done by former referees who use a detailed observation form. This observation is essential for DFB referees.

The DFB's observation form not only covers the performance of the referee, but also regard the performance of the referee assistants in a specific match. Employing interior and external

[19] Although the performance requirements change within professional leagues, a very talented referee will only need six to eight years to be promoted to DFB referee and another two years to become a FIFA referee. Only those who become referee at an early age have a viable chance to be nominated for the FIFA list (Strigel, 1999).

[20] Cf. `www.wahretabelle.de/news/dfl-streitpunkt-schiedsrichter-gehalter-/5480`, last access: June 11th, 2013.

[21] Another incentive for referees to perform well is the possibility to referee international matches. For example, Herbert Fandel, one of the most famous referees in Germany, talks about the special honor to referee a Champions League Final (DFB, 2007a). Furthermore there are yet more potential earnings for outstanding or famous referees after the end of their referee career. For example, the Swiss FIFA referee Urs Meier appears as an expert for referee decisions on German television. Besides the media, even businesses have indicated their interests in successful ex-referees. Since 2005, after retiring from the pitch, former world referee Dr. Markus Merk has worked as a motivational trainer.

1. External factors	2. Interior factors
Teams (local derby, promotion/relegation, neighbors in table)	Foul play (many fouls, unsportsmanlike behavior, dives)
Match ground and weather (slippy canvas)	
Attendance (large spectators setting, noisy fanatics, frantic trainer or substitute)	

Table 2.1: External and interior factors of a match.

factors a match is classified according to definite criteria (cf. table 2.1).[22]

Monitoring is essential, hence the DFB is still trying to increase efficiency of German referees. In the first and second division a so-called coach is appointed, who also supervises the referees. This coach shall review the match with the referee team and prepare the active referees for their next match using video analyzes. Observers and coaches evaluate the match and assess match-winning mistakes negatively. Monitoring is necessary to ensure a fair rating of performance and to discover talents (DFB, 2007b).

2.4 Data

Our data were mostly obtained from the company Impire AG.[23] Among others, Impire specializes in statistics on football matches and also provides data for TV broadcasts and Bundesliga football teams. From Impire, we use data about football matches such as matchday, home and away team, goals and referee decisions (e.g. yellow, yellow-red and red cards, not awarded goals or red cards etc.). In principle, these data are publicly available and it would be possible to collect them "by hand" from different football websites.[24] The German football magazine Kicker[25] follows one of these websites to collect information on referees like age, date of the first Bundesliga match, physical height and profession.[26] In total, this data set contains 4,590 German 1^{st} Bundesliga matches from seasons 1993/94 through 2007/08. The data contain 67 referees who directed at least two matches and who have up to 20 years of experience. Table 2.2 provides an overview of the number of 1^{st} Bundesliga matches that the referees directed each season. The average number of matches increases from 9.13 in 1993/94 to 17.13 in season 2007/08. Even the maximum number of matches for a single referee increases from 12 to 24. One reason for this rise is that the DFB reduced the total number of referees in the 1^{st} Bundesliga. In the season 1993/94, 32 referees were appointed in total whereas only 19 were nominated in 2007/08. Thus, it has become more difficult for referees to be promoted to the 1^{st} Bundesliga,

[22] Besides the external and interior factor it is also possible that a match becomes very sophisticated because of a weak referee performance (BFV, 2001).
[23] Cf. `www.impire.de`.
[24] Other possible data sources are `www.bundesliga.de` or `www.wahretabelle.de`.
[25] Cf. `www.kicker.de`.
[26] Remember that in Germany football referees do not work as professionals, therefore the DFB claim that all DFB referees have a regular occupation.

given that a referee can only be promoted if he outperforms a referee from the 1^{st} Bundesliga.

Year	Mean	SD	Min	Max
1993	9.13	1.83	2	12
1994	10.55	2.48	5	14
1995	12.13	3.78	3	18
1996	13.76	3.54	1	18
1997	14.61	2.15	1	18
1998	13.43	2.82	1	18
1999	14.24	3.93	1	21
2000	15.56	4.04	1	22
2001	14.77	3.92	2	24
2002	15.04	4.17	8	23
2003	15.93	4.59	1	24
2004	16.04	4.26	8	22
2005	17.88	3.7	9	25
2006	17.25	4.55	6	24
2007	17.13	3.97	9	24

Table 2.2: Matches per season and referee.

Table 2.3 describes how often referees umpire the same team within a season. Evidently, there is huge variation between referees and teams. Within 18 teams and 34 match days, the same referee and team meet only 1.34 to 1.82 times on average. Even the maximum lies at only 4 to 5 encounters. That implies that there is sufficient variation between referees and teams within a season, to examine individual referee effects in a next step.

Year	Mean	SD	Min	Max
1993	1.34	0.56	1	3
1994	1.42	0.62	1	4
1995	1.47	0.66	1	4
1996	1.56	0.71	1	4
1997	1.62	0.75	1	4
1998	1.53	0.69	1	4
1999	1.60	0.77	1	5
2000	1.67	0.79	1	4
2001	1.65	0.80	1	5
2002	1.66	0.79	1	5
2003	1.75	0.89	1	5
2004	1.74	0.88	1	5
2005	1.90	0.96	1	5
2006	1.83	0.94	1	5
2007	1.82	0.94	1	6

Table 2.3: Distribution of referees and teams per season.

Figure 2.1 shows the distribution between a referee and the same teams during a referee's tenure. Only few referees manage a particular team more than thirty times. The peak in this distribution is at about eight times, due to the fact, as discussed later in section 2.5.4, that a referee on average only stays in 1^{st} Bundesliga for seven years.

Table 2.4 shows descriptive statistics on referee decisions, taking into account the whole sample and the referee characteristics sample.[27] The latter sample only includes matches for which further information on the referee is available, such as profession, height and football association. Although there is a data loss of about 26% if only the referee characteristics subsample is applied, the descriptive statistics of both samples do not exhibit great differences. These small differences between both samples prevent potential issues when we use referee characteristics instead of referee fixed effects in our second stage regression. Table 2.4 shows that on average, home teams score 0.5 more goals per match than away teams. On average, three goals are awarded in a match, one match boasted as many as eleven goals. The numbers of not awarded goals and awarded penalties are very similar across the two samples.[28] On average, 0.7 penalties failed to be awarded per match, but in one match, this number stood at six. As table 2.4 reveals, referees brandish four yellow cards per match on acreage while the yellow-red and red cards are only drawn 0.1 times each. Lastly, table 2.4 shows that the number of red cards not awarded is 0.3 on average.

[27] Later on, this referee characteristics subsample is used to examine referee properties instead of referee fixed effects.

[28] Not awarded goals, penalties and red cards are given as the sum of wrong and disputable referee decisions.

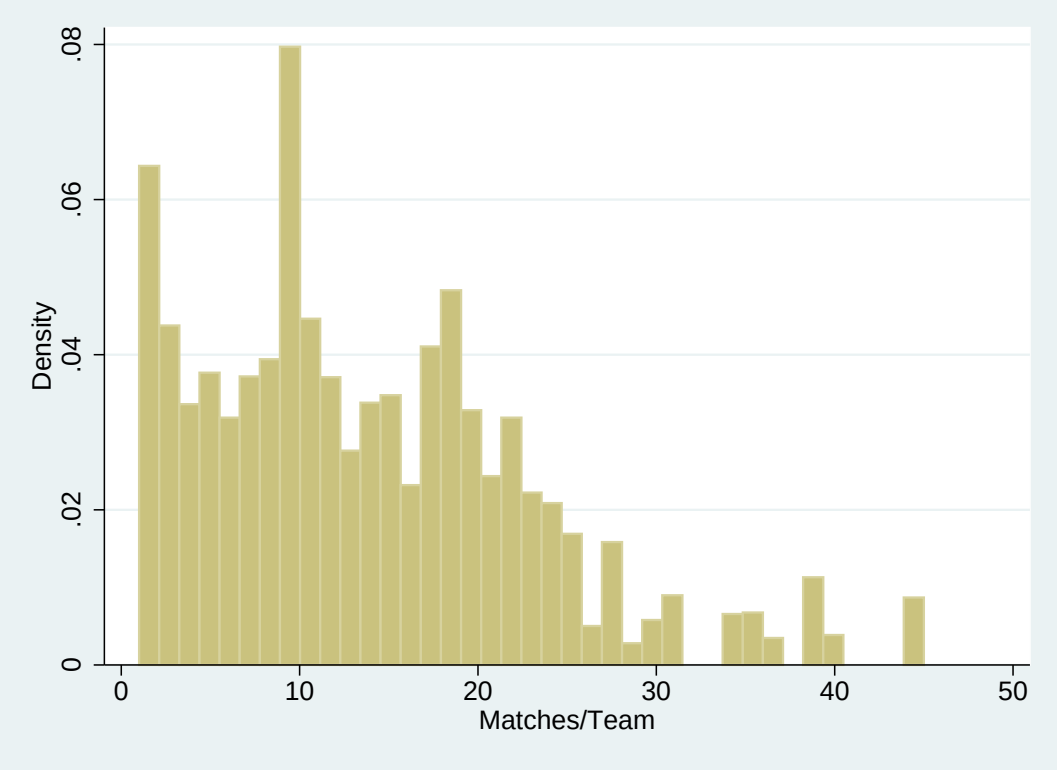


Figure 2.1: Distribution of referees on teams.

	Match-Referee-Sample					Referee-Characteristics-Sample				
	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Goal difference	0.497	1.787	-8	7	4,590	0.493	1.789	-8	7	3,383
Goals	2.866	1.707	0	11	4,590	2.867	1.711	0	11	3,383
Not awarded goals	0.244	0.486	0	3	1,779	0.243	0.483	0	3	1,615
Penalties	0.253	0.496	0	3	4,590	0.263	0.505	0	3	3,383
Not awarded penalties	0.711	0.911	0	6	1,721	0.714	0.911	0	6	1,559
Yellow cards	4.051	1.853	0	10	2,448	4.056	1.844	0	10	2,202
Yellow-red cards	0.125	0.355	0	3	2,448	0.12	0.35	0	3	2,202
Red cards	0.107	0.347	0	2	2,448	0.108	0.348	0	2	2,202
Not awarded red cards	0.332	0.629	0	4	1,721	0.337	0.632	0	4	1,559

Table 2.4: Descriptive statistics of dependent variables.

Table 2.5 displays the frequency of the match outcome variable, the result from the perspective of home teams. Home teams win three points in 48% of the matches. In 26% of the matches, the teams draw. Home teams lost only 26% of their matches between 1993/94 and 2007/08.[29] Again, table 2.5 shows no marked differences in frequencies between the whole sample and the referee characteristic subsample.

	Match-Referee Sample		Referee-Characteristics Sample	
Result	Absolute	Percent	Absolute	Percent
3	2,191	47.73	1,615	47.74
1	1,210	26.36	874	25.84
0	1,189	25.90	894	26.43

Table 2.5: Frequencies of the dependent variable result.

2.5 Individual Effects

This section illustrates the empirical results for individual referee effects. First, we estimate whether referees indeed have an individual effect on match outcome or referee decisions caused by personal characteristics. With respect to selection criteria, education and DFB evaluation, we expect to find no significant individual referee effects. But as we demonstrate in the next section, there is substantial empirical evidence of the joint significance of referee fixed effects.

This empirical approach owes to the literature on individual effects of managers (CEO, CFO etc.). Therefore, this section begins with a brief literature review of the most important studies on estimating individual manager effects on firm performance.

This strand of research was pioneered by Abowd et al. (1999). The authors first measure individual worker effects and firm fixed effects on wage. This means they examine the variation in personal wage rates holding firm effects constant and variation in firm wage rates holding person effects constant.

Bertrand and Schoar (2003) enhance this approach by estimating whether individual managers affect firm performance. Particularly they investigate how manager personalities, as opposed to firm, industry or market factors, explain unobserved differences in firm's success. The authors are interested in quantifying how much of the observed variation in firm performance can be explained by manager fixed effects. Therefore, they only use data on managers who change their jobs and work for at least two firms. Their variables of interest are different firm policy variables like investment or cash flow. In a first step, the authors run a regression without manager fixed effects (CEO, CFO, other). Next, they include CEO fixed effects only at first, followed by CEO, CFO and other fixed effects. An individual manager effect is found if the *p-value* from the *F-test* for joint significance of the manager fixed effects is significant and the explanatory power increases, after the manager fixed effects are included. Next, Bertrand

[29] This could be a small hint for home advantage.

and Schoar (2003) estimate different management styles while they interpret the correlation between the manager fixed effects for the different policy variables. Lastly, they repeat their first regression on corporate policy variables. But instead of manager fixed effects, they control for manager properties like birth cohort, tenure and possession of an MBA. Finding evidence that individual managers affect firm performance, the authors conclude that these manager effects are attributable to observable individual characteristics like education and age.

One problem of these results is that the estimated manager effects could be due to a selection bias: successful firms select successful managers. These possible endogeneity problems were addressed by Fee et al. (2010) and Graham et al. (2012).

Fee et al. (2010) capture the problem of CEO turnovers endogeneity and the difficulty of isolating manager effects. It is ambiguous whether variations in firm performance are the result of a particular management style or if instead there is a firm policy change which is accompanied with a CEO change. They use exogenous (death, health problems) as well as endogenous CEO turnovers and replicate the results of Bertrand and Schoar (2003). Fee et al. (2010) infer that it is uncertain whether varieties in firm performance are referable to individual manager effects or whether it is the effect from an underlying endogenous process.

Graham et al. (2012) examine the role of unobservable time-invariant firm and manager effects. For this purpose they use variations in executive pay as the dependent variable and examine whether fixed effects affect the interpretation and contribution of traditional explanatory variables.

Bennedsen et al. (2007), too, seek evidence that managers have notable effects on firm performance. But rather than looking at personal characteristics of CEOs like age or MBA, they are interested in whether CEO deaths or deaths of immediate family members (children, parents, mother-in-law) affect firm performance. The authors mention two advantages of this unusual approach. First, the above mentioned shocks will definitively impinge on the managers' performance because of the CEO death itself or because CEOs are distracted by the death of a family member. Second, they expect that these shocks will only influence the managers but have no direct effect on firm performance. This approach was chosen to again solve possible endogeneity problems because firms typically do not fire or appoint managers randomly.

2.5.1 Empirical Strategy

In contrast to the studies on manager effects, problems with endogeneity do not exist in our analyzes of individual referee effects. An advantage of the football setting is that the combination of home and away teams as well as the referee, changes every match day. Furthermore, a neutral institution (the DFB) prepares the referees' schedule.[30] Equation 2.1 illustrates our empirical approach in simplified terms:

[30] To counter the problem that the decision which referee is appointed for which match is based on prior performance the following regressions are clustered at team level.

$$y_{it} = \alpha_t + \gamma_i + \gamma_j + \beta X_{ijt} + \lambda_{Referee} + \epsilon \tag{2.1}$$

The dependent variable y_{it} stands for match outcome or referee decisions for team i at match t, α_t denotes time fixed effects and γ_i and γ_j are team fixed effects for the two contestants in a match. The vector X_{ijt} contains all time-variant control variables like the recent performance of both teams, the relevance of the match as well as the number of drives within a match. Lastly, referee fixed effects are denoted by $\lambda_{Referee}$.

In the final analyzes, for every variable of interest, the following equation is estimated by ordinary least squares regression:

$$\begin{aligned} y_{it} = \beta_0 + \alpha_t + \gamma_i + \gamma_j + \beta_1 * \chi_{it} + \beta_2 * \chi_{jt} \\ + \beta_3 * \tau_i + \beta_4 * \tau_j + \beta_5 * \mu_{it} + \beta_6 * \mu_{jt} \\ + \beta_7 * \theta_{it} + \beta_8 * \theta_{jt} + \lambda_{Referee} + \epsilon_{it} \end{aligned} \tag{2.2}$$

The variable y_{it} alternatively denotes variables like result, goal difference and referee decisions such as awarded yellow, yellow-red and red cards etc. Further, we note for every team whether it plays at home or away and control for this information. Season fixed effects are denoted by α_t, γ_i are team fixed effects and γ_j are fixed effects for the opposing team. Equation 2.2 is intended to capture factors that are known ex ante and that might influence the result or the events of the match. One important aspect is the recent performance of the teams, as captured by χ_{it} and χ_{jt}. These are vectors for the playing team and its opponent consisting of time-varying variables for short-term, medium-term and long-term past performance. A team's long-term performance is measured by its average league position in the last three seasons. Medium-term performance refers to a team's success during the current season, as measured by the average number of points in both home and away matches. Finally, as our measure of short-term performance, the average points are accumulated up to the observation of interest. Here, we use performance dummy-variables that describe the strength of a team over the last four matches. Specifically, we create a dummy variable for every team and for each of the last four matches of the following form: the team's recent performance prior to matchday t is $(y_{it-1}, y_{it-2}, y_{it-3}, y_{it-4})$ where $y_{it-k} \in 3, 1, 0$ for $k = 1, ..., 4$ indicates the number of points attained in the match which is played k match days before match day t. The same is done for the opponent, yielding in $3^4 = 81$ history dummy variables for each team.[31]

τ_i and τ_j are the relative budgets within a season for team and opponent, as calculated by the fraction of a team's absolute budget to the average absolute budget in the league in a given season. μ_{it} and μ_{jt} are variables for the relevance of a match (championship, relegation etc.) for both teams.

As already discussed (cf. section 2.3), it is important to know which kind of match is played. To describe this "match type", different variables are used to depict the strength and incentives

[31] This approach refers to the work of Hentschel et al. (2012).

of the two teams playing. First, the three performance measures for home and away teams are applied. Second, we control for whether a match is crucial for one of the two contestants. Third, θ_{it} and θ_{jt} signify the number of drives for both teams within a match. The drives are introduced to capture at least one match specific variable that might influence the match outcome and is independent of the referee. Finally, we control for referee fixed effects and therefore include referee dummy variables that are indicated by $\lambda_{Referee}$.

2.5.2 Results

Before discussing the results of equation 2.2, we reiterate the hypothesis of this study: We expect that referees in football matches act impartially and do not influence the match outcome. To estimate the individual influence of a referee, equation 2.2 is first run without the referee dummy variables. Subsequently, equation 2.2 is repeated with referee fixed effects. Evidence of significant influence of referees on our dependent variables is found if the adjusted R^2 increases after referee fixed effects are included and if the F-test for joint significance of these referee dummy variables is significant.[32]

Equation 2.2 is applied not only to match outcomes (result, goal difference) but, in a further step, also to referee decisions such as awarded yellow, yellow-red and red cards, awarded goals and penalties, as well as not awarded red cards, goals and penalties. Although, these decisions are at the discretion of the individual referee, we expect them to be made neutrally and strictly in accordance with DFB regulations. The results of these estimations are displayed in table 2.6 for home teams and in table 2.7 for away teams. The first three columns (model (I)) in each tables describe the result of equation 2.2, the next three columns (model (II)) describe the results for an extension of equation 2.2 by the time-variant referee characteristic years of experience. This is done as a robustness check. Even if we control for experience, individual effects from referees are found. The last columns (model (III)) of tables 2.6 and 2.7 describe the results for a model which, instead of relying on dummy variables to control for the short term performance of both teams, uses the average number of points won in the last four matches. This is done as a further robustness check to counter the potential problem of the much reduced number of degrees of freedom caused by this large number of dummy variables. The main focus of the discussion, however, is on the first three columns of tables 2.6 and 2.7.

The fourth row in each of of tables 2.6 and 2.7 show that the F-test for joint significance of the referee dummy variables are significant for the dependent variable result. The p-value for home teams is 0.000 and remains significant if we control for time-variant referee information. Further, we find a small increase in the adjusted R^2 by 1.2%, too. The findings for match result for away teams are less clear. As table 2.7 shows, the p-value for joint significance is 0.000 but the explanatory power does not increase if equation 2.2 is estimated for away teams. Only in model (III) do we find a significant referee effect on result for away teams.

[32] An OLS approach may justifiably be used even though the dependent variable has only three realizations because at the end of a match, three points are always better than one or zero points (cf. Angrist and Pischke, 2008). Nevertheless the following results were verified with an ordered probit model, which confirms our OLS estimates.

As the findings for home teams show, the F-test for joint significance of the referees are highly significant for every dependent variable like result, yellow card, goal or penalty. However a real individual effect exists only if the explanatory power increases after the referee dummy variables are added to the regression. We find that adjusted R^2 increases from 0.083 to 0.084 for the most interesting variable result. Stressed again that there should be no systematic effects, it is striking that the effects on other dependent variables are even higher. For example, the adjusted R^2 for an awarded yellow card increases from 0.044 to 0.050 which implies a percentage increase by 13.6%. If equation 2.2 is estimated for an awarded red card, the explanatory power increases by 87.5% and we even find a four percentage increase for an awarded goal. The explanatory power for the referee decision penalty increases by 73.3% once we control for referee fixed effects. In sum, we find significant individual effects from referees for home teams for result, goal difference and the following referee decisions: yellow card, yellow-red card, red card, awarded goal and penalty. These findings even hold for the other two models (II) and (III) where, however, the increase in explanatory power for the variable result is again very small.

Table 2.7 displays the findings of equation 2.2 for away teams. Again, we detect many significant F-tests for the joint significance of the referee variables. However, compared to the results for home teams, there are some differences in the increase of the explanatory power if we add referee fixed effects as shown in the third column of model (I). Indeed, we find no significant referee effects for the variable result inasmuch as the explanatory power does not increase once we control for referee fixed effects. Further, we find no significant effects for awarded yellow-red and red cards in away matches. The effect on goal difference yields a four percentage increase of the same direction as for home teams. Substantial effects are found for awarded yellow card, where the adjusted R^2 increases by 76%, 62 percentage points more than in home matches. Furthermore, adding the referee variables increases the explanatory power for the referee decision not awarded penalty by 5%. Again, we find similar results in the models (II) and (III). In sum, the results for away teams are markedly different from those for home teams. First, we find no significant referee effects for the variable result for away teams. Second, the total number of significant individual referee effects is smaller than for home teams.

Nevertheless, overall we find that German 1^{st} Bundesliga referees wield an individual influence on the game. Contrary to our expectation of referee neutrality, the empirical results indicate that this assumption is unwarranted for most referee decisions and in particular for match outcome.

Next, we are interested in the estimated referee fixed effects for match outcome and referee decisions to examine how heterogenous referees are with respect to these different decisions.

Table 2.8 shows descriptive statistics of the estimated referee fixed effects for match outcome and referee decisions. Although the average values of the fixed effects are relatively small, we find relatively large differences between the minimum and maximum values of the fixed effects. This is a hint of heterogeneity among referees. Again, German referees receive the same training, all their matches are monitored and their remuneration is equal within the divisions. Further

officials, trainer and fans would expect that referees are neutral and behave in the same manner. But as table 2.8 reveals, we find large differences for individual referee decisions and match outcome, moreover, there are also differences between home and away teams. Here, the average fixed effects for the match outcome variables result and goal difference differ only slightly. Yet the differences between home and away teams are much greater with respect to awarded and not awarded cards, as well as not awarded goals and penalties.

Home	(I)			(II)			(III)		
	F-test	N	Adj. R^2	F-test	N	Adj. R^2	F-test	N	Adj. R^2
Result	-	4,182	0.083	-	4,169	0.082	-	4,182	0.081
Result	*0.0000 (53.70, 32)*	*4,182*	*0.084*	*0.0000 (96.03, 32)*	*4,169*	*0.083*	*0.0000 (54.37, 32)*	*4,182*	*0.082*
Goal difference	-	4,182	0.098	-	4,169	0.100	-	4,182	0.095
Goal difference	*0.0000 (54.94, 32)*	*4,182*	*0.102*	*0.0000 (54.61, 32)*	*4,169*	*0.101*	*0.0000 (20.03, 32)*	*4,182*	*0.099*
Yellow card	-	2,231	0.044	-	2,231	0.044	-	2,231	0.047
Yellow card	*0.0000 (23.53, 26)*	*2,231*	*0.050*	*0.0000 (62.87, 26)*	*2,231*	*0.049*	*0.0000 (120.23, 26)*	*2,231*	*0.054*
Yellow-red card	-	2,231	0.025	-	2,231	0.026	-	2,231	0.025
Yellow-red card	*0.0000 (10.15, 26)*	*2,231*	*0.027*	*0.0000 (10.16, 26)*	*2,231*	*0.027*	*0.0000 (46.37, 26)*	*2,231*	*0.029*
Red card	-	2,231	0.008	-	2,231	0.008	-	2,231	0.012
Red card	*0.0000 (14.70, 26)*	*2,231*	*0.015*	*0.0000 (36.65, 26)*	*2,231*	*0.015*	*0.0000 (40.55, 26)*	*2,231*	*0.019*
Not awarded red card	-	1,855	0.028	-	1,855	0.029	-	1,855	0.028
Not awarded red card	*0.0000 (9.76, 26)*	*1,855*	*0.028*	*0.0137 (15.03, 26)*	*1,855*	*0.027*	*0.0000 (6.11, 26)*	*1,855*	*0.027*
Goal	-	4,182	0.076	-	4,169	0.075	-	4,182	0.074
Goal	0.0000 (87.23, 32)	*4,182*	*0.079*	*0.0000 (82.35, 32)*	*4,169*	*0.078*	*0.0000 (326.14, 32)*	*4,182*	*0.076*
Not awarded goal	-	1,903	-0.008	-	1,903	-0.008	-	1,903	-0.009
Not awarded goal	*0.0000 (24.99, 26)*	*1,903*	*-0.007*	*0.0000 (26.84, 26)*	*1,903*	*-0.007*	*0.0000 (17.82, 26)*	*1,903*	*-0.005*
Penalty	-	4,182	0.015	-	4,169	0.014	-	4,182	0.012
Penalty	*0.0000 (60.37, 32)*	*4,182*	*0.026*	*0.0000 (107.70, 32)*	*4,169*	*0.026*	*0.0000 (3,496.13, 32)*	*4,182*	*0.022*
Not awarded penalty	-	1,855	0.040	-	1,855	0.041	-	1,855	0.043
Not awarded penalty	*0.0000 (114.37, 26)*	*1,855*	*0.045*	*0.0000 (21.14, 26)*	*1,855*	*0.045*	*0.0000 (44.01, 26)*	*1,855*	*0.044*

a. Regressions with referee fixed effects are in italics.
b. Standard errors are clustered at the team level.
c. Reported are F-tests for the joint significance of the referee fixed effects.
d. For each F-test the p-value, the number of constraints and
the value of the F-statistic are reported.
e. Models (II) even controls for time-variant referee characteristics (age, experience).
f. Models (III) controls for average points in last four matches instead of match dummies.

Table 2.6: Referee fixed effects in home matches.

Away	(I)			(II)			(III)		
	F-test	N	Adj. R^2	F-test	N	Adj. R^2	F-test	N	Adj. R^2
Result	-	4,182	0.082	-	4,169	0.081	-	4,182	0.084
Result	*0.0000 (28.20, 32)*	*4,182*	*0.082*	*0.0000 (34.05, 32)*	*4,169*	*0.081*	*0.0000 (23.03, 32)*	*4,182*	*0.085*
Goal difference	-	4,182	0.099	-	4,169	0.098	-	4,182	0.097
Goal difference	*0.0000 (111.19, 32)*	*4,182*	*0.103*	*0.0000 (63.51, 32)*	*4,169*	*0.101*	*0.0000 (35.98, 32)*	*4,182*	*0.102*
Yellow card	-	2,231	0.021	-	2,231	0.028	-	2,231	0.024
Yellow card	*0.0000 (5.86, 26)*	*2,231*	*0.037*	*0.0000 (11.68, 26)*	*2,231*	*0.037*	*0.0000 (30.78, 26)*	*2,231*	*0.044*
Yellow-red card	-	2,231	0.001	-	2,231	0.000	-	2,231	0.016
Yellow-red card	*0.0000 (19.04, 26)*	*2,231*	*-0.004*	*0.0000 (13.79, 26)*	*2,231*	*-0.005*	*0.0000 (205.93, 26)*	*2,231*	*0.012*
Red card	-	2,231	0.015	-	2,231	0.015	-	2,231	0.009
Red card	*0.0000 (10.22, 26)*	*2,231*	*0.011*	*0.0000 (6.41, 26)*	*2,231*	*0.010*	*0.0000 (10.90, 26)*	*2,231*	*0.005*
Not awarded red card	-	1,894	0.015	-	1,894	0.018	-	1,894	0.035
Not awarded red card	*0.0000 (31.18, 26)*	*1,894*	*0.015*	*0.0000 (23.59, 26)*	*1,894*	*0.027*	*0.0000 (19.79, 26)*	*1,894*	*0.045*
Goal	-	4,182	0.076	-	4,169	0.076	-	4,182	0.075
Goal	*0.0000 (14.42, 32)*	*4,182*	*0.079*	*0.0000 (35.13, 32)*	*4,169*	*0.078*	*0.0000 (13.53, 32)*	*4,182*	*0.077*
Not awarded goal	-	1,896	-0.010	-	1,896	-0.011	-	1,896	-0.012
Not awarded goal	*0.0003 (4.16, 26)*	*1,896*	*-0.010*	*0.0000 (12.85, 26)*	*1,896*	*-0.008*	*0.0000 (34.20, 26)*	*1,896*	*-0.015*
Penalty	-	4,182	0.000	-	4,169	0.000	-	4,182	0.004
Penalty	*0.0000 (17.55, 32)*	*4,182*	*0.000*	*0.0000 (17.26, 32)*	*4,169*	*-0.001*	*0.0000 (15.61, 32)*	*4,182*	*0.003*
Not awarded penalty	-	1,894	0.041	-	1,894	0.040	-	1,894	0.037
Not awarded penalty	*0.0000 (9.30, 26)*	*1,894*	*0.043*	*0.0000 (13.27, 26)*	*1,894*	*0.043*	*0.0000 (36.25, 26)*	*1,894*	*0.036*

a. Regressions with referee fixed effects are in italics.
b. Standard errors are clustered at the team level.
c. Reported are F-tests for the joint significance of the referee fixed effects.
d. For each F-test the p-value, the number of constraints and
the value of the F-statistic are reported.
e. Models (II) even controls for time-variant referee characteristics (age, experience).
f. Models (III) controls for average points in last four matches instead of match dummies.

Table 2.7: Referee fixed effects in away matches.

	Home			
	Mean	SD	Min	Max
Result	-5.80e-10	.1672753	-1.415832	1.29575
Goal difference	-1.99e-10	.2486666	-1.411123	1.716284
Yellow card	-1.82e-09	.1885183	-1.025217	.5073101
Yellow red card	2.35e-11	.0346896	-.0806199	.2976502
Red card	1.56e-10	.0351929	-.0595983	.3171462
Not awarded red card	-1.13e-12	.065652	-.1065028	.3679781
Goal	-2.08e-10	.1805578	-1.012399	2.156752
Not awarded goal	-2.23e-10	.0550492	-.1632138	.0956036
Penalty	-5.33e-11	.0592822	-.3025232	.6140327
Not awarded penalty	3.68e-10	.0989653	-.3688249	.836126
	Away			
	Mean	SD	Min	Max
Result	-1.65e-10	.1573011	-1.046284	1.110669
Goal difference	-1.09e-09	.2476694	-1.753304	1.490386
Yellow card	1.46e-10	.2383328	-.4724632	1.66844
Yellow red card	1.38e-10	.0271906	-.0638093	.1954479
Red card	7.44e-11	.0263356	-.0898046	.1596946
Not awarded red card	2.88e-11	.0731037	-.2125522	.4440277
Goal	-2.47e-10	.181015	-1.0225	2.143388
Not awarded goal	-1.73e-10	.0508763	-.1465778	.3474265
Penalty	-5.87e-12	.0442329	-.212919	.3629287
Not awarded penalty	-3.50e-10	.1025311	-.404209	.3048013

Table 2.8: Distribution of referee fixed effects for home and away teams.

Next, we take a detailed look at the descriptive statistics for the estimated referee fixed effect for result and again differentiate between home and away teams. As tables 2.9 and 2.10 show, for most 1^{st} Bundesliga teams we find a positive referee fixed effect on result in home matches but a negative effect in away matches. In particular, we find that the most successful teams (Borussia Dortmund and FC Bayern München) always have a positive referee fixed effect.[33] For the other champions (SV Werder Bremen, 1. FC Kaiserslautern, VfB Stuttgart, VfL Wolfsburg) we detect different directions of the referee fixed effects. Borussia Dortmund on average profit from referee influence by 0.03 points in away matches and 0.004 points in home matches. For FC Bayern München it is the other way around. There is a higher referee fixed effect when they play at home (0.01 points) than in away matches (0.004 points). Compared to other top teams like Borussia Dortmund or Bayer Leverkusen, they have the highest fixed effect in home matches. Even teams like SSV Ulm or St. Pauli always benefit from a referee bias. Although they are less successful in the league, a positive referee fixed effect in home and away matches is noticed. However there are also teams which are at disadvantage. On average, we detect a negative fixed effect in both home and away matches (e.g. SC Freiburg, Eintracht Frankfurt and 1. FC Kaiserslautern). Further, we notice that some teams are disadvantaged home matches but benefit on average in away matches (e.g. Hertha BSC Berlin, 1. FC Köln and VfL Wolfsburg).

[33] These two teams most frequently won the championship during the observation period.

Home	Mean	SD	Min	Max
1860 München	.0052523	.1749722	-.5617493	.8673021
Alemannia Aachen	.0896571	.124376	-.0823698	.2495471
Arminia Bielefeld	.0191787	.148957	-.3186742	.6182071
VfL Bochum	-.0247936	.1355948	-.5617493	.2495471
SV Werder Bremen	.0040598	.1684069	-1.416454	.8585901
Energie Cottbus	.0088116	.1362239	-.1961861	.6182071
Borussia Dortmund	.0043385	.1463868	-.4383083	.8585901
Dynamo Dresden	-.0183446	.2798817	-.5617493	.6993073
MSV Duisburg	-.009987	.2024458	-.5617493	1.288919
Fortuna Düsseldorf	-.0094712	.117828	-.1961861	.2295575
FC Bayern München	.0104091	.1555682	-.5617493	.6993073
Eintracht Frankfurt	-.0097971	.2141312	-1.170888	1.288919
SC Freiburg	-.0331324	.1882278	-1.416454	.4370036
Hamburger SV	.0011524	.1502966	-.4394639	.8585901
Hannover 96	-.0073946	.1515385	-.3186742	.6182071
Hertha BSC Berlin	-.0027638	.1243482	-.3186742	.6182071
1. FC Kaiserslautern	-.0117017	.1628301	-.5617493	.8673021
KFC Uerdingen	.0085191	.2075909	-.2827318	.8673021
Karlsruher SC	.0421013	.2669551	-.5617493	1.009511
1. FC Köln	-.0037601	.1806482	-.4394639	.8673021
Bayer Leverkusen	-.0107875	.1535187	-1.170888	.6993073
Borussia Mönchengladbach	-.0079929	.1694887	-.5617493	.8673021
FSV Mainz	.0019298	.1339082	-.1796232	.2495471
1. FC Nürnberg	.0367727	.1688824	-.5617493	.6182071
Hansa Rostock	.0143998	.1557986	-.4394639	.8673021
SSV Ulm	.034103	.0640583	-.0823698	.13769
FC Schalke 04	.0067866	.1551749	-.416149	.8673021
St. Pauli	.0109855	.1368273	-.2827318	.3167353
VfB Stuttgart	-.0009213	.1662575	-.5617493	1.009511
SpVgg Unterhaching	-.0064961	.1132971	-.1961861	.2223046
VfB Leipzig	.0599541	.3730594	-.4383083	.8585901
SG Wattenscheid 09	-.0796515	.3453733	-.5617493	.8585901
VfL Wolfsburg	-.016441	.1262594	-.3186742	.6182071

Table 2.9: Referee fixed effects by home teams.

Away	Mean	SD	Min	Max
1860 München	-.0085872	.1304636	-.7057701	.266419
Alemannia Aachen	-.0280734	.2268007	-.678037	.2054478
Arminia Bielefeld	-.0085152	.1250609	-.3456424	.3438875
VfL Bochum	.0031919	.1396308	-.3456424	.5555463
SV Werder Bremen	-.0195301	.1533629	-.7360463	.1855414
Energie Cottbus	-.0145862	.1672451	-.678037	.2054478
Borussia Dortmund	.030453	.1370438	-.3456424	.5555463
Dynamo Dresden	-.041429	.2816364	-.7360463	.5003968
MSV Duisburg	.0061462	.2167913	-.7360463	1.114005
Fortuna Düsseldorf	-.0108255	.1650759	-.7057701	.1669814
FC Bayern München	.0044773	.1277826	-.605759	.4915302
Eintracht Frankfurt	-.0170931	.1783241	-.7057701	.5555463
SC Freiburg	-.0136305	.196713	-1.048431	.5555463
Hamburger SV	-.003698	.1306014	-.605759	.5003968
Hannover 96	.013683	.1358187	-.678037	.2054478
Hertha BSC Berlin	.002872	.1165343	-.3456424	.3438875
1. FC Kaiserslautern	-.004284	.1488604	-.7360463	.5555463
KFC Uerdingen	-.0397647	.2166637	-.7057701	.4915302
Karlsruher SC	-.0430239	.2202952	-.7360463	.5555463
1. FC Köln	.0297956	.1386513	-.3456424	.5555463
Bayer Leverkusen	.0084856	.1537897	-.605759	.5555463
Borussia Mönchengladbach	.0235815	.1708094	-1.048431	.5555463
FSV Mainz	-.0192729	.1641353	-.678037	.2054478
1. FC Nürnberg	-.0055198	.1768509	-.678037	.5555463
Hansa Rostock	-.0087318	.1197101	-.678037	.2054478
SSV Ulm	.0278322	.0925586	-.0866614	.1855414
FC Schalke 04	.0198137	.1389567	-.7057701	.4915302
St. Pauli	.0054002	.0984007	-.3456424	.1855414
VfB Stuttgart	-.0192192	.1615824	-.7057701	.5555463
SpVgg Unterhaching	.011025	.1092812	-.2026099	.1855414
VfB Leipzig	-.0425806	.4112404	-.7360463	1.107652
SG Wattenscheid 09	-.133586	.2893743	-.7870661	.1855414
VfL Wolfsburg	.0267286	.1503846	-.678037	1.114005

Table 2.10: Referee fixed effects by away teams.

2.5.3 Referee Styles

In the previous section, we found empirical evidence that referees have significant impact on match outcome, match winning variables (goals, penalties) and general decisions (yellow, red cards). Next, we are interested in whether a significant referee fixed effect on match outcome or referee decision has a positive or negative significant effect on other significant referee effects from tables 2.6 and 2.7. Therefore we estimate so-called "referee styles". Hence, the following regression is executed for every significant individual referee fixed effect that we found in the set of regressions above:[34]

$$F.E.(y)_{ijt} = \alpha + \beta F.E.(z)_{ijt} + \epsilon_{ijt} \tag{2.3}$$

where j indexes referees, and y and z are any two variables for match outcome or referee decisions with a significant p-value for the F-test and an increased explanatory power in the results. Since the right hand variable is an estimated coefficient and therefore noisy by definition, a GLS estimation technique is used to account for (possible) measurement error. Further, to ensure a comparability of these different fixed effects, we use the standardized coefficients in equation 2.3. Thus, the dimension for every coefficient is the standard deviation (SD).[35]

Tables 2.11 and 2.12 display the estimated referee styles for home and away teams, respectively. Remember that we only use variables for match outcome and referee decisions with significant individual referee effects. Hence, these two tables look different from each other. The first columns of tables 2.11 and 2.12 display the left hand side variable of equation 2.3. As the second column of table 2.11 shows, referees with a high effect on match results also have a positive effect on most of the other referee decisions in home matches. The individual effect for an awarded yellow card increases by 0.17 SD on average if the fixed effect for result increases by one standard deviation. The effect for an awarded goal is even higher. An increase in the fixed effect for result is associated with an increase in the individual effect for an awarded goal by 0.71 SD. In away matches, the effect of goal difference on the other referee decisions is less obvious.[36] As table 2.12 indicates, we find two negative effects from goal difference. Thus, if the individual effect for goal difference increases by one SD the individual effect for an awarded yellow card decreases by 0.23 SD and the fixed effect for an awarded goal decreases by 0.79 SD. Yet goal difference also has a positive effect: if the fixed effect for goal difference increases by one SD, the fixed effect for a not awarded penalty increases by 0.05 SD.

[34] The variable goal difference is not used for referee styles in home matches, because the result is calculated from goal difference and this would lead to high partial coefficients.

[35] Using Z-scores moreover has the advantage that they follow the normal distribution. Z-scores are calculated as the difference between the variable and its mean divided by its SD.

[36] We only find a significant referee effect for the match outcome variable "goal difference" if we control for away teams.

y_{ijt}	Result	Yellow card	Yellow-red card	Red card	Goal	Penalty	Not awarded penalty
Result	-						
Yellow card	0.17***	-					
Yellow-red card	0.03*	0.2***	-				
Red card	0.26***	0.26***	-0.02	-			
Goal	0.71***	0.43***	0.09**	0.24***	-		
Penalty	0.2***	0.24***	0.02	0.15***	0.3***	-	
Not awarded penalty	-0.08***	0.16***	0.3***	-0.35***	-0.04**	-0.06**	-

$*p < 0.05, **p < 0.01, ***p < 0.001$

Table 2.11: Referee styles in home matches.

y_{ijt}	Goal difference	Yellow card	Goal	Not awarded penalty
Goal difference	-			
Yellow card	-0.23***	-		
Goal	-0.79***	0.34***	-	
Not awarded penalty	0.05**	0.2***	-0.05**	-

$*p < 0.05, **p < 0.01, ***p < 0.001$

Table 2.12: Referee styles in away matches.

Further positive and significant correlations between the individual referee fixed effects are found for home matches. An increase in the individual effect for an awarded yellow card by one SD increases the individual effect for an awarded goal by 0.43 SD. If the individual effect for an awarded yellow-red card increases by one SD, the individual referee effect for not awarded penalties increases by 0.3 SD. Further, if the fixed effect for an awarded red card increases by one SD, the individual effect for an awarded goal increases by 0.24 SD while the individual effect for not awarded penalties decreases by 0.35 SD. We conclude that referees with a significant individual effect on result for home teams have positive effects on referee decisions during a match, too. Referees with individual effects on brandishing a yellow, yellow-red or red card also have positive impacts on awarding goals.

Again, the effects between the significant individual referee effects are less decisive for away teams. If the individual referee effect for an awarded yellow card increases by one SD, the effect for an awarded goal increases by 0.34 SD and the referee fixed effect for a not awarded penalty increases by 0.2 SD.

Altogether, we find different individual referee effects between home and away teams and significant correlations between these single individual referee fixed effects. Once more, there is evidence that referees do not strictly perform to the DFB rules and tend to behave differently across home and away teams.

2.5.4 Referee Characteristics

This section focuses on observable referee characteristics and whether we find some empirical support that these characteristics explain the significant individual referee effects. One possible explanation for this referee bias could be that referees, like judges and managers follow career concerns.[37]

The career concerns model assumes that at the beginning of a career in a principal-agent-relationship, an agent's talent is unobservable.[38] The principal's expectations about the agent's effort in the future are formed on basis of an agent's effort today. Therefore, uncertainty about an agent's ability declines over time (Borland, 1992). This implies that an agent can increase her future incomes by expending high effort today. However, the influence on future incomes decrease towards the end of an agent's career. Thus, at the beginning of a career it is worthwhile to invest in reputation and choose a high effort because there is great uncertainty about the agent's talent. At the end of a career, a high effort is merely costly, so the optimal effort becomes very small or even zero (Holmström, 1999). Regarding referees in football, we assume that they follow career concerns, too.[39] So, we assume that referee bias is low in the first years of a referee's career but increases when a referee finishes his career.

Further, a referee fixed effect could be driven a referee's last performance. We assume that if a referee performs bad in his last match, he has an incentive to achieve better in the actual match. On the one hand, we include his Kicker grade from his last match but we also use the difference in matchdays where a referee has to manage a football game.[40] If a referee has to pause for several matchdays that could be a hint for poor performance, too. Further, we control for referee's height. We would expect that a taller referee is faster, has a better overview of the match or may be even more assertive due to his posture.

Table 2.13 displays descriptive statistics for the referees in the data set. The referees were on average born in 1962 and have about seven years of experience in 1^{st} Bundesliga matches, Dr. Markus Merk with 20 years of experience constituting an outlier.[41] The average Kicker grade in our referee sample is 3.3, some referees are graded "very good" while others "fail". On average, referees pause 2.26 matchdays, the maximum being 23 matchdays. The average height is at least 185 centimeters. There is a balance of matches directed by FIFA referees (51%) and non-FIFA referees (49%).[42]

[37] The relevant literature on judges and manager comprises Levy (2005), Miceli and Coşgel (1994), Bertrand and Schoar (2003), Adams et al. (2005), Frank and Goyal (2007), among others.

[38] For a detailed theoretical discussion on career concerns, also see Holmström (1982, 1999). A survey on modeling career concerns in organizations was done by Gibbons and Waldman (1999). Additional theoretical and empirical evidence on career concerns is presented in Gibbons and Murphy (1992), Irlenbusch and Sliwka (2006) and Koch et al. (2009), among others.

[39] Other studies on referee bias also account for career concerns (cf. Nevill et al., 2002; Boyko et al., 2007; Boeri and Severgnini, 2008; Dawson, 2012)

[40] After every match the experts of Kicker magazine grade the performance of the referee on objective criteria using marks of 1 (good performance) to 6 (poor performance).

[41] Experience is calculated as the difference between the current match date and date of individual referee's Bundesliga premiere.

[42] FIFA is a dummy variable which equals one if the referee in a given match is a FIFA referee.

	Mean	SD	Min	Max
Year of birth	1962	6.15	1946	1983
Experience	6.92	4.26	0	20
Kicker	3.3	1.12	1	6
Matchday difference	2.26	1.66	0	23
Height (cm)	185	4.68	172	198
FIFA	Absolute	%		
0	2,250	49.21		
1	2,322	50.79		

Table 2.13: Descriptive statistics of referee characteristics.

Besides these referee characteristics, we are also interested in the referee's professional jobs because on the one hand referees do not work as professionals and on the other hand they are expected to subordinate their working live off the pitch to their work as a referee. Figure 2.2 show that referees who work in the trading and medical sectors constitute the largest fraction.[43] Only 7% of the referees work as engineers, around 13% work in the public sector and 10% are lawyers.

Tables 2.14 to 2.17 describe the effects of referee characteristics on the different match outcome variables and referee decisions for home and away teams. The estimated coefficients are generated from a slightly altered version of equation 2.2. Following Bertrand and Schoar (2003) we substitute the referee fixed effects from equation 2.2 with information from the referee characteristic subsample:

$$\begin{aligned} y_{ijzt} = {} & \beta_0 + \alpha_t + \gamma_i + \gamma_j + \beta_1 * \chi_{it} + \beta_2 * \chi_{jt} \\ & + \beta_3 * \tau_i + \beta_4 * \tau_j + \beta_5 * \mu_{it} + \beta_6 * \mu_{jt} + \beta_7 * \theta_{it} + \beta_8 * \theta_{jt} \\ & + \beta_9 * fifa_{zt} + \beta_{10} * birth_{zt} + \beta_{11} * experience_{zt} + \beta_{12} * height_z \\ & + \beta_{13} * kicker_{zt-1} + \beta_{14} * red_{zt-1} + \beta_{15} * penalty_{zt-1} \\ & + \beta_{16} * matchdaydiff_{zt} + \beta_{17} * X_z + \beta_{18} * Y_z + \epsilon_{izt} \end{aligned} \tag{2.4}$$

where y_{ijzt} stands for the same dependent variables from equation 2.2 and z denotes the referee in that match. Here, the focus lies on the referee characteristics. Thus, $fifa_z$ stands for the above mentioned dummy variable indicating whether a match is managed by a FIFA referee. Our expectation is that FIFA referee status is negatively correlated with the dependent variables because the status is precious and FIFA referees are therefore especially interested in maintaining a good reputation. $birth_{zt}$ is referee's year of birth and $experience_{zt}$ denotes the years of referee's experience. Assuming that referees follow career concerns we expect negative

[43] The professions are subsumed as trading (commercial clerk, banker, controller, business economist, master of communications, public administration specialist, export and import merchant, key account manager, financial adviser, sales manager) medical (doctor, dentist), lawyer (jurist, chief of chancery), public (teacher, public administration specialist, chief of regulatory agency, policeman), engineer and other (production manager, hotelier, pianist, sports scientist, electrical mechanic, welder).

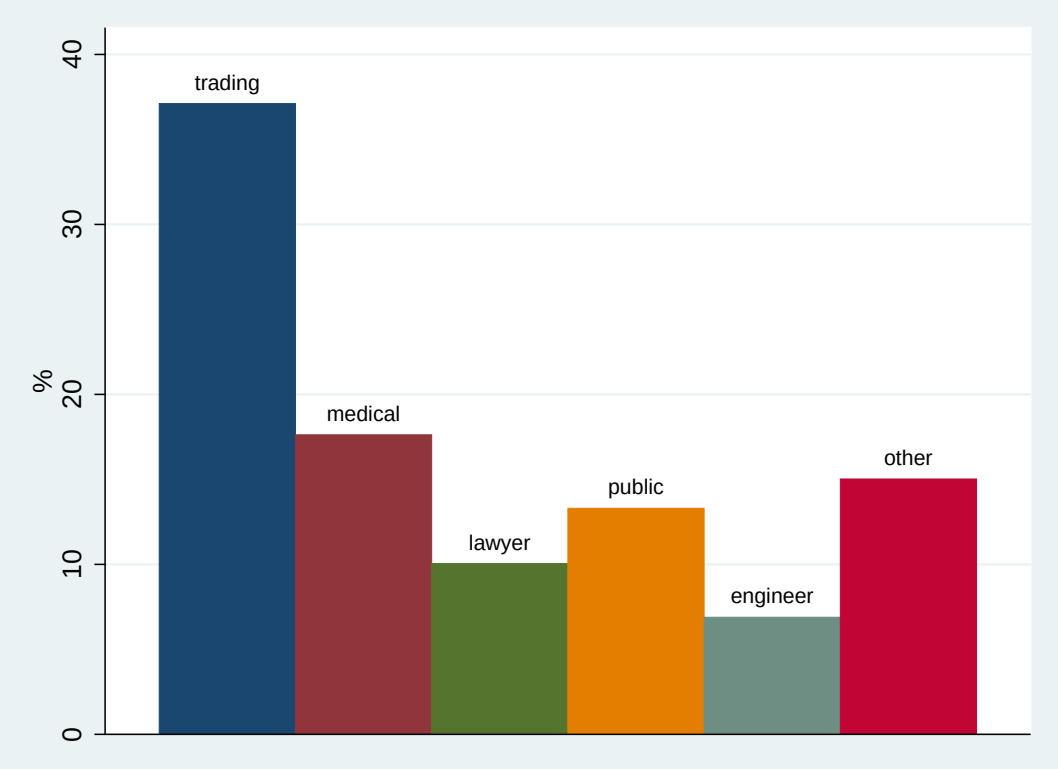


Figure 2.2: Distribution of referee professions.

effects from years of experience, but if we control for a non-linear relationship between experience and the dependent variables, we expect that effect will swap signs at some point. $height_z$ is the variable for a referee's height in centimeters. $kicker_{zt-1}$ is the referee's Kicker mark for the previous match. We expect that a referee with a high kicker mark in the previous match (poor performance) has an incentive to perform better in the current match. Further, we assume that referees try to avoid any patterns in their referee decisions. Therefore, we also include two dummy variables for notable referee decisions in the previous match. red_{zt-1} and $penalty_{zt-1}$ are equal to one if the referee issued a red card or a penalty, respectively, in his last match. Another hint for referee performance might be the time between two consecutive matches that a referee is nominated for. If a referee performs poorly he is prescribed a little break before his next appearance. This pause is indicated by $matchdaydiff_{zt}$. X_z is a vector with information on the regional association of a referee.[44] Y_z is a vector with information about a referee's job off the football pitch, working in a professional job being a requirement for a referee in Germany. To avoid serial correlation we again use robust standard errors.

[44] These associations are Norddeutscher Fußballverband (NFV), Westdeutscher Fußball- und Leichtathletikverband (WFLV), Fußball Regional Verband Südwest (FRVS), Süddeutscher Fußballverband (SFV) and Nordostdeutscher Fußballverband (NOFV).

Tables 2.14 and 2.15 show the results for individual referee effects on home teams. At first, we find no significant effects from year of birth and experience on match outcome and further referee decisions. In a next step, we control for a non-linear (quadratic) influence of experience in equation 2.4 to examine whether referees follow career concerns. Again we find no significant coefficients for experience and therefore no evidence of career concerns.[45]

Although, we expect that referees heed their reputation, the evidence does not indicated so. FIFA referees having more to lose, we would expect a negative effect of our status variable. Yet we find weak effects of ambiguous direction from the FIFA dummy-variable. However, the negative but insignificant effect of the FIFA dummy prevails if we control for referee effects for home teams.

Controlling for height also reveals no significant coefficients for home teams. Furthermore, the magnitude of these coefficients is very small. However, it seems that a referee's previous performance has a small significant impact on referee decisions for home teams in the current match: if the Kicker grade of the last match increases about one grade, the number of awarded yellow-red cards in the current match decreases by 0.0118 units. But the impact on awarded red cards is reverse, a plus of 0.0113 units. Further, we find no impact on match outcome or referee decisions in the current match once a referee brandished a red card in his last match. Yet if a referee awarded a penalty in his last match, the number of awarded yellow cards decreases by 0.151 units and the number of not awarded penalties decreases by 0.0779 units in the current match.

Further, controlling for regional associations leads to some significant differences between these associations for the referee decision "awarded red card". Compared to the association for the Norddeutscher Fußballverband (NOFV), we find significant negative coefficients for the Fußball Regional Verband Südwest (FRVS) and the Süddeutscher Fußballverband (SFV). Moreover, there is a significant positive impact from the Westdeutscher Fußball- und Leichtathletikverband (WFLV) for the referee decision "awarded penalty" compared to the reference category. The magnitude of the effect for this association is much greater than for the other regional associations.

Finally, we also check whether the referee's main professions have significant effects on match outcome and referee decisions. We find negative effects on the referee decisions awarded yellow (-0.417) and yellow-red card (-0.0717) if a referee work as an engineer compared to referees who work in the "other" sector (e.g. hotelier, sports scientist, welder).

In sum, our referee characteristics cannot explain the significant individual referee effects on result for home teams, though, we do find some significant effects for awarded yellow, yellow-red and red cards as well as for awarded and not awarded penalties.

[45] For reasons of clarity these regression results are presented in tables 2.18 and 2.19 in section 2.A1.

	(1) **Result**	(2) **Goal difference**	(3) **Yellow card**	(4) **Yellow-red card**	(5) **Red card**
FIFA	-0.136	-0.151	0.0612	-0.00499	-0.0129
	(0.0726)	(0.116)	(0.0861)	(0.0104)	(0.0160)
Year of birth	0.000701	-0.00521	0.00786	-0.00469	-0.00210
	(0.0193)	(0.0267)	(0.0166)	(0.00274)	(0.00308)
Experience	0.00626	0.00446	-0.00670	-0.00439	0.00126
	(0.0190)	(0.0272)	(0.0139)	(0.00277)	(0.00284)
Height	0.00424	0.0146	0.0131	0.000438	0.000666
	(0.0123)	(0.0178)	(0.0108)	(0.00189)	(0.00153)
Kicker last match	-0.0285	-0.0307	-0.0327	-0.0118*	0.0113*
	(0.0284)	(0.0329)	(0.0391)	(0.00493)	(0.00449)
Red card last match	0.106	0.144	0.0734	0.0128	-0.00900
	(0.0993)	(0.135)	(0.147)	(0.0220)	(0.0217)
Penalty last match	0.107	0.150	-0.151*	0.0260	0.00351
	(0.0889)	(0.121)	(0.0691)	(0.0130)	(0.0174)
Matchday difference	-0.0145	-0.0119	0.00608	-0.00371	0.00202
	(0.0327)	(0.0328)	(0.0263)	(0.00423)	(0.00397)
NFV^{+}	0.0887	0.0835	-0.00156	0.0362	-0.0104
	(0.151)	(0.191)	(0.173)	(0.0313)	(0.0255)
WFLV^{+}	-0.0388	-0.192	0.0355	0.00573	-0.0376
	(0.166)	(0.230)	(0.157)	(0.0289)	(0.0258)
FRVS^{+}	-0.0180	-0.168	-0.194	0.0110	-0.0590*
	(0.142)	(0.190)	(0.180)	(0.0310)	(0.0267)
SFV^{+}	-0.0112	-0.0907	-0.188	0.0234	-0.0480*
	(0.123)	(0.148)	(0.132)	(0.0242)	(0.0219)
Trading^{++}	-0.162	-0.300	-0.0431	-0.0262	-0.0316
	(0.111)	(0.171)	(0.114)	(0.0177)	(0.0259)
Medical^{++}	-0.153	-0.212	0.100	-0.0284	-0.0157
	(0.127)	(0.161)	(0.106)	(0.0182)	(0.0246)
Lawyer^{++}	-0.0509	-0.0581	0.0617	-0.0338	-0.00163
	(0.185)	(0.237)	(0.150)	(0.0253)	(0.0373)
Public^{++}	0.0463	0.0266	-0.0352	-0.0439	-0.0430
	(0.200)	(0.273)	(0.196)	(0.0343)	(0.0367)
Engineer++	-0.300	-0.413	-0.417*	-0.0712**	-0.0104
	(0.270)	(0.327)	(0.175)	(0.0207)	(0.0346)
N	1,928	1,928	1,928	1,928	1,928
Adj. R-sq	0.100	0.133	0.039	0.027	0.004

a. Bold printed models have significant individual referee effects in the first stage.

b. Standard errors in parentheses, c. Standard errors are clustered at team level.

d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.14: Referee characteristics in home matches I.

	(6) Not awarded red card	(7) **Goal**	(8) Not awarded goal	(9) **Penalty**	(10) **Not awarded penalty**
FIFA	-0.00930	-0.121	-0.0351	0.0209	0.0419
	(0.0334)	(0.0805)	(0.0233)	(0.0190)	(0.0464)
Year of birth	-0.00220	-0.00583	-0.00407	0.00143	0.00777
	(0.00484)	(0.0197)	(0.00381)	(0.00385)	(0.00919)
Experience	-0.00110	0.00257	-0.00487	-0.00403	0.00719
	(0.00568)	(0.0187)	(0.00453)	(0.00443)	(0.00915)
Height	0.000200	0.0112	0.00162	-0.00353	-0.00291
	(0.00367)	(0.0134)	(0.00299)	(0.00232)	(0.00458)
Kicker last match	-0.00481	-0.0150	-0.00251	0.00480	0.000358
	(0.00970)	(0.0340)	(0.00755)	(0.00744)	(0.0143)
Red card last match	-0.0307	0.0762	-0.00816	-0.0294	-0.0328
	(0.0361)	(0.103)	(0.0224)	(0.0213)	(0.0463)
Penalty last match	0.00671	0.106	-0.0259	-0.00371	-0.0779*
	(0.0244)	(0.0928)	(0.0248)	(0.0188)	(0.0370)
Matchday difference	0.000805	0.00396	-0.000270	-0.0134	0.0172
	(0.0105)	(0.0274)	(0.00807)	(0.00707)	(0.0118)
NFV^{+}	-0.0452	0.110	0.00781	0.0370	0.0715
	(0.0532)	(0.145)	(0.0599)	(0.0363)	(0.0769)
WFLV^{+}	0.0140	-0.101	0.0451	0.122**	0.0289
	(0.0558)	(0.166)	(0.0473)	(0.0373)	(0.0876)
FRVS^{+}	-0.0111	-0.0444	0.0172	0.0242	-0.0145
	(0.0598)	(0.136)	(0.0518)	(0.0395)	(0.0795)
SFV^{+}	-0.0631	-0.0849	0.0720	0.0000639	0.0575
	(0.0590)	(0.125)	(0.0396)	(0.0283)	(0.0582)
Trading^{++}	-0.0875	-0.139	-0.000605	0.00374	-0.0200
	(0.0503)	(0.127)	(0.0422)	(0.0297)	(0.0812)
Medical^{++}	-0.0404	-0.128	0.0309	0.00898	0.0679
	(0.0414)	(0.0821)	(0.0276)	(0.0269)	(0.0559)
Lawyer^{++}	-0.0540	0.0767	0.00760	0.00636	-0.0660
	(0.0729)	(0.151)	(0.0496)	(0.0532)	(0.114)
Public^{++}	-0.0185	0.119	0.0660	0.0653	0.0521
	(0.0860)	(0.186)	(0.0744)	(0.0451)	(0.130)
Engineer^{++}	0.00805	-0.414	-0.0507	-0.000775	-0.0274
	(0.0983)	(0.233)	(0.0647)	(0.0543)	(0.104)
N	1,601	1,928	1,654	1,928	1,601
Adj. R-sq	0.032	0.096	-0.014	0.018	0.058

a. Bold printed models have significant individual referee effects in the first stage.

b. Standard errors in parentheses. c. Standard errors are clustered at team level

d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.15: Referee characteristics in home matches II.

Tables 2.16 and 2.17 display the results of equation 2.4 for away teams. Again, the year of birth and experience have no significant effects on match outcome or referee decisions. Once more, if we introduce the quadratic terms of experience, we find a significant increasing effect from experience for the referee decision "not awarded red card".[46] The tendency not award such cards increases as soon as a referee surpasses seven years of experience, which is the average amount of experience in our sample. So, we do find some evidence that referees follow career concerns if we run equation 2.4 for away teams.

Once again we fail to find any significant effects from the FIFA dummy variable. However, compared to the results for home teams, FIFA referees have positive impact on our dependent variables on more occasions for away teams. Moreover, the number of awarded yellow cards decreases by 0.00911 units if the period of time between two matches directed by the same referee increases by one matchday.

Again, we examine any influence of the referees' regional associations. We see that the number of awarded yellow cards increases by 0.372 units if a referee is a member of the Norddeutscher Fußballverband (NFV) compared to members of the Nordostdeutscher Fußballverband (NOFV).

There is some evidence that a referee's profession can suitably explain match outcome and referee decisions for away teams. The goal difference increases by 0.282 or 0.227 units respectively if a referee works in the trading or medical sector, compared to the category of reference. If a referee works as a lawyer, the number of awarded red cards decreases by 0.0564 units. The number of awarded yellow cards decreases by 0.471 units if the referee on the pitch works in the public sector. Moreover, if a referee works as an engineer, the number of awarded yellow-red cards significantly decreases by 0.0674 units and the number of awarded goals decreases by 0.432 units. But there are also significant and positive effects from this profession. The number of not awarded goals increases by 0.109 units and the number of not awarded penalties increases by 0.254 units.

Again, our referee characteristics reveal only marginally significant explanations for the significant referee effects for away teams. Indeed, it seems that a referee's profession plays a more important role for match outcome and referee decisions for away teams than for home teams.

[46] For reasons of clarity these regression results are presented in tables 2.20 and 2.21 in section 2.A1.

	(1) Result	(2) **Goal difference**	(3) **Yellow card**	(4) Yellow-red card	(5) Red card
FIFA	0.0889	0.151	0.0881	-0.000597	0.00514
	(0.0977)	(0.121)	(0.0754)	(0.0108)	(0.0134)
Year of birth	-0.000315	0.00615	0.0234	-0.00160	0.00255
	(0.0124)	(0.0208)	(0.0140)	(0.00204)	(0.00430)
Experience	0.000404	-0.00436	0.00240	-0.00113	0.000621
	(0.0142)	(0.0200)	(0.0136)	(0.00228)	(0.00347)
Height	-0.00143	-0.0149	0.0221	0.00102	-0.00182
	(0.0104)	(0.0126)	(0.0123)	(0.00180)	(0.00195)
Kicker last match	0.0285	0.0312	0.0103	0.00214	0.00598
	(0.0296)	(0.0391)	(0.0396)	(0.00570)	(0.00484)
Red card last match	-0.0573	-0.133	-0.0948	0.0274	0.00506
	(0.0902)	(0.148)	(0.102)	(0.0200)	(0.0187)
Penalty last match	-0.119	-0.155	-0.125	-0.0193	-0.00198
	(0.0728)	(0.0763)	(0.0662)	(0.0154)	(0.0129)
Matchday difference	0.00657	0.0108	-0.0366	-0.00911**	-0.00706
	(0.0220)	(0.0250)	(0.0249)	(0.00316)	(0.00461)
NFV^{+}	-0.0266	-0.0779	0.372*	-0.0122	0.0317
	(0.144)	(0.209)	(0.170)	(0.0339)	(0.0283)
WFLV^{+}	0.0399	0.198	0.152	-0.0187	0.0258
	(0.136)	(0.211)	(0.168)	(0.0253)	(0.0235)
FRVS^{+}	-0.0195	0.144	-0.229	-0.0102	-0.0224
	(0.144)	(0.215)	(0.206)	(0.0308)	(0.0232)
SFV^{+}	0.00278	0.0779	-0.146	0.00864	-0.0169
	(0.0935)	(0.149)	(0.123)	(0.0275)	(0.0263)
Trading^{++}	0.170	0.282*	-0.201	-0.0247	-0.0254
	(0.0860)	(0.128)	(0.123)	(0.0216)	(0.0164)
Medical^{++}	0.166	0.227*	-0.0258	-0.00430	0.0115
	(0.0810)	(0.102)	(0.123)	(0.0205)	(0.0191)
Lawyer^{++}	0.0420	0.0523	-0.333	-0.00699	-0.0564*
	(0.174)	(0.222)	(0.177)	(0.0363)	(0.0260)
Public^{++}	-0.112	-0.0453	-0.471*	0.0283	-0.0374
	(0.140)	(0.191)	(0.191)	(0.0332)	(0.0253)
Engineer^{++}	0.186	0.443	-0.339	-0.0674*	0.00819
	(0.195)	(0.246)	(0.216)	(0.0269)	(0.0350)
N	1,928	1,928	1,928	1,928	1,928
Adj. R-sq	0.110	0.135	0.029	0.011	0.008

a. Bold printed models have significant individual referee effects in the first stage.

b. Standard errors in parentheses. c. Standard errors are clustered at team level.

d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.16: Referee characteristics in away matches I.

	(6) Not awarded red card	(7) **Goal**	(8) Not awarded goal	(9) Penalty	(10) **Not awarded penalty**
FIFA	0.0367	-0.121	0.0256	-0.0232	0.0787
	(0.0327)	(0.0829)	(0.0233)	(0.0208)	(0.0568)
Year of birth	0.00639	-0.00577	0.00170	0.00345	-0.0130
	(0.00673)	(0.0191)	(0.00384)	(0.00623)	(0.0118)
Experience	0.00639	0.00271	-0.00109	0.00176	-0.0115
	(0.00738)	(0.0146)	(0.00389)	(0.00471)	(0.00871)
Height	0.00241	0.0108	-0.00447	-0.00217	-0.00681
	(0.00339)	(0.0101)	(0.00248)	(0.00426)	(0.00607)
Kicker last match	-0.00764	-0.0153	-0.00928	0.00931	0.00229
	(0.0105)	(0.0335)	(0.00844)	(0.00748)	(0.0156)
Red card last match	0.00386	0.0703	0.00239	0.0102	-0.0292
	(0.0450)	(0.108)	(0.0315)	(0.0307)	(0.0346)
Penalty last match	0.0137	0.109	-0.0128	-0.0140	-0.0684
	(0.0299)	(0.0685)	(0.0196)	(0.0164)	(0.0379)
Matchday difference	-0.00182	0.00398	0.00756	0.00295	0.0109
	(0.00668)	(0.0179)	(0.00811)	(0.00763)	(0.0124)
NFV^{+}	-0.0595	0.102	-0.0374	-0.0330	-0.169
	(0.0527)	(0.147)	(0.0583)	(0.0444)	(0.0958)
WFLV^{+}	-0.0415	-0.0997	-0.0438	-0.0190	-0.0476
	(0.0551)	(0.150)	(0.0456)	(0.0370)	(0.115)
FRVS^{+}	-0.0297	-0.0283	-0.0294	0.00802	0.0376
	(0.0466)	(0.189)	(0.0415)	(0.0572)	(0.113)
SFV^{+}	-0.0172	-0.0770	-0.0535	-0.0589	-0.0185
	(0.0473)	(0.103)	(0.0364)	(0.0426)	(0.0916)
Trading^{++}	0.0171	-0.123	0.0514	-0.00422	0.0200
	(0.0448)	(0.126)	(0.0344)	(0.0319)	(0.0785)
Medical^{++}	0.0164	-0.135	0.0281	-0.00500	-0.0786
	(0.0455)	(0.0821)	(0.0325)	(0.0297)	(0.0823)
Lawyer^{++}	-0.0454	0.0855	0.0186	0.00306	0.206
	(0.0477)	(0.188)	(0.0545)	(0.0450)	(0.109)
Public^{++}	-0.0283	0.138	0.0142	0.0478	0.120
	(0.0802)	(0.199)	(0.0559)	(0.0476)	(0.103)
Engineer^{++}	-0.107	-0.432*	0.109*	0.0633	0.254**
	(0.0632)	(0.210)	(0.0486)	(0.0786)	(0.0824)
N	1,651	1,928	1,654	1,928	1,651
Adj. R-sq	0.015	0.097	-0.009	0.001	0.049

a. Bold printed models have significant individual referee effects in the first stage.

b. Standard errors in parentheses. c. Standard errors are clustered at team level.

d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.17: Referee characteristics in away matches II.

2.6 Conclusion

The aim of this study is to examine whether referees have a significant individual influence in German 1^{st} Bundesliga matches and how referee characteristics help to explain these individual referee effects. Recent studies, examining the length of extra time at the end of the second half, penalties or bookings, have found a referee bias. In this paper, focus lies on the final outcome (result) of a match and referee decisions like bookings, dismissals, awarded goals or penalties, and whether these variables are significantly influenced by referees. In sum, we find evidence that referees have a significant impact on these decisions while general DFB or FIFA guidelines state an impartial behavior. Yet the more crucial finding is that referees have significant effects on the result of a match. In addition, there is evidence that referees treat home and away teams in a different way. Further, we conclude that teams are subjected to referee bias in different ways, which do not solely depend on whether they are the home or away team. Thus, some teams, including FC Bayern München, Borussia Dortmund or St. Pauli, on average benefit from referee bias while others are systematically disadvantaged by referees (e.g. 1. FC Kaiserslautern, SC Freiburg).

Similarly to studies which examine a manager's influence on firm performance, we hypothesize that referees have different "referee styles" and therefore also study the correlation between the individual referee fixed effects. Using only significant individual referee effects, we find different styles for home and away teams. Thus, referees with significant individual effects for result for home teams also have significant effects on other referee fixed effects like awarded yellow and red cards, as well as awarded goals. These correlations between the different individual referee effects are less pronounced if we control for away teams.

In a next step, we use referee properties instead of referee fixed effects to examine our first stage regression. Here, we investigate whether these characteristics can explain the significant individual effects on match outcome variables and further referee decisions. Moreover, we hypothesize that referees follow career concerns. Allowing for a non-linear relationship between experience and the dependent variables, we however find only limited evidence to confirm this hypothesis. Neither does our second hypothesis on referee behavior find empirical support: There are no significant results to suggest that referees worry about their reputation because we find no significant effects of our FIFA dummy variable. By contrast, we find further explanations for the referee fixed effects looking at the referee's profession. Referees who work in the trading or medical sector as well as engineers differ significantly from those referees who work in other professional jobs. Yet this only holds true for selected referee decisions and mostly for away teams.

In the end, we have to acknowledge that observable referee characteristics are insufficient to explain our findings of significant referee effects. Thus, we have to assume that there are also unobservable characteristics which impact upon a referee's individual effect on match outcome. Meanwhile, the DFB and Deutsche Fußball Liga have resolved to reduce such individual referee effects. In time for the 2012/13 season a basic financial security for DFB referees was imple-

mented to answer (amongst other objectives) the UEFA's calls for the professionalization of referees.[47] Further, the referees are provided their own physiotherapists at the respective match venue. Therefore, future work should double-check this individual referee influence with new data and validate whether these new facilities for referees are helpful to limit the significant individual influence.[48]

[47] E.g. FIFA referees receive €40,000 and 1^{st} Bundesliga referees are paid €20,000.

[48] Other examples of suitable measures would be the implementation of a fourth official at the sideline or the equipment of referees with headsets to simplify their communication with the assistant referees.

2.A1 Additional Results

	(1) Result	(2) Goal difference	(3) Yellow card	(4) Yellow-red card	(5) Red card
FIFA	-0.141	-0.147	0.00996	-0.0188	-0.0138
	(0.101)	(0.145)	(0.115)	(0.0165)	(0.0173)
Year of birth	0.00108	-0.00553	0.0119	-0.00361	-0.00203
	(0.0207)	(0.0283)	(0.0176)	(0.00257)	(0.00305)
Experience	0.00957	0.00163	0.0282	0.00501	0.00188
	(0.0463)	(0.0565)	(0.0407)	(0.00809)	(0.00458)
Experience2	-0.000180	0.000153	-0.00189	-0.000510	-0.0000338
	(0.00184)	(0.00209)	(0.00184)	(0.000448)	(0.000228)
Height	0.00425	0.0146	0.0131	0.000447	0.000447
	(0.0123)	(0.0179)	(0.0109)	(0.00187)	(0.00153)
Kicker last match	-0.0286	-0.0305	-0.0340	-0.0122*	0.0112*
	(0.0283)	(0.0330)	(0.0390)	(0.00494)	(0.00450)
Red card last match	0.106	0.144	0.0692	0.0117	-0.00908
	(0.101)	(0.136)	(0.146)	(0.0216)	(0.0217)
Penalty last match	6 0.107	0.150	-0.151*	0.0260	0.00351
	(0.0889)	(0.121)	(0.0684)	(0.0129)	(0.0174)
Matchday difference	-0.0143	-0.0121	0.00858	-0.00304	0.00206
	(0.0323)	(0.0327)	(0.0276)	(0.00449)	(0.00388)
NFV^{+}	0.0840	0.0875	-0.0510	0.0229	-0.0113
	(0.159)	(0.195)	(0.170)	(0.0333)	(0.0268)
WFLV^{+}	-0.0424	-0.189	-0.00187	-0.00433	-0.0383
	(0.176)	(0.244)	(0.149)	(0.0297)	(0.0264)
FRVS^{+}	-0.0150	-0.171	-0.162	0.0195	-0.0584*
	(0.137)	(0.181)	(0.189)	(0.0324)	(0.0271)
SFV^{+}	-0.0138	-0.0885	-0.216	0.0160	-0.0485*
	(0.127)	(0.161)	(0.133)	(0.0222)	(0.0231)
Trading^{++}	-0.157	-0.303	-0.000340	-0.0147	-0.0309
	(0.106)	(0.170)	(0.114)	(0.0225)	(0.0291)
Medical^{++}	-0.149	-0.216	0.148	-0.0155	-0.0148
	(0.125)	(0.165)	(0.129)	(0.0240)	(0.0283)
Lawyer^{++}	-0.0461	-0.0621	0.112	-0.0203	-0.000743
	(0.182)	(0.237)	(0.150)	(0.0320)	(0.0402)
Public^{++}	0.0549	0.0193	0.0553	-0.0196	-0.0414
	(0.199)	(0.272)	(0.220)	(0.0444)	(0.0429)
Engineer^{++}	-0.294	-0.418	-0.355	-0.0545*	-0.00928
	(0.260)	(0.322)	(0.192)	(0.0261)	(0.0385)
N	1,928	1,928	1,928	1,928	1,928
Adj. R-sq.	0.100	0.133	0.039	0.027	0.004

a. Bold printed models have significant individual referee effects in the first stage.
b. Standard errors in parentheses. c. Standard errors are clustered at team level.
d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.18: Additional results on referee characteristics in home matches I.

	(6) Not awarded red card	**(7) Goal**	(8) Not awarded goal	**(9) Penalty**	**(10) Not awarded penalty**
FIFA	-0.0231	-0.119	-0.0261	0.0290	0.0636
	(0.0417)	(0.103)	(0.0309)	(0.0234)	(0.0580)
Year of birth	-0.00114	-0.00596	-0.00479	0.000804	0.00609
	(0.00448)	(0.0210)	(0.00408)	(0.00411)	(0.00902)
Experience	0.00822	0.00144	-0.0112	-0.00948	-0.00751
	(0.0121)	(0.0448)	(0.0128)	(0.00887)	(0.0207)
Experience2	-0.000508	0.0000616	0.000341	0.000296	0.000801
	(0.000741)	(0.00174)	(0.000558)	(0.000382)	(0.00116)
Height	0.000205	0.0112	0.00162	-0.00353	-0.00292
	(0.00367)	(0.0134)	(0.00298)	(0.00231)	(0.00456)
Kicker last match	-0.00510	-0.0150	-0.00225	0.00501	0.000827
	(0.00963)	(0.0340)	(0.00756)	(0.00751)	(0.0141)
Red card last match	-0.0318	0.0763	-0.00759	-0.0288	-0.0310
	(0.0362)	(0.104)	(0.0228)	(0.0211)	(0.0474)
Penalty last match	0.00692	0.106	-0.0260	-0.00370	-0.0782*
	(0.0244)	(0.0928)	(0.0249)	(0.0188)	(0.0373)
Matchday difference	0.00156	0.00388	-0.000659	-0.0138	0.0161
	(0.0108)	(0.0276)	(0.00786)	(0.00707)	(0.0116)
NFV^{+}	-0.0588	0.111	0.0172	0.0448	0.0930
	(0.0634)	(0.158)	(0.0591)	(0.0386)	(0.0657)
WFLV^{+}	0.00450	-0.0998	0.0520	0.128**	0.0440
	(0.0587)	(0.180)	(0.0486)	(0.0380)	(0.0820)
FRVS^{+}	-0.00246	-0.0454	0.0119	0.0192	-0.0282
	(0.0608)	(0.131)	(0.0530)	(0.0378)	(0.0889)
SFV^{+}	-0.0703	-0.0840	0.0772	0.00433	0.0688
	(0.0615)	(0.133)	(0.0386)	(0.0287)	(0.0504)
Trading^{++}	-0.0759	-0.141	-0.00835	-0.00293	-0.0382
	(0.0569)	(0.146)	(0.0479)	(0.0296)	(0.0822)
Medical^{++}	-0.0270	-0.129	0.0223	0.00149	0.0467
	(0.0515)	(0.0948)	(0.0328)	(0.0276)	(0.0531)
Lawyer^{++}	-0.0409	0.0751	-0.00138	-0.00145	-0.0866
	(0.0793)	(0.169)	(0.0547)	(0.0512)	(0.117)
Public^{++}	0.00545	0.116	0.0495	0.0512	0.0143
	(0.104)	(0.215)	(0.0767)	(0.0476)	(0.137)
Engineer^{++}	0.0245	-0.416	-0.0618	-0.0105	-0.0534
	(0.102)	(0.249)	(0.0710)	(0.0519)	(0.108)
N	1,601	1,928	1,654	1,928	1,601
Adj. R-sq.	0.032	0.095	-0.015	0.018	0.058

a. Bold printed models have significant individual referee effects in the first stage.
b. Standard errors in parentheses. c. Standard errors are clustered at team level.
d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.19: Additional results on referee characteristics in home matches II.

	(1) Result	**(2) Goal difference**	**(3) Yellow card**	(4) Yellow-red card	(5) Red card
FIFA	0.0988	0.153	0.0614	0.00337	0.00364
	(0.0942)	(0.118)	(0.0720)	(0.0116)	(0.0188)
Year of birth	-0.00108	0.00598	0.0254	-0.00191	0.00267
	(0.0121)	(0.0216)	(0.0153)	(0.00212)	(0.00428)
Experience	-0.00634	-0.00583	0.0206	-0.00383	0.00164
	(0.0280)	(0.0424)	(0.0347)	(0.00463)	(0.00751)
Experience2	0.000366	0.0000799	-0.000988	0.000147	-0.0000553
	(0.00157)	(0.00207)	(0.00150)	(0.000222)	(0.000389)
Height	-0.00144	-0.0149	0.0221	0.00102	-0.00182
	(0.0104)	(0.0126)	(0.0123)	(0.00180)	(0.00195)
Kicker last match	0.0288	0.0313	0.00963	0.00224	0.00594
	(0.0295)	(0.0392)	(0.0399)	(0.00575)	(0.00490)
Red card last match	-0.0564	-0.132	-0.0972	0.0277	0.00493
	(0.0889)	(0.147)	(0.103)	(0.0200)	(0.0190)
Penalty last match	-0.119	-0.155	-0.125	-0.0193	-0.00198
	(0.0729)	(0.0763)	(0.0662)	(0.0154)	(0.0129)
Matchday difference	0.00607	0.0107	-0.0352	-0.00931**	-0.00699
	(0.0220)	(0.0254)	(0.0252)	(0.00316)	(0.00451)
NFV^{+}	-0.0171	-0.0758	0.346	-0.00838	0.0302
	(0.167)	(0.225)	(0.171)	(0.0335)	(0.0313)
WFLV^{+}	0.0471	0.199	0.132	-0.0158	0.0247
	(0.139)	(0.220)	(0.164)	(0.0246)	(0.0264)
FRVS^{+}	-0.0256	0.143	-0.212	-0.0127	-0.0215
	(0.146)	(0.221)	(0.201)	(0.0316)	(0.0258)
SFV^{+}	0.00809	0.0791	-0.160	0.0108	-0.0177
	(0.102)	(0.158)	(0.125)	(0.0272)	(0.0270)
Trading^{++}	0.162	0.280	-0.179	-0.0280	-0.0242
	(0.0996)	(0.149)	(0.113)	(0.0217)	(0.0216)
Medical^{++}	0.157	0.225	-0.000662	-0.00802	0.0129
	(0.0955)	(0.116)	(0.133)	(0.0198)	(0.0230)
Lawyer^{++}	0.0323	0.0502	-0.307	-0.0109	-0.0549
	(0.184)	(0.238)	(0.178)	(0.0354)	(0.0269)
Public^{++}	-0.129	-0.0491	-0.424*	0.0213	-0.0348
	(0.163)	(0.218)	(0.173)	(0.0329)	(0.0356)
Engineer^{++}	0.174	0.440	-0.306	-0.0722*	0.0100
	(0.202)	(0.254)	(0.211)	(0.0276)	(0.0393)
N	1,928	1,928	1,928	1,928	1,928
Adj. R-sq.	0.109	0.135	0.028	0.010	0.008

a. Bold printed models have significant individual referee effects in the first stage.
b. Standard errors in parentheses. c. Standard errors are clustered at team level.
d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.20: Additional results on referee characteristics in away matches I.

	(6) Not awarded red card	(7) **Goal**	(8) Not awarded goal	(9) Penalty	(10) **Not awarded penalty**
FIFA	0.0818	-0.123	0.0101	-0.0274	0.0644
	(0.0401)	(0.0916)	(0.0289)	(0.0223)	(0.0687)
Year of birth	0.00278	-0.00565	0.00294	0.00378	-0.0119
	(0.00609)	(0.0211)	(0.00396)	(0.00646)	(0.0117)
Experience	-0.0237*	0.00375	0.00926	0.00465	-0.00193
	(0.0103)	(0.0401)	(0.0117)	(0.0101)	(0.0156)
Experience2	0.00162*	-0.0000566	-0.000558	-0.000157	-0.000516
	(0.000703)	(0.00175)	(0.000590)	(0.000441)	(0.000810)
Height	0.00241	0.0108	-0.00448	-0.00217	-0.00681
	(0.00339)	(0.0101)	(0.00252)	(0.00425)	(0.00604)
Kicker last match	-0.00715	-0.0153	-0.00968	0.00920	0.00214
	(0.0107)	(0.0334)	(0.00851)	(0.00737)	(0.0156)
Red card last match	0.00690	0.0702	0.00143	0.00988	-0.0301
	(0.0455)	(0.109)	(0.0316)	(0.0307)	(0.0352)
Penalty last match	0.0139	0.109	-0.0128	-0.0140	-0.0685
	(0.0299)	(0.0685)	(0.0198)	(0.0164)	(0.0379)
Matchday difference	-0.00335	0.00405	0.00812	0.00317	0.0113
	(0.00643)	(0.0183)	(0.00807)	(0.00780)	(0.0121)
NFV^{+}	-0.0158	0.100	-0.0521	-0.0371	-0.183
	(0.0506)	(0.151)	(0.0593)	(0.0472)	(0.0998)
WFLV^{+}	-0.00694	-0.101	-0.0551	-0.0221	-0.0586
	(0.0544)	(0.155)	(0.0494)	(0.0384)	(0.119)
FRVS^{+}	-0.0569	-0.0274	-0.0202	0.0106	0.0462
	(0.0498)	(0.193)	(0.0445)	(0.0585)	(0.112)
SFV^{+}	0.00830	-0.0778	-0.0617	-0.0611	-0.0266
	(0.0456)	(0.106)	(0.0367)	(0.0432)	(0.0928)
Trading^{++}	-0.0214	-0.122	0.0644	-0.000688	0.0323
	(0.0497)	(0.137)	(0.0393)	(0.0345)	(0.0793)
Medical^{++}	-0.0254	-0.134	0.0428	-0.00102	-0.0653
	(0.0516)	(0.0882)	(0.0397)	(0.0309)	(0.0842)
Lawyer^{++}	-0.0901	0.0870	0.0331	0.00721	0.221
	(0.0505)	(0.196)	(0.0581)	(0.0493)	(0.108)
Public^{++}	-0.108	0.140	0.0412	0.0553	0.146
	(0.0921)	(0.223)	(0.0648)	(0.0576)	(0.108)
Engineer^{++}	-0.162*	-0.431	0.128*	0.0684	0.272**
	(0.0705)	(0.210)	(0.0542)	(0.0811)	(0.0888)
N	1,651	1,928	1,654	1,928	1,651
Adj. R-sq.	0.019	0.096	-0.009	0.000	0.049

a. Bold printed models have significant individual referee effects in the first stage.
b. Standard errors in parentheses. c. Standard errors are clustered at team level.
d. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
e. + category of reference is NOFV, ++ category of reference is Other.

Table 2.21: Additional results on referee characteristics in away matches II.

Chapter 3
Are football referees really neutral or do they have prejudices?

Abstract

Recent research on referee bias in sports relies on principal-agent-theory to explain biased referee decisions. By contrast, we use a game-theoretic model that originates from a racial-discrimination setting, to examine whether football referees are biased. In equilibrium, we find that the "fail rate" of a referee must be equal for home and away teams. We test this result with data from German football using a simple non-parametric Pearson χ^2 test.

Keywords discrimination, favoritism, referees, football, Pearson χ^2 test
JEL Classification D70, J00, L83

3.1 Introduction

Recent research on referee bias in football shows several ways to estimate influence factors on referee behavior. These studies alternatively use the extra time at the end of the second half, the number of penalties or bookings as indicators of potential referee bias. Usually a parametric procedure like OLS regression or a probit model is estimated to assess which explanatory variables have a statistically significant impact on referee bias in support of home teams. Most studies in that field stress, amongst other factors, that referee bias is a reaction to social pressure from the crowd in the stadium (e.g Lucey and Power, 2004; Page and Page, 2010) or depends on the referee's experience (e.g Boyko et al., 2007; Dawson, 2012).

The theoretical framework in that research (e.g Dohmen, 2003, 2008; Rickman and Witt, 2008) often rests upon a principal-agent relationship between a referee and the national football association. Our contribution to that field of research is that instead we use a game-theoretic approach to explain referee behavior, which allows us to test our results with a simple non-parametric method. In our mixed-strategy equilibrium, a referee who is neutral and non-prejudiced treats home and away teams in the same manner. By means of an empirical test, we compare the average fraction of wrong decisions over total referee decisions ("fail rate") across groups (e.g. home and away teams), finding differences across these groups. There are two possible explanations. First, these differences are due to statistical discrimination. In other words, the variation is due to different ways of playing a match. For example, one would presume that home teams often play more offensively in their own stadium and therefore players from the away team are more often forced to stop their rivals by fouls etc. Here, a referee is more often obliged to penalize a player from the away team. The second reason could be that a referee has prejudices against particular types of teams (e.g. away, poor or favorite teams), causing him to make biased decisions and driving the significantly different fail rate.

Besides the debate on individual wrong referee decisions, a new discussion on referees has emerged during the last years. For instance, at the end of the first half of season 2013/14, Herbert Fandel, the DFB's head of referees had to admit that the referee's performances were unsatisfactory in recent matches, a conspicuous number of wrong decisions, particularly on offsides, having been made. Related to these events, more and more football players lament some referees' arrogant manner.[1] This growing dissatisfaction with referee behavior in dealing with football players is most prevalent within the so-called smaller teams. Here, the players feel more readily discriminated against by the referee. In one example, a referee allegedly refereed to a match as a "Drecks-Kick", which roughly means that he deemed the quality of the match to be very low and he was disgusted at having to referee the match.[2] These developments motivate our investigation into whether referees in football have prejudices against teams or matches. One would expect that not to be the case, given that referees are monitored and evaluated after every match and must perform consistently well in order to remain in 1^{st} Bundesliga service or

[1] Cf. `www.spiegel.de/sport/fussball/fussball-bundesliga-schiedsrichter-chef-fandel-raeumt-fehler-ein-a-940523.html`, last access: January 16th, 2014.

[2] Cf. `www.tagesspiegel.de/sport/fall-gagelmann-spieler-von-augsburg-und-hertha-klagen-ueber-respektlose-schiedsrichter/9055556.html`, last access: January 16th, 2014.

to be promoted to Fifa referee.

To the best of our knowledge, our study is the first to use a game-theoretic model which has its origin in a racial-discrimination environment. Knowles et al. (2001) develop a model that detects racial behavior by police officers searching motor vehicles. They find that in equilibrium the guilt probability of carrying drugs does not significantly differ across groups (e.g. "African-Americans" and "Whites") if police officers have no prejudices. The authors define prejudices as different costs of searching an "African-American's" as opposed to a "White" driver's vehicle. The authors then check their hypothesis using a simple non-parametric test (Pearson χ^2 test), the advantage being that the test only requires data on race and guilt probabilities.

We build on this model in a football referee setting, finding that in equilibrium the fail rate must be equal for home and away teams if the referee is unbiased. Yet the model's applications are not limited to testing for differences between home and away teams. We also distinguish according to crucial versus non-crucial matches, favorite home teams, racetrack in stadiums, well-attended versus poorly attended matches and first versus second half of the season.

To test our hypothesis, we use a data set that was provided by Impire AG. The data cover seasons 1999/00 through 2006/07 of the German 1^{st} Bundesliga.[3] We use information on referee decisions like awarded and not awarded goals, as well as penalties and not awarded red cards and whether these decisions were right, wrong or disputable.[4] Later, to build our subgroups, we use details on the relative budget of the teams, the number of spectators and whether the stadium has a racetrack around the pitch.

In the empirical section, we also apply a Pearson χ^2 test to support our null hypothesis which is that the fail rate of referee decisions does not differ across groups. We are able to reject it for several of the subgroups, most notably for the comparison between home and away teams. In particular, this is true for disputably awarded and not awarded goals as well as not awarded penalties. Moreover, if we use the total number of referee decisions, we also find significant evidence that referees treat home and away teams differently. Comparing the proportions in other subgroups, we find some empirical evidence of biased referee behavior, too. If we compare well-attended and poorly attended matches, we can also reject the null hypothesis for the referee decision on awarded goals and the total number of disputable decisions. Similar results are found for matches played in stadiums with, as opposed to without, a racetrack around the pitch. Here, we find significant p-values for the total number of disputable decisions, too and in particular for not awarded disputable goals. Further, we find one significant Pearson χ^2 test statistic each for the subgroups "second half of the season", "rich" and "local derby".

In sum, we are able to confirm earlier results that referees tend to penalize home teams differently from away teams. Further, we find evidence that referee bias is attributable to additional factors, including team budget and special types of matches.

[3] We only have data on referee decisions for that period.
[4] This classification was done by the experts from Impire.

The remainder of this paper is organized as follows: Section 3.2 describes our model on referee prejudices in football. In section 3.3, the data set is described and descriptive statistics are presented. Section 3.4 lays out the empirical strategy and section 3.5 presents our empirical findings. Section 3.6 concludes.

3.2 The Model

3.2.1 Related Literature on Prejudice

We use a model on prejudice that Knowles et al. (2001) developed to estimate whether police officers are racially biased in motor vehicles searches or rather, they are interested in motor vehicle searches for contraband and whether there are significant differences between searches of African-American motorists and White motorists. The model features mixed-strategy equilibria. In their model, police officers maximize their numbers of successful searches and the motorist's decision on whether or not to carry contraband depends on the likelihood of being searched. Further, a police search decision is made on the basis of different drivers characteristics (e.g. type of car, license number) and race.[5] Then Knowles et al. (2001) define prejudices as different costs to police officers of searching drivers of different race. The model's main result is that if two subgroups searched in equilibrium and police officers have the same costs of searching these subgroups, then the number of successful searches is equal across these subgroups - assuming that police officers are non-prejudiced. Further, the authors mention that statistical discrimination must exist as a property of the equilibrium. So if the two groups have different probabilities of carrying drugs in equilibrium, it is possible that one group is searched more often than the other. However, this is essential to assert equal guilt proportions across groups in equilibrium. The model's predictive power is not constrained to the subgroup "race". Knowles et al. (2001) also use subgroups like sex, type of car and time of day to determine whether police officers are prejudiced. Finally, they suggest the Pearson χ^2 test to differentiate between statistical discrimination and racial prejudice.

Also in a vehicle search setting, Dharmapala and Ross (2004) use the KPT (Knowles-Persico-Todd) model as a basis but model two different pay-offs for the motorists. First, they include a probability that considers that the police cannot control every motorist, which yields motorists who do not randomize their carriage of contraband. Second, they include two different levels of offense severity. Although the authors extend the KPT-model in two ways, they do use KPT's original empirical test, except that here the validity of this empirical test depends, amongst other factors on which types of equilibria exist.[6]

Anwar and Fang (2005) also build on the work of Knowles et al. (2001). They extend the model for motorists as well as for police officers. Like Dharmapala and Ross (2004), the authors also account for the possibility that police officer differ in their behavior depending on their own

[5] These characteristics are observable for the police but not for the econometrician.

[6] With this "new" model, Dharmapala and Ross (2004) find multiple equilibria, namely "Fully Randomizing Equilibria" that are equal to the equilibria of the KPT model, and "Equilibria with Randomization over Low-Level Offenses", which include the possibility that some motorists will always carry drugs.

race. Concerning motorists, Anwar and Fang (2005) allow for more information on their characteristics that influences the likelihood of carrying drugs. Further, they include an equilibrium average success rate of police officers against motorists.[7] Within their empirical method they test for monolithic police behavior as well as for racial prejudices. Compared to the test from the KPT model, they also include the average success rate in their empirical tests. In the end, they find empirical evidence that police officers do not behave monolithic and that no relative racial prejudice exists if officers have different races.

Antonovics and Knight (2009), too, build upon the KPT model on motor vehicle searches and include information on the race of the police officers and the motorists. Then, they use a probit model to test for any statistical discrimination, as opposed to so-called preference-based discrimination, in the search probabilities for motorists of different races. Statistical discrimination would be present if search decisions were independent of the police officer's race. If by contrast, motorists are more likely to be searched if their race differs from the officer's, Antonovics and Knight (2009) assume preference-based discrimination. The probit model controls for the motorists' and police officers' race, driver characteristics, and police officers' race dependent costs and a dummy variable for a mismatch between the driver's and the officer's race. Further, the authors assume of normality and random matching of officers and drivers. Thus, their approach even holds if driver characteristics are unobservable to the econometrician. In the end, they find that the likelihood of a motorist being searched is significantly higher if her race differs from the officer's.

3.2.2 Referee Prejudice in Football

We also use the KPT-model to provide a theoretical framework that describes unbiased referee behavior. In the following, this model yields a simple empirical test suitable to answering our research question with our data. The football players are all of type (c, t). $t \in \{H, A\}$, which is observable by the referee, denotes the player's team, where "H" stands for a player from the home team and "A" denotes a player from the away team. The one-dimensional variable c refers to all other match characteristics, including player characteristics that are observable by the referee but unobservable or only partially observable by the econometrician.[8] These c are used by the referee to decide on penalties, red cards, yellow and yellow-red cards, as well as goals. $F(c, H)$ and $F(c, A)$ denote the distributions of c in home and away teams.

The referees penalize football players if they violate the rules and each referee can face a player of any type (c, t). Referees minimize their total number of wrong decisions and their costs for making a decision. These costs include the costs for a wrong decision (α) but also individual costs of penalizing a player from home and away teams. These individual costs are denoted by l_t, where l_H and l_A may differ because of social pressure from the crowd or because of a referee's prejudices against certain teams. If a referee is assumed to be neutral these costs would be

[7] The authors work with the assumption that officers with prejudices will search minorities more often although the probability of success is smaller. This method has its origin in (Becker, 1957).

[8] Match characteristics like a hotly contested match with many fouls, the number of spectators or the positions both teams in the table. Player characteristics denote for example a player who has attracted attention for nagging. That is, we assume that c is always positive.

the same.[9] The costs of a wrong decision (α) are equal for all referees.[10] This assumption is based on the fact that in professional football, referees and their assistants are monitored and evaluated by official referee observers (match assessors) and these performance evaluations are used to decide whether a referee stays in the Bundesliga, is relegated to a lower league or is promoted to FIFA referee.

Football players consider the probability of being penalized in deciding whether or not to cheat (e.g. a dive or violent conduct). If they do not cheat and are not penalized, their pay-off is zero.

If they cheat and they are penalized, their pay-off is $-j(c,t)$, whereas if they are not penalized, their pay-off is $v(c,t)$. The costs of being detected and penalized (e.g. a red card for a dive and the loss of reputation) are denoted by $j(c,t)$, $v(c,t)$ is the utility gained from cheating (e.g. the chance of scoring from an awarded penalty after a dive).

Further, we include private information on the football players. We assume that football players have different "moral costs" of being detected as a "cheat". These costs are denoted by m^i and they are random from a referee's viewpoint. We denote with $\gamma(c,t)$ the probability that a referee penalizes a football player of type (c,t). A player's expected payoff from cheating is:

$$\gamma(c,t)(-j(c,t)) + (1-\gamma(c,t))v(c,t) - m^i \geq 0. \tag{3.1}$$

If equation 3.1 is greater than zero, the player decides to cheat (e.g. dive). Equating the left hand side of 3.1 to zero, we find a threshold value $\overline{m}$, such that a player cheats iff $m^i < \overline{m}$. Thus, our threshold value is denoted by

$$m^i \leq -\gamma(c,t)j(c,t) + v(c,t) - \gamma(c,t)v(c,t) := \overline{m}(\gamma(c,t)).$$

The derivatives for $\overline{m}(\gamma(c,t))$ are

$$\frac{\partial \overline{m}\gamma(c,t)}{\partial \gamma(c,t)} = \overline{m'}(\gamma(c,t)) = -j(c,t) - v(c,t) < 0$$

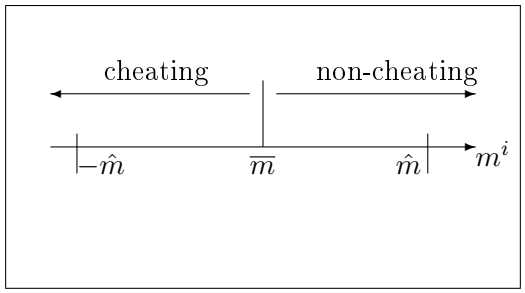
and

$$\frac{\partial^2 \overline{m}\gamma(c,t)}{\partial \gamma^2(c,t)} = \overline{m''}(\gamma(c,t)) = 0.$$

These results are necessary for computing the minimum costs of a referee decision in the next step.

[9] Like KPT, we assume that $l_H < 1$ and $l_A < 1$ because if a referee had costs of 1 for one or both teams, these teams would always cheat and that would be an uninteresting case (Dharmapala and Ross, 2004).
[10] For simplicity, we suppose that $\alpha = 0$ but the results also hold if we consider $0 < \alpha < 1$.

Further, m^i is uniformly distributed and lies in the interval [-$\widehat{m}$, $\widehat{m}$].



That is, we denote the distribution off m^i by $F(m^i) = \frac{1}{2} + \frac{1}{2\hat{m}}m^i$.[11]

The threshold $\overline{m}(\gamma(c,t))$ determines the fraction of football player of class (c,t) who cheat. If a referee detects these players, he makes the right decisions and behaves as expected.

Referees choose the probability $\gamma(c,t)$ of penalizing each football player of type (c,t). Therefore the referee minimizes his fail rate and costs of a wrong decision plus his individual costs of penalizing a player from team t at $F(m^i = \overline{m}(\gamma(c,t)))$:[12]

$$min_{\gamma(c,H),\gamma(c,A)} \sum_{t=H,A} \int [(1 - F(\overline{m}(\gamma(c,t)))) + l_t]\gamma(c,t)f(c|t)dc. \tag{3.2}$$

Optimization yields the following first order condition:

$$-f(m^i)m'(\gamma(c,t))\gamma(c,t) + (1 - F(\overline{m}(\gamma(c,t)))) + l_t = 0.$$

Solving for $\gamma(c,t)$ under the sufficient conditions $v(c,t) \geq (\frac{1}{2}+l_t)2\hat{m}$ for $\gamma(c,t) > 0$ and $-\frac{v(c,t)}{2\hat{m}} \leq \frac{1}{2} + l_t + \frac{j(c,t)}{\hat{m}}$ for $\gamma(c,t) \leq 1$, we find an optimal penalizing probability for referees:

$$\gamma^*(c,t) = (\frac{v(c,t)}{2\hat{m}} - \frac{1}{2} - l_t)\frac{\hat{m}}{j(c,t) + v(c,t)}. \tag{3.3}$$

Computing the partial derivative of $\gamma^*(c,t)$ with respect to l_t, we get

$$\frac{\partial\gamma^*(c,t)}{\partial l_t} = -\frac{\hat{m}}{j(c,t) + v(c,t)} < 0. \tag{3.4}$$

So if $l_H > l_A$, we find $\gamma(c,H) < \gamma(c,A)$. This means that a referee with higher costs of penalizing players from the home team exhibits a lower probability of sentencing these players.

[11] The partial derivative of $F(m^i)$ is denoted by $f(m^i) = \frac{1}{2\hat{m}}$.

[12] Remember that the fail rate is the average fraction of wrong decisions over total referee decisions. That also include the case of a referee failing to penalize a blamable player.

Next, we introduce two definitions of biased referees. First, a referee is defined as biased if he has preferences for wrong decisions pertaining to special types of teams (e.g home, away). These preferences are modeled as different individual costs for a decision against one of the two teams:

Definition 1: A referee is biased if $l_H \neq l_A$.

Second, we define statistical discrimination as the case when $\gamma(c, H) \neq \gamma(c, A)$ while $l_H = l_A$. For example, there are different styles of playing (e.g. home teams play more offensively) and for this reason it is possible that referees have to make more or fewer decisions against one of the two teams:

Definition 2: Assume $l_H = l_A$. Then an outcome exhibits statistical discrimination if $\gamma(c, H) \neq \gamma(c, A)$.

Referring to (Knowles et al., 2001), we assume that neutral referees respond in the same way to a cheating player in both subgroups. This implies that the fail rate should not significantly differ across these subgroups and so we find in equilibrium that

$$l_t = f(m^i)m'(\gamma(c,t))\gamma(c,t) + F(\overline{m}(\gamma(c,t))) - 1. \tag{3.5}$$

We denote the right hand side of equation (3.5) as $\Delta(\overline{m}(\gamma(c,t)))$. Then, it follows that

$$\Delta(\overline{m}(\gamma(c,H))) = l = \Delta(\overline{m}(\gamma(c,A))) \tag{3.6}$$

if $l_H = l_A = l$, meaning that referees are unprejudiced. Again, this does not imply that $\gamma^*(c, H) = \gamma^*(c, A)$. The equilibrium probability of being penalized may differ between home and away teams, for example due to different playing styles. Suppose that home teams play more offensively. Then away teams may be tempted to commit more fouls, increasing their probability of receiving yellow cards. So, different $\gamma(c,t)$ are driven by the observable match-characteristics c. In other words, the players of both teams are in the same match with the same match circumstances (stadium, audience, weather,...) and the referees have the same sets of information on this match for both groups. Although one team is penalized more often, we assume that the fail rate should be ultimately equal across the two subgroups since an unbiased referee makes no difference between players of the two teams.

As in Knowles et al. (2001), equation 3.6 yields a test for prejudice that is applicable even if we have no information on c and on γ^*. All we need is data on each team's frequency of being wrongly penalized conditional on the total number of sanctions.

$$D(t) = \int \Delta(\overline{m}\gamma(c,t))\frac{\gamma(c,t)}{\int \gamma(s,t)f(s|t)ds}dc. \tag{3.7}$$

Using 3.6 to substitute for $\Delta(\overline{m}(\gamma(c,t)))$, we get

$$D(H) = l = D(A),$$

which is the prediction that we test with our data in section 3.5.

An advantage of this model is that t is not limited comparing the "fail rates" between home and away teams. Additionally, we build subgroups for successful and unsuccessful teams, favorites and underdogs, crucial and non-crucial matches, local derbies and non-local derbies, poor and rich teams, as well as well-attended and poorly attended matches. In section 3.5, we test equation 3.6 with a non-parametric test. But prior to discussing the results, we describe our data and present descriptive statistics for our subgroups and referee decisions.

3.3 Data

The data, for our analyzes of referee prejudices were mostly collected from Impire AG.[13] The company specializes in statistics on football matches and supplies the data for TV broadcasts and Bundesliga football teams. In principle, the data are publicly available as it would be possible to collect them "per hand" from different football websites.[14] From Impire AG, we use data on referee decisions like awarded and not awarded penalties, as well as goals and not awarded red cards, knowing in each case whether these decisions are right, wrong or disputable. Further, we have data on the relative budget of the teams, the number of spectators and the presence of a racetrack around the pitch.[15] This set of information allows us to build the additional subgroups mentioned earlier. Table 3.1 illustrates the distribution of the subgroups in our sample. The distribution for home and away teams is excluded because every team plays at home and away the same number of times in each season.

In total, we observe 2,448 matches. Of these 1,292 matches are categorized as crucial in that at least one team fights against relegation, for the championship or for qualification for an international competition. Further, we identify 176 local derbies.[16] Despite its small size, this subgroup is interesting because we would assume that these matches carry a special atmosphere for spectators, players, coaches, managers and even referees. Another subgroup is constituted by matches with a high attendance in the stadium. We categorize a match as "well-attended" if the attendance-to-capacity ratio is higher than the median of 89.5%, which yields 1,217 well-attended matches. Further differentiation is made between home teams known as the "favorite" as opposed to the "underdog". We identify an "underdog" as a team that has won fewer points in the last four matches than its opponent. Thus, we have a total of 1,363 matches with a favorite home team and 1,085 matches in which we call the home team the "underdog". Moreover, we build a subgroup regarding the teams' relative budget in a season. We find 833 matches with a

[13] Cf. `www.impire.de`.
[14] One of such website is maintained by German football magazine Kicker (`www.kicker.de`). Other possible data sources are `www.bundesliga.de` or `www.wahretabelle.de`.
[15] These data were collected from Kicker. Information on budgets was collected from several football magazines (e.g. special issues) as the teams are not obliged to report their financial data in Germany.
[16] This classification is based on a list of famous German local derbies by Mechtel et al. (2011).

team that is poor in the sense that their relative budget is smaller than the median of 89.8%. Further, we divide the matches according to the presence of a racetrack in the stadium. Thus, we have 1,802 matches played in a stadium without a racetrack and 646 matches played in a stadium with a racetrack around the pitch.

Subgroup	Absolute	%
No crucial match	1,151	47.02
Crucial match	1,297	52.98
Non-local derby	2,272	92.81
Local derby	176	7.19
Well-attended matches	1,231	50.29
Poorly attended matches	1,217	49.71
Favorite	1,363	55.68
Underdog	1,085	44.32
Poor	833	34.03
Rich	1,615	65.97
No racetrack	1,802	73.61
Racetrack	646	26.39

Table 3.1: Distribution of subgroups (excl. home/away).

In our empirical analyzes, we also distinguish between matches played in the first versus the second half of the season. Further, we build a group of matches played in the last two matchdays of a season because these are really relevant matches that are even held simultaneously.

Tables 3.2 through 3.5 show the average fail rate across our subgroups. Table 3.2 only focuses on the different fractions for home and away teams since this is the group where we are most interested in. Here, we find variation in the average ratio of referee decisions for home versus away teams for a large number of decisions. Firstly, the average ratio for not awarded disputable goals differs by about 3.49 percentage points between home and away teams. Similar differences are found for awarded and not awarded penalties and the total number of disputable decisions. Even greater differences are found for awarded and not awarded disputable penalties, as well as for not awarded red cards. The number of awarded disputable penalties differs by about 10.34 percentage points between home and away teams. The average ratio for not awarded disputable penalties varies by about 7.28 percentage points. Lastly, we find a disparity of about 4.88 percentage points in the average ratio for not awarded red cards.

Referee Decisions	Home	Away
Goals	2.83 (13.43)	2.01 (11.45)
Not aw. goals	23.16 (41.71)	27.01 (44.37)
Disp. goals	3.84 (15.65)	2.36 (12.78)
Not aw. disp. goals	11.58 (31.32)	20.83 (40.18)
Penalties	8.58 (27.48)	8.0 (27.19)
Not aw. penalties	19.83 (37.66)	26.12 (42.46)
Disp. penalties	23.88 (42.1)	17.75 (38.14)
Not aw. disp. penalties	34.64 (44.88)	37.93 (46.8)
Not aw. red card	35.09 (46.46)	40.3 (48.17)
Disp. not aw. red card	58.15 (48.03)	50.68 (48.71)
Total decisions	8.26 (19.71)	9.93 (24.16)
Disp. total decisions	12.63 (24.0)	12.13 (25.89)

Table 3.2: Average rate (in %) of wrong referee decisions for home/away (standard deviation in parentheses).

Table 3.3 exhibits the average ratios for referee decisions on goals across the other subgroups. Referring to awarded and disputably awarded goals, we do not find large differences in the average ratios across the subgroups. By contrast, as the third column of table 3.3 shows, there are huge differences for not awarded goals across the subgroups. There is a difference of 11.07 percentage points between local and non-local derbies. The subgroups "well-attended matches", "favorite" and "poor" display differences of about five percentage points. Finally comparing the average rates of referee decisions between the last two matchdays and all other matchdays, we find a difference of about 6.33 percentage points.

Subgroups	Goals	Not awarded goals	Disputable goals	Not awarded disputable goals
No crucial match	2.18 (9.76)	26.96 (44.08)	2.98 (11.39)	18.43 (37.96)
Crucial match	2.55 (9.2)	23.97 (42.08)	3.05 (10.33)	13.54 (33.0)
Non-local derby	2.33 (9.48)	26.07 (43.43)	3.12 (10.95)	16.21 (35.9)
Local derby	2.85 (10.03)	15.0 (35.11)	1.7 (9.21)	10.0 (27.54)
Poorly attended matches	1.95 (8.52)	27.64 (44.13)	2.81 (10.38)	15.16 (35.08)
Well-attended matches	2.79 (10.42)	22.89 (41.71)	3.23 (11.29)	16.52 (35.88)
Favorite	2.53 (9.99)	27.56 (44.02)	2.9 (10.61)	14.87 (34.3)
Underdog	2.17 (8.89)	22.6 (41.63)	3.17 (11.13)	16.99 (36.85)
Poor	2.27 (9.18)	28.67 (44.63)	2.99 (9.97)	18.33 (37.73)
Rich	2.42 (9.69)	23.79 (42.2)	3.03 (11.26)	14.62 (34.29)
No racetrack	2.28 (9.16)	25.05 (42.99)	2.88 (10.64)	14.48 (34.64)
Racetrack	2.61 (10.46)	26.27 (43.22)	3.4 (11.39)	19.77 (37.55)
First half of the season	2.04 (9.41)	26.0 (43.58)	2.98 (11.25)	17.7 (37.09)
Second half of the season	2.7 (9.63)	24.76 (42.54)	3.06 (10.42)	14.06 (33.81)
No last two matchdays	2.41 (9.66)	24.92 (42.71)	3.0 (10.71)	15.98 (35.59)
Last two matchdays	1.81 (7.01)	31.25 (47.09)	3.29 (12.73)	13.54 (33.72)

Table 3.3: Average rate (in %) of wrong decisions on goals (standard deviation in parentheses).

Concerning referee decisions on awarded and not awarded penalties, we find differences in their average ratio, too. There is a difference of about three percentage points for the average ratio of an awarded penalty for each of the subgroups "well-attended", "favorite" and "racetrack". Further, the average rate for an awarded penalty varies by about six percentage points between local and non-local derbies, as well as for matches played during the last two matchdays compared to all other matches of the season. The third column of table 3.4 indicates the differences in the average ratio of not awarded penalties for the subgroups. Again, we find a large difference for well-attended matches and matches from the last two matchdays. Comparing the average ratio of awarded disputable penalties, we find that this ratio varies by about 5.83 percentage points between local and non-local derbies. The subgroup for matches played on either of the last two matchdays shows a difference of about 9.72 percentage points in the average ratios. The last column of table 3.4 exhibits the average ratios for not awarded penalties. Again, there is a large difference between local and non-local derbies. Moreover, we find that the average ratio differs by 4.17 percentage points between poor and rich teams.

Subgroups	Penalties	Not awarded penalties	Disputable penalties	Not awarded disputable penalties
No crucial match	6.9 (24.84)	21.66 (37.98)	24.07 (42.05)	37.78 (44.72)
Crucial match	9.01 (27.64)	23.01 (38.63)	20.55 (39.21)	35.4 (44.07)
Non-local derby	7.51 (25.55)	22.75 (38.54)	21.72 (40.2)	35.9 (44.27)
Local derby	13.27 (33.5)	18.08 (35.59)	27.55 (44.56)	43.57 (45.11)
Poorly attended matches	9.46 (28.56)	25.34 (40.44)	20.69 (39.59)	36.12 (44.79)
Well-attended matches	6.56 (23.89)	19.92 (36.32)	23.76 (41.58)	36.81 (44.04)
Favorite	9.37 (28.39)	23.66 (39.06)	22.77 (41.17)	36.01 (44.15)
Underdog	6.2 (23.29)	20.81 (37.36)	21.49 (39.88)	37.11 (44.67)
Poor	7.57 (26.0)	24.08 (38.87)	19.2 (38.62)	33.83 (43.0)
Rich	8.22 (26.54)	21.43 (38.01)	23.7 (41.49)	38.0 (45.07)
No racetrack	8.97 (27.75)	22.17 (38.18)	21.79 (40.43)	37.03 (44.56)
Racetrack	5.59 (22.36)	22.99 (38.77)	23.29 (41.1)	34.99 (43.85)
First half of the season	7.55 (25.78)	21.46 (37.8)	23.74 (41.77)	38.13 (44.9)
Second half of the season	8.45 (26.92)	23.26 (38.82)	20.75 (39.43)	34.95 (43.83)
No last two matchdays	8.4 (26.9)	22.71 (38.45)	22.88 (40.96)	36.28 (44.29)
Last two matchdays	2.63 (16.22)	17.47 (36.14)	13.16 (34.26)	39.92 (45.75)

Table 3.4: Average rate (in %) of wrong decisions on penalties (standard deviation in parentheses).

Finally, table 3.5 shows the average rate for not awarded red cards and the total number of referee decisions. As the last two columns indicate, we fail to find large differences comparing the total number of referee decisions across the subgroups. However, there does exist a substantial difference for not awarded and disputable not awarded red cards. Especially for the subgroups "local derby", "favorite" and "last two matchdays", we find differences in the average fraction of these two referee decisions. The average ratio differs between local and non-local derbies by about twelve and eleven percentage points, respectively. Similarly, this is true for the average ratio within the subgroup for the last two matchdays. Here the averages differ by 22.69 and 24 percentage points, respectively. Comparing favorite home teams with "underdogs", the average

ratios differ by about 4 and 3.48 percentage points.

Subgroups	Not awarded red card	Disputable not awarded red card	Total decisions	Disputable total decisions
No crucial match	37.95 (47.1)	55.09 (47.16)	8.05 (16.23)	12.1 (18.98)
Crucial match	39.85 (46.68)	51.96 (47.73)	9.44 (16.46)	12.35 (18.59)
Non-local derby	40.02 (46.92)	52.39 (47.41)	8.83 (16.37)	12.12 (18.79)
Local derby	28.05 (44.79)	63.41 (47.47)	8.17 (16.35)	13.66 (18.42)
Poorly attended matches	39.45 (47.23)	54.59 (47.83)	8.72 (16.66)	11.46 (18.92)
Well-attended matches	38.73 (46.58)	52.29 (47.24)	8.84 (16.06)	13.02 (18.59)
Favorite	40.79 (47.3)	51.78 (47.71)	9.41 (17.18)	12.17 (19.33)
Underdog	36.75 (46.19)	55.26 (47.18)	7.98 (15.24)	12.31 (18.04)
Poor	39.09 (47.43)	53.03 (48.02)	9.33 (16.53)	12.7 (19.38)
Rich	39.02 (46.53)	53.43 (47.21)	8.5 (16.27)	12.0 (18.45)
No racetrack	37.95 (46.91)	54.97 (47.71)	8.82 (16.64)	12.58 (19.25)
Racetrack	42.33 (46.56)	48.2 (46.54)	8.68 (15.58)	11.27 (17.32)
First half of the season	38.81 (47.27)	54.15 (47.27)	7.88 (15.85)	12.13 (19.08)
Second half of the season	39.23 (46.54)	52.59 (47.69)	9.68 (16.82)	12.34 (18.46)
No last two matchdays	40.08 (46.95)	52.19 (47.46)	8.92 (16.47)	12.24 (18.72)
Last two matchdays	17.39 (38.76)	76.09 (42.29)	6.57 (14.47)	12.16 (19.54)

Table 3.5: Average rate (in %) of wrong decisions on not awarded red cards and total decisions (standard deviation in parentheses).

In the next section, we aim to establish whether the differences in the average ratios of referee decisions across the subgroups are statistically significant to find evidence that referees are prejudiced (biased). Failure to find statistical significance would imply statistical discrimination, which is a property of the equilibrium of our model, meaning that the differences in the average ratios are due to match characteristics or to different playing styles. Therefore, we define an outcome-based test in section 3.4 and present its results in section 3.5.

3.4 Empirical Strategy

The test for referee bias compares fail rates across groups with different observed match characteristics only observable to the referee. Using the model of section 3.2.2, we apply a very strong assumption. Independently of the set of characteristics, the fail rate should be the same across groups. To test our hypothesis we only need the posterior frequency of wrong decisions, conditional on the total number of referee decisions across the subgroups (Knowles et al., 2001).

A standard procedure to test this hypothesis would be to estimate a logit- or a probit-model. However, as mentioned earlier, the parametric method requires more information, especially data about all incidents during a match. A much simpler way to test our hypothesis is to conduct a non-parametric Pearson χ^2 test. It compares the ratio of right referee decisions "within conditioning cells against the ratio that would be expected under the null hypothesis of no association between rightly penalized and the conditioning characteristics"(Knowles et al., 2001, p.217). The test statistic for the hypothesis of no association between fail rate and team

is given by

$$\sum_{t \in T} \frac{(\hat{p}_t - \hat{p})^2}{\hat{p}_t} \sim \chi^2(T-1)$$

where T is the cardinality of the set of team categories, $\hat{p}_t$ and $\hat{p}$ are conditional and unconditional estimated ratios of referee decisions, respectively.

Remember that using a non-parametric test for our analyzes has several advantages. First, we need no further information on the match characteristics that are used by the referee to penalize or not to penalize a player. Second, non-parametric tests do not require any assumptions on the distribution, e.g. normal distribution, of our "dependent" variable.

In the following, we execute the Pearson χ^2 test for all referee decisions in our sample. The results for our subgroups, especially for the group "home/away", are presented in the next section in tables 3.6 through 3.8.

3.5 Results

In section 3.3, we found large differences between average ratios of referee decisions across our subgroups. Now we apply the Pearson χ^2 test, to determine whether these differences are statistically significant or rather due to statistical discrimination.

Table 3.6 exhibits the results for the referee decisions on goals. For the group "home/away" we find a statistically significant difference for disputably awarded and not awarded goals. Thus, we can reject the null-hypothesis of equal fail rates for home and away teams. We find yet two other subgroups where the null hypothesis can be rejected. First, the test statistic is significant for the subgroup "well-attended", implying that the number of spectators might yet have an impact on that referee decision. Second, a similar result is found for matches in the first versus the second half of the season. Referring the referee decision "not awarded disputable goals", we find a significant p-value for our subgroup "racetrack". Thus the fail rate differs significantly between stadiums with and without a racetrack around the pitch.

Subgroups	Goals	Not awarded goals	Disputable goals	Not awarded disputable goals
Home/away	12.46 (0.053)	3.46 (0.325)	**27.71 (0.026)**	**2.51 (0.049)**
No crucial/crucial	17.7 (0.059)	1.18 (0.759)	9.76 (0.370)	5.52 (0.137)
Non-local derby/local derby	8.23 (0.606)	2.56 (0.464)	11.84 (0.223)	3.69 (0.297)
Poorly attend./well-attend.	**20.08 (0.029)**	2.53 (0.469)	5.81 (0.759)	1.03 (0.795)
Favorite/underdog	7.56 (0.672)	3.32 (0.345)	6.46 (0.694)	1.4 (0.705)
Poor/rich	3.79 (0.956)	2.87 (0.412)	11.08 (0.270)	2.94 (0.401)
No racetrack/racetrack	13.35 (0.205)	3.01 (0.390)	7.53 (0.582)	**12.13 (0.007)**
First/second half of the season	**20.25 (0.027)**	1.46 (0.691)	15.65 (0.074)	1.84 (0.606)
No/last two matchdays	8.34 (0.595)	1.45 (0.695)	7.95 (0.540)	4.28 (0.233)

Table 3.6: Results from Pearson χ^2 tests for referee decisions on goals (p-values in parentheses).

Table 3.7 displays the results for awarded and not awarded penalties. Again, the second row of this table shows the estimated Pearson χ^2 test statistic for home and away teams. We find a significant p-value for not awarded penalties. The decision on not awarded penalties is associated with two more significant differences. As the third column of table 3.7 shows, we also must reject our null-hypothesis that the fail rate for not awarded penalties is equal across our subgroups "local derby" and "rich".

Subgroups	Penalties	Not awarded penalties	Disputable penalties	Not awarded disputable penalties
Home/away	2.51 (0.285)	**12.67 (0.049)**	4.64 (0.099)	9.22 (0.162)
No crucial/crucial	1.98 (0.577)	8.87 (0.262)	3.34 (0.342)	9.22 (0.324)
Non-local derby/local derby	2.34 (0.504)	**14.31 (0.046)**	1.4 (0.706)	6.62 (0.578)
Poorly attend./well-attend.	2.83 (0.418)	9.53 (0.217)	1.96 (0.580)	8.86 (0.354)
Favorite/underdog	2.69 (0.443)	7.51 (0.378)	2.21 (0.531)	6.29 (0.614)
Poor/rich	0.96 (0.810)	**15.13 (0.034)**	2.11 (0.550)	9.77 (0.282)
No racetrack/racetrack	2.18 (0.535)	3.76 (0.807)	1.45 (0.695)	3.63 (0.889)
First/second half of the season	1.13 (0.770)	7.06 (0.422)	2.05 (0.561)	7.14 (0.522)
No/last two matchdays	2.04 (0.564)	2.62 (0.918)	3.08 (0.379)	3.37 (0.909)

Table 3.7: Results from Pearson χ^2 tests for referee decisions on penalties (p-values in parentheses).

The results of the Pearson χ^2 tests for not awarded red cards and total referee decisions are presented in table 3.8. Although we found differences in the average ratios for not awarded red cards in section 3.3, these distinctions are not significant across our subgroups. Yet if we use the fraction of total wrong referee decisions on total referee decisions, we find significant p-values for our subgroup "home/away". The same is true for the total number of disputable referee decisions. Again referring the total number of disputable decisions, we also find significant p-values for two other subgroups, well- versus poorly attended matches and matches that are played in stadiums with, as opposed to a racetrack. There is also a barely significant Pearson χ^2 test statistic for our subgroup "crucial" with respect to the total number of disputable referee decisions.

Subgroups	Not awarded red card	Disputable not awarded red card	Total decisions	Disputable total decisions
Home/away	6.99 (0.136)	5.24 (0.264)	**46.12 (0.000)**	**61.03 (0.000)**
Not crucial/crucial	5.19 (0.520)	5.41 (0.368)	25.13 (0.291)	**37.72** (0.049)
Non-local derby/local derby	4.26 (0.642)	3.64 (0.603)	21.41 (0.495)	29.75 (0.234)
Poorly attend./well-attend.	2.55 (0.863)	6.35 (0.274)	27.16 (0.205)	**46.12 (0.006)**
Favorite/underdog	4.01 (0.675)	2.61 (0.760)	25.51 (0.273)	28.93 (0.267)
Poor/rich	4.52 (0.607)	2.31 (0.804)	20.82 (0.532)	23.18 (0.567)
No racetrack/racetrack	8.41 (0.210)	7.16 (0.209)	20.83 (0.531)	**40.75 (0.024)**
First/second half of the season	4.6 (0.596)	4.68 (0.456)	26.8 (0.219)	31.25 (0.181)
No/last two match days	6.94 (0.327)	6.54 (0.257)	15.28 (0.850)	20.59 (0.715)

Table 3.8: Results from Pearson χ^2 tests for referee decisions on not awarded red cards and total decisions (p-values in parentheses).

Altogether, we find empirical evidence that referees treat home and away teams in different ways. That is true for disputably awarded and not awarded goals as well as not awarded penalties. Moreover, looking at the total number of referee decisions, we also find significant results for our subgroup "home/away". Comparing the ratios in other subgroups, we find some empirical evidence for divergent referee behavior, too. The subgroup for well-attended matches has two significant p-values, leading us to reject the null-hypothesis concerning the referee decision on awarded goals and the total number of disputable decisions. Similar results are found for our subgroup "racetrack". Here we reject the null hypothesis both for the total number of disputable decisions and for not awarded disputable goals. Further, we find one significant p-value for each of the subgroups "second half of the season", "rich" and "local derby". All other differences in the average ratios that are found in section 3.3 are not statistically significant and therefore attributable to statistical discrimination.

3.6 Conclusion

This paper has analyzed whether referees have prejudices regarding home and away teams in German 1^{st} Bundesliga matches. We use a game-theoretic model that is usually applied in a racial-discrimination setting. From the equilibrium of our model, we develop the hypothesis that if referees are neutral, the fraction of wrong referee decisions over total referee decisions (fail rate) will not differ significantly across home and away teams and other subgroups like "crucial", "local derby", "well-attended", "favorite", "rich" and "racetrack".

Yet comparing the average ratios of wrong referee decisions, we find substantial differences across all our subgroups. Subsequently we use a simple Pearson χ^2 test to establish whether these differences are statistically significant and thus constitute evidence of prejudiced referees.

The test yields significant results for referee decisions on goals, penalties and the total number of referee decisions. Especially for our subgroup "home/away", we find various significant p-values, which leads to the conclusion that referees judge home and away teams differently.

Although we do not find too many significant differences in the fail rates across our subgroups, the evidence of prejudice that we do have is remarkable, given that referees are expected to decide neutrally and to treat all teams equally.

In accordance with studies on referee bias (e.g Dohmen, 2003; Boyko et al., 2007; Page and Page, 2010), we find evidence that referees may be influenced by social pressure from the crowd, seeing that we also find significant p-values for our subgroups "well-attended" and "racetrack". Moreover, there is some statistical evidence of different referee behavior between poor and rich teams, matches in the first and second half of the season as well as local and non-local derbies.

All other differences in the mean ratios of wrong referee decisions must be due to statistical discrimination and can be explained by different (aggressive/defensive) ways of playing.

Chapter 4

The Impact of Intermediate Information on Effort Provision in Soccer[1]

Abstract

Tournament theory draws numerous conclusions for effort provision with regard to information prior to and during contests. Reduced effort is expected when the contestants are heterogeneous ex ante or intermediate information is available as these factors might indicate that the outcome of a contest is already certain. This paper applies detailed within-tournament information on intermediate score and the effort of the contestants to empirically test these assumptions. We use running data from substituted soccer players of the German Bundesliga and find only weak evidence for the negative effect of ex ante heterogeneity on effort while intermediate information measured by the goal difference at the time of the substitution significantly affects effort. Players provide highest effort when their team is leading by one goal and reduce effort when the team is trailing behind. This behavior can be explained by prospect theory. Players value potential losses higher than potential gains and adjust effort accordingly. Once the game appears to be decided players reduce effort independent of which team is in the lead.

Keywords tournaments, incentive effects, intermediate information, heterogeneity, effort
JEL Classification J00, L83

[1] This chapter is co-authored by Christian Deutscher and Sandra Hentschel.

4.1 Introduction

Rank-order tournaments are a popular field of research in labor economics and sports. This popularity follows from two facts: on the one side tournaments are part of our everyday lives and on the other side tournament theory provides several well formulated and empirically - or experimentally - testable hypotheses. Theoretical considerations circle around the seminal work by Lazear and Rosen (1981) who show that under certain conditions rank order tournaments are efficient in inducing effort by workers. However, disparity in ability or the availability of intermediate information on the performance or relative rank of contestants can reduce incentive effects of tournaments (McLaughlin, 1988). In a two player tournament low chances to win presumably result in reduced effort by the less capable contestant in order to save effort costs. The more capable competitor anticipates this and decreases effort as well. Accordingly, in asymmetric contests incentive effects are small. Intermediate information is expected to result in similar effects. Once significant information on relative performance becomes available both the leading and trailing contestant reduce effort - even if the contestants had been homogeneous ex ante.

An extensive literature analyzing the effect of heterogeneity on effort has emerged in the last decades. Most of these studies confirm the negative effect of heterogeneity as proposed by theory.[2] Concerning the effect of intermediate information on effort less evidence is provided by the existing literature as empirical studies on this topic are rare at best (Genakos and Pagliero, 2012; Casas-Arce and Martínez-Jerez, 2009). Experimental studies on this subject often investigate under which circumstances it is efficient to reveal interim results though only few studies provide experimental or empirical evidence of their assumptions.[3] Despite the extensive literature analyzing tournament designs and their incentive effects there are still some open questions - especially with respect to incentive effects of dynamic contests. Although it is rudimental to know how agents respond to relative performance in previous stages of the competition (Genakos and Pagliero, 2012, p.783) evidence concerning the effect of intermediate information on effort in dynamic tournaments is rare. Most of the existing studies focus on determinants of effort that are known prior to a contest while only few explicitly consider within-tournament dynamics (e.g. Genakos and Pagliero, 2012; Lynch, 2005; Berger and Nieken, 2014). Furthermore empirical studies investigating effort effects are often faced with the problem of how to measure effort respectively how to separate the incentive effects of a tournament from the ability effects (e.g. Sunde, 2009; Berger and Nieken, 2014; Wicker et al., 2013). Hence there is little empirical evidence on the impact of interim results on effort of contestants.

This study attempts to close this gap by using detailed running data of professional kickers playing in the German Bundesliga and extensive within-game information. Detailed game level statistics on the running distance and the number of high intensity runs and sprints became available recently for each player fielded in a Bundesliga match and are used as our proxies for effort. Focusing on effort provided by individual players who were substituted in during the course of a match enables us to disentangle incentive effects due to intermediate results from

[2] Cf. section 4.2.1.
[3] See e.g. Aoyagi (2010); Ederer (2010); Gershkov and Perry (2009).

further aspects influencing effort, e.g. ex ante heterogeneity of the two competing teams or the "intensity" of the match prior to the substitution. Thus this study adds new insights into the effect of interim results and ex ante heterogeneity.

Our results indicate strong incentive effects of intermediate results which are measured by the score of the match at the time of substitution. Supporting loss aversion our findings show effort to be highest when the respective team is leading by one goal. In contrast to the interim score results for ex ante heterogeneity are mixed and depend on the respective model.

We proceed as follows: In the next section we present literature on the effect of heterogeneity and intermediate results on performance. Afterwards we discuss our proxy for effort, describe our data set and present descriptive statistics of the variables of interest (section 4.3). Section 4.4 explains the empirical method used in this article. In the fifth section we present the results of several estimations while the last section concludes.

4.2 Literature Review

Tournament theory suggests intermediate information to have a similar effect on effort as heterogeneity and identifies both to be important drivers of effort provision. Most of the existing studies confirm propositions made by the theory. Literature can be distinguished by the impact of heterogeneity and intermediate results. Hence an overview regarding empirical studies is presented in the next two sections separately. First we present studies with respect to the impact of ex ante heterogeneity. Afterwards we focus on literature on intermediate information and its incentive effects. Some of the presented studies investigate both determinants of effort and therefore are mentioned in both sections.

4.2.1 Impact of Heterogeneity on Effort

Literature concerning asymmetric tournaments can be categorized into experimental and empirical studies which in turn can be classified in firm and sports studies.

Bull et al. (1987) experimentally investigate incentive effects of tournaments and find that in asymmetric contests effort levels of disadvantaged agents are much higher than predicted by the theory while the behavior of the advantaged participants is in accordance with the theory. Schotter and Weigelt (1992) analyze the impact of affirmative action programs and equal opportunity laws - which are modeled as rank order tournaments - on effort of heterogeneous agents. In general the results are consistent with assumptions made by the theory.

Backes-Gellner and Pull (2013) investigate both theoretically and empirically the role of employee heterogeneity concerning the performance of sales representatives. The empirical results are highly consistent with tournament theory: the performance of sales representatives is negatively related to heterogeneity. However, this effect depends on the several aspects, e.g. the number of prizes and participants of the tournament.

A wide range of literature about incentive effects of asymmetric contests uses non-experimental field data of sporting competitions as detailed information on both the tournament and the contestants often is publicly available (Kahn, 2000). Frick et al. (2008), for example, investigate the impact of heterogeneity in ability on effort provision with data from the German Bundesliga. Their analyzes on game level basis suggests that ex ante heterogeneity significantly reduces effort provided by both teams. Bach et al. (2009) confirm these results with an analyzes of Olympic rowing regattas. The authors show that the more capable oarsmen row faster times when the heterogeneity of the starting field decreases while underdogs always provide highest effort.

Sunde (2009) and Lallemand et al. (2008) make use of tennis data in order to analyze the effect of ability differences between the two players on effort. While Sunde's results confirm that highest effort is exerted in homogeneous contests and therefore support theoretical assumptions, Lallemand et al. (2008) find that in uneven matches favorites (underdogs) win more (less) games, i.e. perform better (worse). They conclude that ability differences tend to have greater influence on the outcome of a match than effort differences.

Brown's empirical analyzes of golf data (Brown, 2011) confirms the previous results. Brown analyzes the adverse incentive effect of superstars in tournaments, more precisely the impact of Tiger Woods on effort provided by the other golfers. Results indicate that the presence of Tiger Woods significantly decreases performance of the other competitors. This negative effect is strongest for the higher - skilled players.

Berger and Nieken (2014) study handball teams and whether they react to heterogeneity and intermediate information measured by the half-time score of a match. They find that the intensity of a match respectively half-time is negatively related to ex ante heterogeneity of the teams. However, the results indicate that this effect is mostly driven by the favorite team.

4.2.2 Literature on intermediate information

The experimental study conducted by Bull et al. (1987) also investigates the impact of information on intermediate rank and performance on future effort provision. Results suggest that providing information does not influence effort by the agents. Schotter and Weigelt (1992) confirm this finding with experimental data. Gürtler and Harbring (2010) analyze formally as well as experimentally whether a principal's feedback policies affect agents' performance. Results are in line with their prediction that revealing information by the principal is optimal only if the agents are rather homogeneous. Once intermediate information indicates large differences in performance it is detrimental to effort provision.

Ludwig and Lünser (2012) experimentally study the impact of intermediate performance information on effort in symmetric two-stage tournaments. Results indicate that if contestants can observe each other's effort in the first stage, the competitor who is trailing tends to increase and the one who is leading tends to decrease effort compared to the initial stage. The larger the observed differences in effort the lower the impact on effort in the second stage.

Azmat and Iriberri (2010) make use of data that result from a natural experiment at a high school. They study whether feedback information on relative performance affects students' behavior. They find a significant increase in student's grades if feedback information is provided, especially for high ability students. Individuals appear to have natural competitive preferences which is why they respond to additional information. This implies that releasing additional information increases (decreases) benefits for those students being ahead (being behind).

Casas-Arce and Martínez-Jerez (2009) investigate incentive effects of heterogeneity in multi-period tournaments both theoretically and empirically and thereby make use of data from sales contests of a commodity manufacturer. Results indicate that the effect of releasing intermediate performance information is similar to that of ex ante heterogeneity as leading contestants reduce effort with increasing distance to the closest follower. However, trailing competitors decrease effort only if the distance to a better rank is very large.

Although there are a lot of sports studies investigating incentive effect of sport competitions, only few focus on the impact on intermediate results on effort exertion of the contestants. Lynch (2005), for example, examines incentive effects of horse races. As organizers of horse races use handicaps to improve homogeneity among the starting field, the focus of his study does not lie on heterogeneity of the starting field but the closeness of a race and its impact on effort. Results show that jockeys increase effort when the distance between them and their closest competitors is comparably small, indicating that interim information significantly affects effort exerted by the jockeys.

Even though Berger and Nieken (2014) focus their study on ex ante heterogeneity they also investigate whether the score at the end of the first half of a handball game affects the intensity of the second half. As the coefficient for the half-time score is insignificant the authors conclude that additional information on the winning probability of a team does not affect the intensity of a match.

We continue research concerning the incentive effects of intermediate information by using extensive match-level data from substituted soccer players. Before we present our empirical model and results, section 4.3 presents our data and descriptive statistics.

4.3 Data and Descriptive Statistics

Our data covers detailed pre, post and within-game information for each match of the seasons 2011/12 through 2013/14 of the German Bundesliga. This professional soccer league comprises 18 teams, playing each other twice (once at each team's stadium), resulting in 306 matches per season and 918 match observations overall. Prior to every match a team's coach has to choose 11 players for the starting lineups. In the course of a match he is allowed to replace respectively substitute up to three players.[4]

[4] Up to seven players are allowed to sit on the bench.

On match level we have information on the performance respectively effort exerted by each player on the field. This information refers to the running distance and the number of sprints and intensive runs a player performs in the course of a match. Overall our dataset contains 25,381 player-match observations for 772 different players. Furthermore detailed information on the course of the score of a match and the number of substitutions and sending-offs of both teams was derived from the league's official website at `www.bundesliga.de`.

Unfortunately, the match-level statistics per player always refer to the whole time the respective player is on the field and are unavailable for sub periods of games. Hence it is not possible to estimate the incentive effect of interim results for players who are in the starting team as at the beginning of a match information on interim results doesn't exist yet. Therefore the following analyzes focuses exclusively on *substituted* players because at the time of the substitution intermediate information does already exist. The probability of a team to win an ongoing match crucially depends on the goal difference at the particular time of a match. Therefore we use the score at the time of a substitution as our measure for intermediate information.[5]

Besides intermediate information there is another and even more important key variable, namely effort. So far only few studies have focus on effort of individual soccer players or teams. In the next section we briefly describe how these studies measure effort before we explain our procedure. There is an extensive literature investigating different determinants of performance, success or productivity of teams or individual athletes.[6]

4.3.1 Measuring Effort in Sports

There is an extensive literature investigating different determinants of performance, success or productivity of teams or individual athletes.[7] However, only few studies focus explicitly on *effort*. This can mainly be attributed to the fact that effort is hard to measure as it "is [often] not directly observable by the principal or the audience (including the econometrician), which constitutes the major empirical problem for testing the incentive effect" (Sunde, 2009, p. 3200). Even though sports data provide manifold and extensive statistics, it is often not clear what

[5] Frick et al. (2008) as well as Berger and Nieken (2014) also refer to the interim score as their measure for intermediate information.

[6] Nuesch (2009), for example, analyze the effect of demographic diversity on team performance measured by the final score of a match. Franck and Nüesch (2010) focus on the impact of talent disparity on team productivity and use the same dependent variable as Nuesch (2009), namely the final score of a match represented by the goal difference. In a further study Franck and Nüesch (2011) investigate how wage dispersion affects team productivity. Here the authors use a season-level data set and measure productivity by the ratio of achieved points at the end of a season and the maximum number of possible points. In contrast to these studies Frick (2011) does not focus on aggregate team performance but individual performance and tests whether the contract length affects player performance, measured by the average grade a player received from the soccer magazine KICKER in a given season.

[7] Nuesch (2009), for example, analyze the effect of demographic diversity on team performance measured by the final score of a match. Franck and Nüesch (2010) focus on the impact of talent disparity on team productivity and use the same dependent variable as Nuesch (2009), namely the final score of a match represented by the goal difference. In a further study Franck and Nüesch (2011) investigate how wage dispersion affects team productivity. Here the authors use a season-level data set and measure productivity by the ratio of achieved points at the end of a season and the maximum number of possible points. In contrast to these studies Frick (2011) does not focus on aggregate team performance but individual performance and tests whether the contract length affects player performance, measured by the average grade a player received from the soccer magazine KICKER in a given season.

the best way to measure effort is (Berger and Nieken, 2014). Still, many measures used in prior studies lack a clear indication that they measure effort.

Some sports studies argue that overall team effort can be derived from the intensity of a match which in turn can be approximated by the number of penalties a team received for fouls or other rule violations. Frick et al. (2008), for instance, use the number of cards (yellow, yellow/red, red) a soccer team receives per match as a measure for the intensity of a match respectively team effort. Berger and Nieken (2014) argue in a similar way and use the number of 2-minute suspensions per handball match and team as a measure for "defensive effort". Although they state that this kind of effort merges fluently into sabotage activities, they can show that it is positively related to the winning probability of a team. Therefore they conclude that the number of 2-minute suspensions represents a good proxy for the intensity of a team's play in handball.

Other sports studies use measures that rely on the outcome of a contest to analyze incentive effects of tournaments. Frick and Prinz (2007), for example, study running data and use the running times as their dependent variable while Sunde (2009) and Lallemand et al. (2008) analyze tennis data and estimate incentive effects on a player's (average) number of games won per match. Sunde (2009) thereby constitutes that it is important to disentangle the capability from the incentive effect. However, distinguishing ability and incentive effects is problematic as effort often is unobservable. Therefore he presents a model to show that effort can be identified by separating the competing players into favorites and underdogs and investigating them separately.

All of these studies have in common that they are very cautious with respect to the denotation of the chosen variable as effort. It seems that they estimate incentive effects in a rather indirect way: the first approach referring to the intensity of a match which in turn is related to overall effort, and the second one using the outcome of a contest to separate effort from ability effects afterwards. Due to technological advancements, recently there are extensive match-level statistics for German Bundesliga matches publicly available. These statistics refer to overall team as well as individual player performance and include information on e.g. running distance, number of sprints and intensive runs, duels won, passes played, goal shots etc. and therefore provide an opportunity to measure effort in a more direct way.

The study conducted by Wicker et al. (2013) is one of the first that apply these kinds of statistics to provide an innovative measure for effort. They use information on the number of intensive runs and on the running distance per game and player to capture effort. The authors state that this procedure has the advantage that "() ... a player can choose the level of intensive runs without touching a ball and being productive. To put it differently, an individual can reach his maximum effort independent of his level of ability" (Wicker et al., 2013, p.131). The focus of the study is on the impact of effort on a player's market value respectively salary. They find that a player's market value is not affected by effort.

Similar to Wicker et al. (2013), we use running statistics to measure effort. Aim of our study is to test the impact of intermediate results on effort. We distinguish between three measures for effort: the distance covered by a player (*distance*), the number of sprints (*sprints*) and the number of intensive runs (*runs*) a player performs in the course of a match.

Since the function of goalkeepers differs significantly from the tasks of the other players and does not seem to be related to running we excluded them from the analyzes. In the following we only consider defenders, midfielders and strikers. As the intermediate score varies only for those players who are substituted in the course of a match, we focus on this sub-sample. We therefore test the effect of intermediate score at the time of the substitution on the effort provided by the respective players.

4.3.2 Descriptive Statistics

In the vast majority of matches (85%) coaches exploit the maximum of possible substitutions. Our sample contains only one match where a coach did not substitute at all. On average there are 2.82 substitutions per match and team. Overall, we observe 5,185 substitutions. Almost all substitutions take place in the second half of a match. Only 4.24% of the substitutions occur in the first half of a match. Most substitutions take place in the 46th minute (8.81%) that is during the half time break.

Evaluating effort by substituted players requires some constrains. First, results might be distorted if players are included who only appeared for a few minutes in the game as for those players the variance of the considered statistics is extremely high. We set the limit for minimum appearance in the game at 15 minutes to reduce the impact of noise. Second, players who get substituted during the first half might systematically differ from their second half counterparts as they can recover during the half time break to put forth additional effort during the rest of the game. Excluding goalkeepers and players who played less than 15 minutes and/or were substituted in the first half reduces the data set to 2,802 observations.[8] For undocumented reasons, information on the distance run, number of intensive runs and sprints was unavailable for particular cases so that the final data set contains 2,768 observations regarding the effort measures distance and runs and 2,747 observations concerning the variable sprints. Distance, runs and sprints depend critically on the minutes played by a player. Therefore we divided these variables through the number of minutes played.

Table 4.1 shows descriptive statistics on the variables distance, runs and sprints (per minute played) for the described sub-sample as well as the number of minutes played by these players. On average a substituted player runs 123 meters per minute and does roughly 3 intensive runs and 1 sprint every 4 minutes. As we expect the prospect of winning or losing a match to critically impact effort we apply the score (i.e. the goal difference) at the time of the substitution as our

[8] Overall, 2,383 observations were excluded. These exclusions subdivide into goalkeepers: 20 observations (4 observations refer to goalkeepers who played less than 15 minutes and further 4 observations to goalkeepers who were substituted in first half), players who played less than 15 minutes: 2,148 observations and players who were substituted in first half: 223 observations.

major control variable capturing intermediate information.

	Obs	Mean	SD	Min	Max
Distance	2,768	0.123	0.012	0.079	0.166
Runs	2,768	0.740	0.214	0.100	1.765
Sprints	2,747	0.235	0.111	0.021	0.765
Minutes played	2,768	27.396	9.642	15	45

Table 4.1: Descriptive statistics on effort variables (per minute played) and minutes played.

There are two ways to operationalize the goal difference: one can determine the (absolute) goal difference or generate dummy-variables for each goal difference. Following assumptions by tournament theory, we expect that a large goal difference at the time of substitution has a negative effect on effort, irrespective of whether the team is leading or trailing the match. Therefore we use dummy variables for the respective goal differences instead of a variable representing the absolute goal difference.

Figure 4.1 shows that most of the substitutions take place when the respective team trails by one goal. This is not surprising as substitutions most often take place to change the course of the game. There are very few observations referring to a goal difference larger than 3 (-3). Hence we pool all observations greater or equal to 3 (-3).

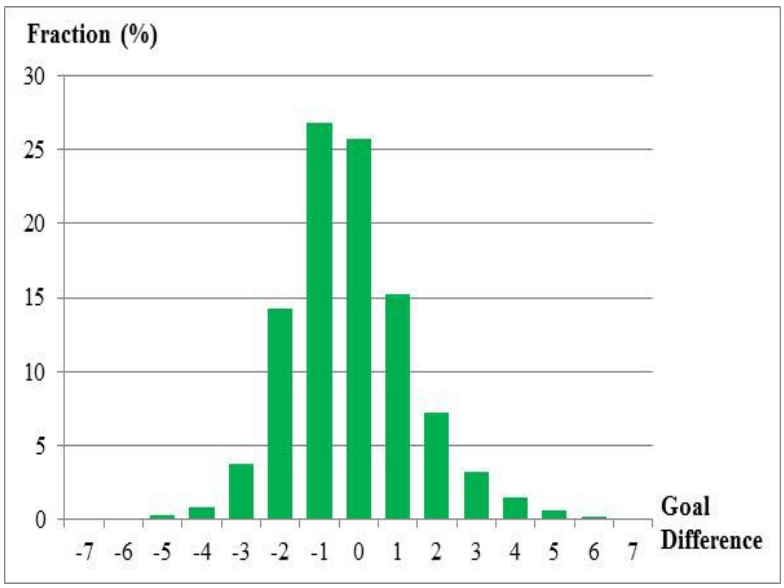


Figure 4.1: Distribution of goal difference at the time of substitution.

Figure 4.2 shows the average running distance, number of runs and sprints per minute in relation to the goal difference at the time of substitution. For all effort measures highest values are reached when the team is leading by one or two goals and lowest when the respective team is trailing or - except for sprints - leading by 3 or more goals at the time of substitution.

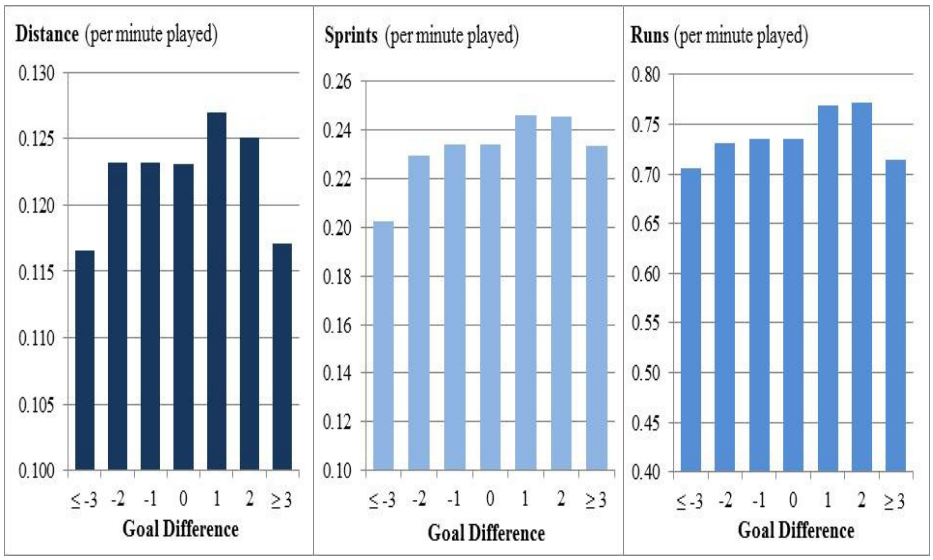


Figure 4.2: Average distance, runs and sprints per goal difference at time of substitution.

In addition to intermediate information further aspects determine individual effort and need to be controlled for. As already mentioned, tournament theory predicts lower effort levels for asymmetric contests. Therefore we control for the ex ante heterogeneity (*heterogeneity*) between the two teams, which we operationalize via betting odds. We rely on information from the website `www.betexplorer.com`. Betting odds have proven to be a good measure in displaying ex ante strength of the teams.[9] We measure heterogeneity as the absolute difference between the winning probabilities of the two teams, which can easily be drawn from betting odds by bookmakers. We assume a negative impact of ex ante heterogeneity on effort.

Furthermore we control for the remaining number of minutes to be played at the time of the substitution (*remaining*) and the number of sending-offs that the respective team received prior to the substitution (*sendingoffs*). A player who is substituted in the 46th minute has to pace himself for a longer period than a player who enters the pitch in the 76th minute and therefore has to choose a lower effort level per minute. After a dismissal the remaining players have to compensate the loss of a player and therefore (should) run or sprint more (often). We hypothesize a positive impact of sending offs on the effort provided by the remaining players.[10]

[9] Cf. Garicano et al. (2005); Deutscher et al. (2013); Frick et al. (2008); Berger and Nieken (2014) for research applying betting odds.

[10] The impact of sending-offs for the respective opponent has no significant impact on the findings to follow and has thus been neglected in the estimations to follow.

There are a lot of studies confirming a "home advantage" in soccer[11] that is the fraction of home team wins is considerably larger than home team losses.[12] Several theories circulate around this phenomenon. One refers to the role of the crowd. Social support by the home fans might influence players' behavior, resulting in higher effort and better performance (Holder and Nevill, 1997). Additionally home teams usually prefer a more offensive style of play. Implementing a rather defensive style of play in turn is accompanied by fewer runs for away teams. Therefore a dummy variable is included that indicates if the substituted player is a member of the away team or not (*away*). We assume a negative impact of being in the away team.

Furthermore it is necessary to control for the intensity of a match prior to a substitution because it makes a difference if a player is substituted in a match where both teams act very cautiously respectively defensively or if he enters a match with two very offensive playing teams. The goal difference does not capture this intensity as it indicates the difference of the goals scored by the two teams and not the total of all goals scored in a particular match which better reflects the intensity of a match. Since the total number of goals scored prior to a substitution might be correlated with the goal difference at the time of the substitution and the remaining minutes, we implement a different indicator displaying intensity. Effort provided by the replaced player is our proxy for the intensity of a match (*effort_replaced*). When the replaced player has shown a high effort level, the match is expected to be more intense than in the case of low effort exerted by the replaced player. We expect a positive effect of this variable on the effort by the substituted player. The effort by the replaced player always refers to the estimated effort measure: when we use distance as the dependent variable, effort by the replaced player also refers to distance, when the dependent variable is runs, effort by the replaced refers to runs.

Finally we also include two variables capturing the respective matchday and its square as well as team, opponent and player dummies to control for unobserved team, opponent and player effects.

Table 4.2 shows descriptive statistics of the control variables. Average descriptive statistics appear to be similar for the replaced and substituted player in terms of running distance, while the number of runs and sprints is significantly higher for the substituted player.

[11] Cf. Courneya and Carron (1992); Clarke and Norman (1995); Nevill et al. (1996); Nevill and Holder (1999); Nevill et al. (2002).

[12] In our data set, 45% of the matches end with home wins, 24% with a tie and only 30% with a home loss.

	Obs	Mean	SD	Min	Max
Heterogeneity	2,768	-0.002	0.364	-0.865	0.865
Remaining	2,768	29.462	9.676	17	47
Sendingoffs	2,768	0.040	0.203	0	2
Away	2,768	0.517	0.500	0	1
Effort_replaced distance	2,768	0.123	0.009	0.090	0.159
Effort_replaced runs	2,768	0.690	0.175	0.111	1.422
Effort_replaced sprints	2,765	0.212	0.090	0.013	0.542

Table 4.2: Descriptive statistics of control variables.

4.4 Results

In order to test the impact of intermediate information on individual effort we apply OLS regression analyzes. We present three models as we use three dependent variables.

We estimated Random and Fixed Effects regression. The Hausman Test was significant at 10% level, indicating that fixed effects were slightly more appropriate. Overall, fixed and random effects estimations provide very similar results.

Concerning the impact of intermediate information on effort, table 4.3 shows that players provide highest effort when their team is leading by one goal. We choose a tied score as our reference category (*goaldiff = 0*). Except for the number of sprints the coefficient of *goaldiff = +1* is highly significant and positive in all of the other models. Compared to a tied score a player runs about 3.5 meters more per minute and provides one additional intensive run every 3 minutes when leading by one goal.

Interestingly, trailing by one goal leads to significant less effort compared to a balanced score or to leading the match. In all models the coefficients for *goaldiff = -1* are negative and significant. For a scoring system that incentivizes offense (3 points for a win, 1 point for a draw and 0 points for a loss) these findings support the idea of loss aversion (Kahneman and Tversky, 1979).[13] Individuals care more about losing something than winning the exact same thing. A team that leads a match by only one goal runs the risk of losing 2 points in case the opponent scores a single goal. This threat of losing two points should have a stronger incentive than the possibility of winning two more points by scoring a goal in case of a tied score. The same should hold true when comparing the incentive effects of a tied score with the situation of trailing by one goal. Again, losing 1 point reflects a higher value than gaining an additional point. Hence, we expect effort to be higher when the score is tied compared to when the team is trailing by one goal. For our estimations both assumptions are supported by the data, since compared to a tied score effort is significantly higher when the team leads and significantly lower when the

[13] For findings from professional golf see Pope and Schweitzer (2011).

team trails by one goal.

Intermediate information indicating a match to already be decided at the time of a substitution (*goaldiff* $\leq$ *-3; goaldiff* $\geq$ *+3*) has a negative effect: the coefficients are negative for all three models and highly significant for distance, but only slightly significant respectively insignificant for runs.

Besides intermediate information further variables impact the effort level provided by the player: the remaining playing time and the effort of the replaced player are significant in all models. The more minutes are to be played from the time of the substitution the less effort per minute is shown as players have to economize. The intensity of the game proxied by the effort of the replaced player affects strongly effort exerted by the substituted player. Higher effort levels of the replaced players are accompanied by significantly higher effort by the substituted player.

Concerning ex ante heterogeneity, results are mixed even though its coefficient is negative in all models. The coefficient is significant and negative for runs and in part for sprints (Model 3), but insignificant for distance. As heterogeneity is negative for underdogs and positive for favorites this indicates that favorites tend to decrease effort the more likely they will win while underdogs seem to get additional motivation the less likely a win is prior to a match. This result is in line with previous research on heterogeneous contests.[14]

In summary, results for distance and runs are very similar and indicate strong effects of intermediate results on effort. In contrast to runs and distance the goal difference merely affects the number of sprints, especially in the Random Effects Model (Model 6). Although the coefficients for sprints are the same as for the other two effort measures, sprinting seems to differ from running.

For robustness checks we conducted estimations including additional control variables, e.g. manager fixed effects, the kind of substitution (midfielder for defender, midfielder for striker etc.), competing in the Champions or Europe League or the national Cup Competition (DFB Pokal) or the rank of a team prior to the match.[15] While these indicators prove to be insignificant or correlated with the goal difference the findings of intermediate information on effort provision remained robust.

[14] Cf. Bull et al. (1987); Berger and Nieken (2014).
[15] These regression results are presented in tables 4.4 to 4.7 in section 4.A1.

	Fixed Effects			Random Effects		
	(1)	(2)	(3)	(4)	(5)	(6)
	Distance	Runs	Sprints	Distance	Runs	Sprints
Goaldiff $\leq$ -3	-0.0061***	-0.0171	-0.0102	-0.0060***	-0.0159	-0.0128
	(0.0009)	(0.0171)	(0.0094)	(0.0009)	(0.0163)	(0.0089)
Goaldiff = -2	-0.0010*	-0.0411***	-0.0123**	-0.0011*	-0.0257**	-0.0050
	(0.0006)	(0.0113)	(0.0062)	(0.0006)	(0.0108)	(0.0059)
Goaldiff = -1	-0.0012**	-0.0328***	-0.0122**	-0.0010**	-0.0187**	-0.0043
	(0.0005)	(0.0093)	(0.0051)	(0.0005)	(0.0089)	(0.0049)
Goaldiff = +1	0.0034***	0.0303***	0.0088	0.0030***	0.0231**	0.0075
	(0.0006)	(0.0112)	(0.0062)	(0.0006)	(0.0107)	(0.0059)
Goaldiff = +2	0.0013*	0.0269*	0.0079	0.0010	0.0157	0.0034
	(0.0008)	(0.0146)	(0.0081)	(0.0008)	(0.0142)	(0.0078)
Goaldiff $\geq$ +3	-0.0068***	-0.0296*	-0.0142	-0.0065***	-0.0305*	-0.0122
	(0.0010)	(0.0176)	(0.0098)	(0.0009)	(0.0169)	(0.0093)
Heterogeneity	-0.0015	-0.0749**	-0.0285*	-0.0019	-0.0732**	-0.0224
(ex ante)	(0.0017)	(0.0306)	(0.0169)	(0.0016)	(0.0288)	(0.0157)
Remaining	-0.0000**	-0.0030***	-0.0011***	-0.0001***	-0.0033***	-0.0011***
	(0.0000)	(0.0004)	(0.0002)	(0.0000)	(0.0003)	(0.0002)
Sendingoffs	0.0015	0.0525***	0.0040	0.0019**	0.0311*	-0.0029
	(0.0010)	(0.0173)	(0.0095)	(0.0009)	(0.0164)	(0.0089)
Away	-0.0008	-0.0143	-0.0136**	-0.0008	-0.0150	-0.0114**
	(0.0006)	(0.0111)	(0.0061)	(0.0006)	(0.0106)	(0.0058)
Effort_replaced	0.1222***	0.1664***	0.1344***	0.1605***	0.2348***	0.1878***
	(0.0216)	(0.0212)	(0.0226)	(0.0199)	(0.0198)	(0.0210)
Matchday	0.0003***	0.0026*	-0.0011	0.0003***	0.0021	-0.0013*
	(0.0001)	(0.0014)	(0.0008)	(0.0001)	(0.0014)	(0.0007)
Matchday2	-0.0000***	-0.0001*	0.0000	-0.0000***	-0.0000	0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Constant	0.1087***	0.6376***	0.1932***	0.1053***	0.6333***	0.2219***
	(0.0049)	(0.0758)	(0.0411)	(0.0036)	(0.0478)	(0.0252)
Team effects	Yes	Yes	Yes	Yes	Yes	Yes
Opponent effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,768	2,768	2,744	2,768	2,768	2,744
Adj. R-sq	0.450	0.449	0.383	0.127	0.144	0.1

a. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.

b. Weighted with number of minutes played by player.

c. Reference category is Goaldiff=0.

Table 4.3: Regression results - Fixed and Random Effects (weighted with number of minutes played).

4.5 Conclusion

This article examines how soccer players response to intermediate information in the course of a match, focusing on substitute players from the German Bundesliga. In an innovative approach we measure effort via the number of runs and running distance a player covered during a match. Results suggest that effort is highest when the respective team is leading by one goal. In line with prospect theory individuals value potential losses higher than gains. Compared to a tied score effort is higher when a team trails by one goal. Effort declines once intermediate information concerning the score indicates the game to be decided.

Regarding the release of intermediate information the following can be concluded: In case that a contest is already decided it is inadvisable to give information to the contestants. Both the contestant leading and the contestant trailing will decrease effort to save effort costs. If the gap between the competitors is small the contestant in lead should be informed in order to increase effort. On the other hand the trailing contestant should not receive any information concerning the intermediate score. Second if intermediate information is knowledge to the contestants handicapping the contestant with a big lead increases effort by both parties. In case intermediate information suggests the contest to be close the trailing competitors should be incentivized by e.g. bonus pay in case of outperforming the leader.

4.A1 Robustness Checks

	(1) Distance	(2) Runs	(3) Sprints
Goaldiff ≤ -3	-0.0062***	-0.0195	-0.0109
	(0.0009)	(0.0171)	(0.0095)
Goaldiff = -2	-0.0011*	-0.0405***	-0.0118*
	(0.0006)	(0.0114)	(0.0063)
Goaldiff = -1	-0.0013**	-0.0324***	-0.0118**
	(0.0005)	(0.0094)	(0.0052)
Goaldiff = +1	0.0033***	0.0299***	0.0085
	(0.0006)	(0.0112)	(0.0062)
Goaldiff = +2	0.0011	0.0258*	0.0074
	(0.0008)	(0.0147)	(0.0081)
Goaldiff ≥ +3	-0.0070***	-0.0332*	-0.0154
	(0.0010)	(0.0178)	(0.0099)
Heterogeneity (ex ante)	-0.0013	-0.0712**	-0.0274
	(0.0017)	(0.0308)	(0.0170)
Remaining	-0.0000**	-0.0030***	-0.0011***
	(0.0000)	(0.0004)	(0.0002)
Sendingoffs	0.0010	0.0479***	0.0024
	(0.0009)	(0.0174)	(0.0096)
Away	-0.0006	-0.0131	-0.0134**
	(0.0006)	(0.0112)	(0.0062)
Effort_replaced	0.1217***	0.1604***	0.1327***
	(0.0225)	(0.0216)	(0.0227)
Matchday	0.0003***	0.0023	-0.0012
	(0.0001)	(0.0014)	(0.0008)
Matchday2	-0.0000***	-0.0001	0.0000
	(0.0000)	(0.0000)	(0.0000)
CL/EL match before	0.0018***	0.0343***	0.0087
	(0.0007)	(0.0124)	(0.0069)
CL/EL match afterwards	0.0005	0.0041	0.0031
	(0.0007)	(0.0128)	(0.0071)
DFB Cup match before	-0.0005	0.0117	0.0060
	(0.0009)	(0.0167)	(0.0093)
DFB Cup match afterwards	0.0004	0.0206	0.0063
	(0.0008)	(0.0144)	(0.0079)
Constant	0.1076***	0.6290***	0.1814***
	(0.0049)	(0.0760)	(0.0416)
Player effects	Yes	Yes	Yes
Team effects	Yes	Yes	Yes
Opponent effects	Yes	Yes	Yes
Substitution effects	Yes	Yes	Yes
Manager effects	No	No	No
Observations	2,768	2,768	2,744
Adj. R-sq.	0.463	0.450	0.382

a. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.
b. Weighted with number of minutes played by player.
c. Reference category is Goaldiff=0.

Table 4.4: Robustness Check with Fixed Effects I (weighted with number of minutes played).

	(4) Distance	(5) Runs	(6) Sprints	(7) Distance	(8) Runs	(9) Sprints
Goaldiff ≤ -3	-0.0061***	-0.0179	-0.0107	-0.0058***	-0.0181	-0.0134
	(0.0009)	(0.0173)	(0.0096)	(0.0009)	(0.0173)	(0.0096)
Goaldiff = -2	-0.0010	-0.0382***	-0.0133**	-0.0010*	-0.0415***	-0.0147**
	(0.0006)	(0.0116)	(0.0064)	(0.0006)	(0.0116)	(0.0064)
Goaldiff = -1	-0.0011**	-0.0313***	-0.0136**	-0.0013**	-0.0322***	-0.0143***
	(0.0005)	(0.0096)	(0.0053)	(0.0005)	(0.0096)	(0.0053)
Goaldiff = +1	0.0034***	0.0298***	0.0073	0.0035***	0.0355***	0.0095
	(0.0006)	(0.0115)	(0.0063)	(0.0006)	(0.0114)	(0.0063)
Goaldiff = +2	0.0013	0.0293**	0.0084	0.0015*	0.0321**	0.0067
	(0.0008)	(0.0149)	(0.0082)	(0.0008)	(0.0149)	(0.0083)
Goaldiff ≥ +3	-0.0067***	-0.0290	-0.0169*	-0.0067***	-0.0231	-0.0129
	(0.0010)	(0.0180)	(0.0100)	(0.0010)	(0.0179)	(0.0100)
Heterogeneity (ex ante)	-0.0002	-0.0467	-0.0140	0.0001	-0.0354	-0.0165
	(0.0019)	(0.0358)	(0.0198)	(0.0020)	(0.0373)	(0.0208)
Remaining	-0.0000**	-0.0029***	-0.0010***	-0.0000**	-0.0027***	-0.0009***
	(0.0000)	(0.0004)	(0.0002)	(0.0000)	(0.0004)	(0.0002)
Sendingoffs	0.0012	0.0513***	0.0025	0.0011	0.0441**	-0.0028
	(0.0010)	(0.0182)	(0.0100)	(0.0010)	(0.0181)	(0.0100)
Away	-0.0003	-0.0059	-0.0085	-0.0003	-0.0025	-0.0094
	(0.0007)	(0.0124)	(0.0069)	(0.0007)	(0.0128)	(0.0071)
Effort_replaced	0.1139***	0.1600***	0.1332***	0.1090***	0.1350***	0.1042***
	(0.0231)	(0.0221)	(0.0231)	(0.0233)	(0.0223)	(0.0236)
Matchday	0.0002***	0.0035**	-0.0014	0.0003***	0.0039**	-0.0015*
	(0.0001)	(0.0016)	(0.0009)	(0.0001)	(0.0016)	(0.0009)
Matchday2	-0.0000***	-0.0001*	0.0000	-0.0000***	-0.0001**	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
CL/EL match before	0.0016**	0.0327**	0.0082	0.0017**	0.0342***	0.0083
	(0.0007)	(0.0130)	(0.0072)	(0.0007)	(0.0129)	(0.0072)
CL/EL match afterwards	0.0003	0.0049	0.0033	0.0006	0.0092	0.0041
	(0.0007)	(0.0131)	(0.0073)	(0.0007)	(0.0131)	(0.0073)
DFB Cup match before	-0.0006	0.0107	0.0063	-0.0009	0.0042	0.0033
	(0.0009)	(0.0168)	(0.0093)	(0.0009)	(0.0168)	(0.0093)
DFB Cup match afterwards	0.0003	0.0200	0.0064	0.0000	0.0139	0.0032
	(0.0008)	(0.0144)	(0.0080)	(0.0008)	(0.0144)	(0.0080)
League position^{+}	0.0001	0.0002	0.0001	0.0000	-0.0017	-0.0005
	(0.0001)	(0.0013)	(0.0007)	(0.0001)	(0.0014)	(0.0008)
League position^{+} (opponent)	-0.0001	-0.0022*	-0.0010	-0.0001	-0.0025**	-0.0010
	(0.0001)	(0.0011)	(0.0006)	(0.0001)	(0.0011)	(0.0006)
Constant	0.1088***	0.6368***	0.2020***	0.1197***	0.9367***	0.2866***
	(0.0055)	(0.0881)	(0.0481)	(0.0108)	(0.1913)	(0.1061)
Player effects	Yes	Yes	Yes	Yes	Yes	Yes
Team effects	Yes	Yes	Yes	Yes	Yes	Yes
Opponent effects	Yes	Yes	Yes	Yes	Yes	Yes
Substitution effects	Yes	Yes	Yes	Yes	Yes	Yes
Manager effects	No	No	No	Yes	Yes	Yes
Observations	2,684	2,684	2,661	2,684	2,684	2,661
Adj. R-sq.	0.466	0.451	0.382	0.470	0.466	0.392

a. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.
b. Weighted with number of minutes played by player.
c. Reference category is Goaldiff=0.
d. + Position in the table before actual matchday.

Table 4.5: Robustness Check with Fixed Effects II (weighted with number of minutes played).

	(1) Distance	(2) Runs	(3) Sprints
Goaldiff ≤ -3	-0.0061***	-0.0183	-0.0140
	(0.0009)	(0.0164)	(0.0090)
Goaldiff = -2	-0.0009	-0.0287***	-0.0071
	(0.0006)	(0.0109)	(0.0060)
Goaldiff = -1	-0.0008	-0.0206**	-0.0070
	(0.0005)	(0.0090)	(0.0050)
Goaldiff = +1	0.0032***	0.0338***	0.0122**
	(0.0006)	(0.0107)	(0.0059)
Goaldiff = +2	0.0011	0.0229	0.0057
	(0.0008)	(0.0143)	(0.0079)
Goaldiff ≥ +3	-0.0061***	-0.0242	-0.0087
	(0.0009)	(0.0171)	(0.0094)
Heterogeneity (ex ante)	-0.0016	-0.0698**	-0.0241
	(0.0015)	(0.0290)	(0.0159)
Remaining	-0.0000***	-0.0031***	-0.0012***
	(0.0000)	(0.0003)	(0.0002)
Sendingoffs	0.0017*	0.0380**	-0.0056
	(0.0009)	(0.0166)	(0.0091)
Away	-0.0006	-0.0131	-0.0123**
	(0.0006)	(0.0106)	(0.0058)
Effort_replaced	0.1263***	1.8628***	0.4377**
	(0.0213)	(0.4009)	(0.2207)
Matchday	0.0003***	0.0031**	-0.0012
	(0.0001)	(0.0014)	(0.0008)
Matchday2	-0.0000***	-0.0001*	0.0000*
	(0.0000)	(0.0000)	(0.0000)
CL/EL match before	0.0018***	0.0309**	0.0098
	(0.0006)	(0.0121)	(0.0066)
CL/EL match afterwards	0.0000	0.0016	0.0056
	(0.0007)	(0.0123)	(0.0068)
DFB Cup match before	-0.0003	0.0119	0.0046
	(0.0009)	(0.0162)	(0.0089)
DFB Cup match afterwards	0.0003	0.0213	0.0057
	(0.0007)	(0.0139)	(0.0076)
Constant	0.1021***	0.3882***	0.1568***
	(0.0036)	(0.0683)	(0.0376)
Team effects	Yes	Yes	Yes
Opponent effects	Yes	Yes	Yes
Substitution effects	Yes	Yes	Yes
Manager effects	No	No	No
Observations	2,768	2,768	2,747
Adj. R-sq.	0.2105	0.1655	0.0787

a. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.
b. Weighted with number of minutes played by player.
c. Reference category is Goaldiff=0.

Table 4.6: Robustness Check with Random Effects I (weighted with number of minutes played).

	(4) Distance	(5) Runs	(6) Sprints	(7) Distance	(8) Runs	(9) Sprints
Goaldiff ≤ -3	-0.0060***	-0.0146	-0.0136	-0.0058***	-0.0161	-0.0158*
	(0.0009)	(0.0165)	(0.0091)	(0.0009)	(0.0163)	(0.0091)
Goaldiff = -2	-0.0008	-0.0266**	-0.0084	-0.0009	-0.0320***	-0.0103*
	(0.0006)	(0.0110)	(0.0061)	(0.0006)	(0.0109)	(0.0061)
Goaldiff = -1	-0.0006	-0.0186**	-0.0084*	-0.0006	-0.0195**	-0.0087*
	(0.0005)	(0.0092)	(0.0051)	(0.0005)	(0.0091)	(0.0051)
Goaldiff = +1	0.0034***	0.0348***	0.0114*	0.0035***	0.0394***	0.0132**
	(0.0006)	(0.0109)	(0.0060)	(0.0006)	(0.0108)	(0.0060)
Goaldiff = +2	0.0012	0.0266*	0.0066	0.0014*	0.0261*	0.0042
	(0.0008)	(0.0145)	(0.0080)	(0.0008)	(0.0143)	(0.0079)
Goaldiff ≥ +3	-0.0059***	-0.0206	-0.0106	-0.0058***	-0.0129	-0.0055
	(0.0009)	(0.0173)	(0.0095)	(0.0009)	(0.0170)	(0.0094)
Heterogeneity (ex ante)	-0.0004	-0.0412	-0.0128	0.0004	-0.0230	-0.0138
	(0.0018)	(0.0338)	(0.0186)	(0.0019)	(0.0353)	(0.0196)
Remaining	-0.0000***	-0.0031***	-0.0011***	-0.0000**	-0.0028***	-0.0010***
	(0.0000)	(0.0003)	(0.0002)	(0.0000)	(0.0003)	(0.0002)
Sendingoffs	0.0019**	0.0392**	-0.0063	0.0017*	0.0333**	-0.0096
	(0.0009)	(0.0172)	(0.0095)	(0.0009)	(0.0170)	(0.0094)
Away	-0.0003	-0.0053	-0.0087	-0.0001	-0.0001	-0.0094
	(0.0006)	(0.0118)	(0.0065)	(0.0007)	(0.0121)	(0.0067)
Effort_replaced	0.1202***	1.8493***	0.4206*	0.1153***	1.8030***	0.4669**
	(0.0217)	(0.4087)	(0.2250)	(0.0219)	(0.4047)	(0.2244)
Matchday	0.0003***	0.0045***	-0.0012	0.0003***	0.0046***	-0.0013
	(0.0001)	(0.0015)	(0.0008)	(0.0001)	(0.0015)	(0.0009)
Matchday2	-0.0000***	-0.0001**	0.0000*	-0.0000***	-0.0001***	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
CL/EL match before	0.0017**	0.0297**	0.0097	0.0018***	0.0335***	0.0107
	(0.0007)	(0.0125)	(0.0069)	(0.0007)	(0.0124)	(0.0069)
CL/EL match afterwards	-0.0002	0.0005	0.0057	0.0001	0.0062	0.0071
	(0.0007)	(0.0127)	(0.0070)	(0.0007)	(0.0126)	(0.0070)
DFB Cup match before	-0.0004	0.0123	0.0051	-0.0007	0.0015	-0.0002
	(0.0009)	(0.0163)	(0.0089)	(0.0009)	(0.0161)	(0.0090)
DFB Cup match afterwards	0.0002	0.0209	0.0059	-0.0000	0.0133	0.0016
	(0.0007)	(0.0139)	(0.0076)	(0.0007)	(0.0138)	(0.0076)
League position^{+}	0.0001	0.0007	0.0000	0.0000	-0.0017	-0.0010
	(0.0001)	(0.0012)	(0.0007)	(0.0001)	(0.0013)	(0.0007)
League position^{+} (oponent)	-0.0001	-0.0023**	-0.0011*	-0.0001	-0.0029***	-0.0012**
	(0.0001)	(0.0011)	(0.0006)	(0.0001)	(0.0011)	(0.0006)
Constant	0.1018***	0.3986***	0.1827***	0.0872***	0.2615*	0.0921
	(0.0039)	(0.0744)	(0.0410)	(0.0085)	(0.1562)	(0.0817)
Team effects	Yes	Yes	Yes	Yes	Yes	Yes
Opponent effects	Yes	Yes	Yes	Yes	Yes	Yes
Substitution effects	Yes	Yes	Yes	Yes	Yes	Yes
Manager effects	No	No	No	Yes	Yes	Yes
Observations	2,684	2,684	2,664	2,684	2,684	2,664
Adj: R-sq:	0.2121	0.1680	0.0772	0.2333	0.2284	0.1289

a. Standard errors in parentheses, * p<0.1, ** p<0.05, *** p<0.01.
b. Weighted with number of minutes played by player.
c. Reference category is Goaldiff=0.
d. + Position in the table before actual matchday.

Table 4.7: Robustness Check with Random Effects II (weighted with number of minutes played).

Chapter 5

Nobody's Innocent – The Role of Customers in the Doping Dilemma[1]

Abstract

Customers who boycott an organization after some scandal may actually exacerbate the fraud problem they would like to prevent. This conclusion is derived from a game-theoretic model that introduces a third player into the standard inspection game. Focusing on the example of doping in professional sports, we observe that doping is prevalent in equilibrium because customers undermine an organizer's incentives to inspect the athletes. Establishing transparency about doping tests is necessary but not sufficient to overcome this dilemma. Our analyzes has practical implications for the design of anti-doping policies, as well as for other situations of fraudulent activities.

Keywords inspection game, doping, professional sports, scandals, cheating
JEL Classification K42, L83, C72

[1] A later version of this chapter is published in the Journal of Sports Economics, 2014, Doi: 10.1177/1527002514551475. It is co-authored by Berno BÃ¼chel and Eike Emrich.

5.1 Introduction

When fraudulent activities are detected in some organization, the customers have to make a decision. Either they continue the relationship with this organization or they boycott it. Behaving in the latter way, i.e. reacting to scandals with a withdrawal of support, can be expected to reduce the extent of fraudulent activities (since the potential loss increases). As we will show in this paper, however, the effect might just go into the opposite direction: critical customers who withdraw support after a scandal unintendedly trigger fraudulent activities. This conclusion follows from a game-theoretic model which extends the standard inspection game by an additional player.[2] We carefully analyze and discuss this model focusing on the example of doping in professional sports.

Sport events, such as the Olympic Games, have grown to a size of substantial economic importance. Thereby the use of performance-enhancing drugs (doping) is considered as a risk for the sports industry. There are at least three reasons why it is socially desirable to reduce the extent of doping (cf. Preston and Szymanski, 2003.). First, as it is well known, the use of performance-enhancing drugs can lead to serious health problems for the athletes. Second, athletes often serve as role models.[3] Thus, a doped athlete is neither in the best interest of parents nor would she give the right image for a sponsoring company. Closely linked to that point is the third argument: an important character of sport is that it becomes uninteresting if athletes systematically violate the rules.[4] Given these arguments, it is not surprising that even the United Nations and the European Commission (EC) are interested in anti-doping policies.[5] The most important scientific questions on doping concern (i) the actual extent of doping – whether the use of performance-enhancing drugs is an exceptional practice of some delinquent athletes or a common practice – and they concern (ii) instruments to reduce the extent of doping.

Despite the rich set of anecdotal evidence, empirical studies about doping are rare. It seems very hard to collect data of high quality. Those few studies that try to assess the extent of doping empirically, make estimations that often strongly exceed the public perception (Pitsch et al., 2007; Striegel et al., 2010; Pitsch and Emrich, 2011).[6] Theoretical approaches to the doping issue have acknowledged that decisions to dope are not independent of decisions of other actors such as other athletes or control agencies. Game theory provides tools to analyze such situations of strategic interaction. The primary focus is thereby given to the interaction among athletes. Since the pioneer work of Breivik (1992), this interaction is often modeled as a prisoner's dilemma, where to dope is the dominant strategy (cf. Bird and Wagner, 1997;

[2] Inspection games are discussed by Dresher (1962), Maschler (1967), Tsebelis (1989), and Avenhaus et al. (2002), among others.

[3] Results of an online-survey reveal that spectators require that athletes serve as role models for a clean and doping-free sport (Emrich et al., 2014).

[4] As a survey on the Olympic Games shows, spectators, fans etc. want to see records and high performances but only under compliance of the rules (Messing and Müller, 1996).

[5] The United Nations Educational, Scientific and Cultural Organization (UNESCO) has established a sizable fund dedicated to "the Elimination of Doping in Sport"
(cf. `www.unesco.org/new/en/social-and-human-sciences/themes/anti-doping/fund-for-the-elimination-of-doping-in-sport/projects/`, last access: July 1, 2014). The European Commission and its member states are currently developing an anti-doping law based on the view that doping is "seriously undermining the principles of open and fair competition" (cf.
ec.europa.eu/sport/policy/societal_role/doping_en.htm, last access: July, 2014).

[6] E.g. Striegel et al. (2010) found an eight times higher number of drug abuse than it is officially confirmed.

Haugen, 2004; Eber, 2008).[7] Extending this approach, game theory is also used to analyze the interaction between athletes and an organization which decides upon conducting doping tests. This is usually modeled as an inspection game (Berentsen et al., 2008; Kirstein, 2012).[8] In an inspection game, there is no pure strategy Nash equilibrium because athletes want to dope without being detected, while the control organization tries to detect doping without testing clean athletes. Thus, mixed strategy equilibria, respectively, perfect Bayesian equilibria are used, which predict an intermediate level of doping.

We build on the previous game-theoretic work on doping but take the analyzes one level further by introducing *customers* as an additional player into the game. Customers are highly important because they finally make professional sports economically viable. Consider a sports event from which customers turn away their interest. This event does not only suffer of less ticket and merchandise revenues, it will also become less attractive for media companies who report from the event and for companies who sponsor the event. In Appendix 5.A1, we present several pieces of evidence for the importance of customers. In particular, the recent history of the Tour de France, the world's most famous cycling race, suggests that the reaction to the disclosures of systematic doping practices is the *withdrawal of support* from several stakeholders. For many other disciplines and events, this scenario has not happened, but it seems to be always present as a *threat*. Importantly, already the threat of withdrawing support is sufficient to significantly affect the incentives to dope, as we show in this paper. Despite their essential role, previous studies (to the best of our knowledge) have not included customers as a player in an inspection game. This paper closes this gap and explicitly analyzes the *role of customers* for the incentives to dope (respectively to cheat in a different context).

In our model, customers support a sports event as long as there is no doping scandal. After a scandal we assume that customers would withdraw their support (and contrast this case from the benchmark case of non-critical customers who always keep supporting). One might conjecture that the behavior of critical customers induces incentives for organizers and athletes to avoid doping since this increases the costs of doping for both athletes and organizers. However, our analyzes reveals that the opposite is true: Under mild assumptions, the unique outcome of the game is that athletes dope, while organizers make insufficient effort to test them. Because our assumptions are very parsimonious, this result is robust against many changes in the specification of utility. The intuition is simply that customers who can withdraw their support constitute a threat to the organizers such that they avoid uncovering (the full extent of) doping.

We then investigate how to change the institutions in order to support a doping-free equilibrium. It turns out that establishing transparency serves this purpose: if customers can observe whether there were serious doping tests, even if they turned out to be negative, then there is a doping-free equilibrium. However, this equilibrium is not unique – there is still an equilibrium involving doping. To rule out all doping equilibria it would be necessary to have a different kind of customer behavior, not only different institutions. We discuss the real world predictions of

[7] Interaction among heterogeneous athletes is analyzed by Berentsen (2002), Berentsen and Lengwiler (2004), and Kräkel (2007).

[8] The fact that in the inspection game there is only one athlete does not mean that the ideas from the strategic interaction between athletes are neglected. In particular, it is assumed that under no controls athletes prefer to dope, which is based on considerations of competition among athletes.

this model and the practical implications of its results for currently debated anti-doping policies.

The remainder of the paper is organized as follows: Section 5.2 presents the model. Section 5.3 establishes the main results, thereby characterizing the doping equilibrium. Section 5.4 studies a change of institutions that admits a doping-free equilibrium. In Section 5.5, we conclude.

5.2 Model

Considering benefits from professional sports, there is a large set of stakeholders: sports associations, team sponsors, event organizers, event sponsors, media, spectators, anti-doping agencies, doctors, politicians, etc. In our model, we restrict attention to three types of players: Athletes, Organizers, and Customers. Athletes can decide between doping and staying clean, whereas doping is defined as the use of illicit substances or methods.[9]

In our model, Organizers represent those actors who decide whether to conduct serious doping tests or not. Thus, testing stands for systematically attempting to detect and punish doped Athletes. An Organizer in that sense is the world anti-doping agency WADA.[10] In several disciplines, the national anti-doping agencies (NADAs) have a major role in organizing doping tests of their athletes. In other disciplines, the sports associations or the event organizers are the key players in organizing systematic doping tests.[11] Consequently, Organizers in our model represent anti-doping agencies, as well as organizers of sports events and sports associations. Indeed, anti-doping agencies are not independent of these organizations (Eber, 2002; Preston and Szymanski, 2003); and with respect to the decision we study here, they have similar interests or they can simply not conduct serious tests without the collaboration of the event organizers or the sports associations.[12]

Customers can decide upon staying a supporter or to withdraw support, e.g. not to continue watching an event on TV, not to further buy merchandise products, or to quit a membership in a club of supporters. Besides spectators, we can also subsume sponsors and the media (who broadcast or report about the sport events) under the term Customers. A withdrawal of each of these three actors can trigger the withdrawal of the two others. Sport events cannot survive without sponsors, withdrawal of the media restricts the access of the customers, and finally sport is only attractive for sponsors as long as there are customers. To make the arguments as clear as possible we focus on one representative customer and we also study only one representative organizer and one representative athlete (such that the strategic interaction between athletes is only presented in a highly reduced form). The extension to multiple players of a type would not qualitatively affect the results, but it would affect the ease of illustration. Therefore, we interpret the behavior of a representative player as the behavior of the Athletes, the Organizers,

[9] The definition of doping is itself an issue that is worth a discourse (cf. Eber, 2006). The binary decision to dope or not to dope is a simplification of a set of decisions which might also be considered as gradual. The simplification can be justified by at least two reasons. First, it is often unambiguous whether an athlete uses illicit substances or not. Second, there is a subjective interpretation of whether the athlete considers that he/she cheats or not.

[10] The WADA is an international institution founded in 1999 in Lausanne. Its main task is the world-wide coordination of anti-doping activities such as detection, deterrence and prevention. Moreover, the WADA coordinates doping tests with national anti-doping agencies (NADAs).

[11] For a richer description of the institutional setting see Emrich and Pierdzioch (2013).

[12] Eber (2002) suspects that even the WADA is not independent: "The problem is that WADA [...] is a product of the IOC [International Olympic Committee] and is probably far from being independent of it."

and the Customers (in plural).

The timing of the players' actions is as illustrated in Figure 5.1. First, Athletes decide on doping, then Organizers decide on testing, and finally Customers decide upon staying. The information set of the Organizers indicates that they do not observe the action of Athletes. Thus, the moves of Athletes and Organizers can also be considered as being simultaneously.

In our model, testing means that a doped Athlete is detected and punished. If the history in this first stage is $(Dope, Test)$ we call it a "scandal." All other histories, i.e. $(Dope, Notest)$, $(Clean, Test)$, $(Clean, Notest)$, are no scandal. Since doping tests and their outcomes are not transparent to the public, Customers cannot distinguish between the three possible histories if there was no scandal.[13] This is captured by the information set consisting of three nodes. As Figure 5.1 shows, this game has eight potential outcomes, which we label in the following way: d-t-s, d-t-l, d-n-s, d-n-l and c-t-s, c-t-l, c-n-s, c-n-l as also illustrated in Figure 5.1. The depicted payoff vectors are in the order Athletes, Organizers, Customers and only present one possible example.[14] While Athletes and Organizers have two strategies each $\{Dope, Clean\}$, respectively $\{Test, Notest\}$, Customers can choose between two actions in two information sets, which yields four strategies. We denote them by $\{SS, SL, LS, LL\}$, where, for instance, LS stands for action *Leave* in the first information set (after a scandal) and action *Stay* in the second information set (after no scandal). The wording 'leave' is a bit strong in the sense that it is not necessary that Customer support is fully lost with this action, but only that it becomes significantly smaller compared with the action *Stay*.

5.3 The Doping Equilibrium

In our analyzes, we focus on pure strategies and employ the notion of subgame perfect Nash equilibrium (SPNE). When the extension to behavioral strategies, where agents can continuously mix between actions, and the refinement of perfect Bayesian equilibrium (PBE) yield different results, we make this explicit. We will introduce assumptions on the players' preferences step-by-step to clarify that mild assumptions are sufficient for some results (equilibrium), while stronger assumptions are needed for others (inefficiency).

5.3.1 Existence of a Doping Equilibrium

We are most interested in the kind of Customers who withdraw their support after a scandal but not otherwise. This idea is covered by Assumption A1 which makes mild assumptions on the preferences of all the players.

[13] The fact that sometimes sport events publicly announce the number of tests they have carried through does not contradict this assumption. Still, Customers do not know whether the Athletes have been seriously and systematically tested.

[14] The specification of explicit payoffs or utility levels forces us to make many assumptions that are not at all necessary for the derivations of the model implications. The set of assumptions we will really use leaves room for many preference orderings and only one of them is represented by the example payoffs in Figure 5.1. The advantage of such a parsimonious approach is that eventually derived results are robust against changes of specification details.

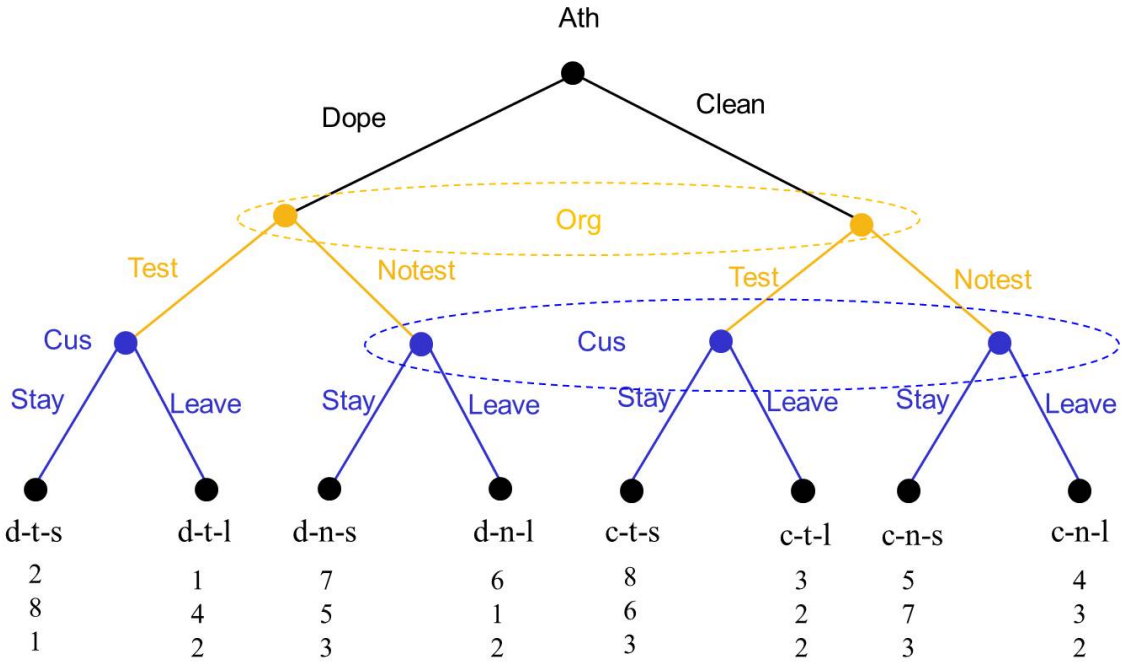


Figure 5.1: Structure of the game and an example for payoffs.

Assumption 1. *For the players' preferences we assume the following:*

Ath: *d-n-s* $\succ^{Ath}$ *c-n-s, i.e. Athletes prefer to dope if not tested; and c-t-s* $\succ^{Ath}$ *d-t-l, i.e. Athletes prefer to be clean and tested, while Customers stay, over being doped and tested, while Customers leave.*

Org: *d-n-s* $\succ^{Org}$ *d-t-l, i.e. a scandal combined with the loss of Customers is worse for the Organizers than undetected doping where Customers stay; and c-t-s* $\succ^{Org}$ *c-n-l, i.e. testing clean Athletes with Customers support is better for the Organizers than not testing clean Athletes when Customers leave.*

Cus: *d-t-l* $\succ^{Cus}$ *d-t-s, i.e. Customers prefer to withdraw support after a scandal; and d-n-s* $\succ^{Cus}$ *d-n-l, c-t-s* $\succ^{Cus}$ *c-t-l, and c-n-s* $\succ^{Cus}$ *c-n-l, i.e. Customers prefer to stay if there is no scandal.*

The Assumptions A1 are easy to justify. The assumption that Athletes dope if there are no tests follows from the standard assumption in the literature that the benefits of doping exceed the costs, even if there were tests (e.g. Maennig, 2002).[15] Organizers might existentially depend on Customers' support such that they would probably prefer any outcome where Customers stay (i.e. d-t-s, d-n-s, c-t-s, and c-n-s) over any outcome where support is withdrawn (i.e. d-t-l, d-n-l, c-t-l, and c-n-l). This also means that testing is not too expensive in the sense that the withdrawal of customer support is worse than conducting tests. This assumption needs not be satisfied for sport events that do not belong to professional sports. The preference of the Customers to leave after a scandal means that they are bothered by doping scandals, rather

[15] This interaction between several Athletes is often modeled as a prisoner's dilemma. There the dominant strategy is to dope, as we assume this behavior here for the case of no tests and one representative Athlete. In reality, there are also Athletes who are unconditional non-dopers (Pitsch et al., 2010). Their (trivial) behavior is not studied within our model.

than enjoying them.[16] Finally, Customers' preferences to stay if there was no scandal reflect the general interest in sports based on the view that Customers are unable to distinguish between undetected doping and clean sport even ex post. That is, their payoff of staying a supporter does not depend on whether d-n-s, c-t-s, or c-n-s is reached because they cannot distinguish between them.[17] And similarly, their payoff of withdrawing their support would not depend on whether outcome d-n-l, c-t-l, or c-n-l is reached.[18] Given A1 a Customer will stay a supporter if and only if there was no scandal – the behavior under scrutiny. The following proposition shows that then outcome d-n-s – i.e. Athletes dope, Organizers do not test, and Customers stay supporters – is an equilibrium outcome.

Proposition 1 (doping equilibrium)**.** *Under Assumptions A1* $s^* := (Dope, Notest, LS)$ *is a SPNE.*

The proofs of this and all other propositions are collected in Appendix 5.A2. The intuition for Proposition 1 becomes apparent when considering the strategic interaction between Athletes and Organizers, given the Customers' behavior. Using the example payoffs from Figure 5.1, the following Matrix (5.1) is induced by Customers who stay if and only if there is no scandal. This can be contrasted with Matrix (5.2) that is obtained in the benchmark case that Customers unconditionally stay.

		Org	
		Test	Notest
Ath	Dope	1,4	**7,5**
	Clean	8,6	5,7

(5.1)

		Org	
		Test	Notest
Ath	Dope	2,8	7,5
	Clean	8,6	5,7

(5.2)

In the benchmark case, best response dynamics always follow a cycle, as it can be seen from Matrix (5.2). This is the classic observation in the inspection game that there is no pure strategy Nash equilibrium. In mixed strategies there would be an equilibrium where the probability of doping for our example payoffs is one over four. The strategic interaction in our model only differs from the benchmark case concerning the payoff in the upper left matrix entry, which is due to customers who leave after a scandal. As it can be seen from Matrix (5.1), this breaks the cycle of deviations (in the best response dynamics) and yields the equilibrium in pure strategies established by Proposition 1. In words, Customers who leave after a scandal establish a threat

[16] In reality there might be Customers who enjoy (doping) scandals. We will consider such customers and, equivalently, uncritical customers, who always stay supporters, as a benchmark later on. However, we study a more critical kind of Customers here.

[17] Basically, this assumption also means that Customers do not respond to what they infer about the behavior of other players. Alternatively, we could assume that Customers also withdraw their support in absence of positive doping tests, if they infer the use of doping by analyzing the situation of strategic interaction. This alternative assumption and its implications are discussed in Subsection 5.3.2.

[18] A similar interpretation holds if we consider sponsors and media companies in the role of the Customers. Moreover, there is a second interpretation of this assumption for these actors. It might be that they are able to distinguish ex post between different outcomes, but do not strongly care about doping as long as it is not officially detected.

to the Organizers such that they prefer not to detect doped athletes, even if they had done so in case that the Customers were uncritical (i.e. in the benchmark case where the Customers always stayed). Thus, the explanation for our qualitatively new result is that the introduction of (critical) Customers undermines the Organizers' incentives to uncover (the full extent of) doping because Organizers anticipate that they would suffer losses in the case of scandals. As a consequence, Athletes are not seriously tested and therefore decide to dope.[19]

Next, we discuss the robustness (and the practical implications) of the finding.

5.3.2 Robustness of the Doping Equilibrium

Proposition 1 only serves as a clear empirical prediction if there are no other equilibria and if its statement is robust to specification details. We first address uniqueness of the doping equilibrium and then study its robustness with respect to a continuous (instead of binary) action space, an imperfect test technology, and a different type of Customer.

Uniqueness Concerning uniqueness, we show that assumptions that are standardly made in inspection games are sufficient to exclude other equilibria.[20] These assumptions are collected in A2.[21]

Assumption 2 (inspection)**.** *In the inspection game the following assumptions are made on the preferences of Athletes and Organizers:*

Ath: $c\text{-}t\text{-}s \succ^{Ath} d\text{-}t\text{-}s$ *and* $c\text{-}t\text{-}l \succ^{Ath} d\text{-}t\text{-}l$*, i.e. Athletes prefer not to dope if there are tests; and* $d\text{-}n\text{-}s \succ^{Ath} c\text{-}n\text{-}s$ *and* $d\text{-}n\text{-}l \succ^{Ath} c\text{-}n\text{-}l$*, i.e. Athletes prefer to dope if there are no tests.*

Org: $d\text{-}t\text{-}s \succ^{Org} d\text{-}n\text{-}s$ *and* $d\text{-}t\text{-}l \succ^{Org} d\text{-}n\text{-}l$*, i.e. Organizers prefer to test the Athletes if they are doped; and* $c\text{-}n\text{-}s \succ^{Org} c\text{-}t\text{-}s$ *and* $c\text{-}n\text{-}l \succ^{Org} c\text{-}t\text{-}l$*, i.e. Organizers prefer not to test if Athletes are clean.*

The Assumptions A2 are partially redundant with Assumptions A1, but further specify that Athletes prefer not to dope if tested and that Organizers prefer to test if and only if Athletes are doped. This reflects that Organizers are willing to detect doping, while testing is costly. The example payoffs provided in Figure 5.1 satisfy both A2 and A1.

Proposition 2 shows that the mild Assumptions A1 and the standard Assumptions A2 are powerful enough to rule out any equilibrium besides the previously found doping equilibrium.

Proposition 2 (uniqueness)**.** *Suppose Assumptions A1 and A2 hold.*
Then $s^* = (Dope, Notest, LS)$ *is the unique SPNE.*

---

[19] This result is not due to other explanatory factors since under our assumptions testing can be almost costless, the benefits of doping need not be high, and the disutility of being detected can be huge. Importantly, our argument is not that doped Athletes produce higher performances which creates utility for the Customers or Organizers, although this idea would not alter the result.

[20] In an inspection game an inspectee has to decide whether to comply or deviate from a norm, while an inspector can choose between inspecting or not inspecting the action. To embed this standard game into our notation we would consider the Athlete as the inspectee, the Organizer as the inspector and for the Customers which are standardly excluded, we would assume constant behavior. That is, our model differs from the standard inspection game only in that Customers sometimes withdraw support, while standardly Customers always stay supporter or, alternatively, they never support.

[21] Usually, the inspection game is represented by numerical payoffs. This implies additional assumptions to the ones collected here. However, those additional assumptions are neither consensual in the literature, nor are they necessary for our results (as long as we obtain pure strategy equilibria).

Next, we address robustness of this result against three natural variations of the model.

Continuous Actions In reality customers might decide on the extent to which they still support, which is a richer action space than just the binary choice of *Stay* or *Leave*.

In order to relax the assumptions of binary actions for each player, we consider behavioral strategies. Under Assumptions A1 and A2, however, this does not affect the result. In equilibrium, Athletes dope with probability one, Organizers test with probability zero and Customers certainly leave after a scandal and stay in the absence of a scandal.

If Customers *preferred* to partially reduce their support after a scandal, the question is whether the reduction is negligible such that we are in the benchmark case or whether the reduction is significant such as in our model where Organizers try to avoid scandals (cf. Assumptions A1: d-n-s $\succ^{Org}$ d-t-l). In the former case we would have an equilibrium in which players randomize, in the latter case the doping result holds.

Imperfect Test Technology Unrealistically, we have assumed that the test technology is free of errors. Extending our game to allow for false-positive and false-negative tests which occur with some probability ε, leads to a more realistic model, but not to a different result. As it can be shown using the example payoffs, the unique SPNE is that Athletes dope, Organizers do not test, and Customers stay in the absence of a scandal as long as $\varepsilon < \frac{1}{2}$.[22]

Sophisticated Customers A crucial assumption throughout our analyzes is that customers are unable to distinguish between undetected doping and clean sport even ex post such that they prefer outcome d-n-s over d-n-l. Our motivation is not the literal (game-theoretic) interpretation that Customers infer that there must be a high level of doping but do not care as long as there is no scandal (even if this might be true for some media companies or sponsors who we also consider in the role of Customers). Rather, we consider less sophisticated Customers who do not draw these inferences and therefore stay supporters in the absence of positive doping tests, which is arguably much more realistic than Customers who leave in that case. For example, most football fans do not seem to be trying to infer the underlying level of doping and to turn away their interest from competitions where it can be suspected that there are insufficient doping controls. However, it is a game-theoretically natural and economically interesting exercise to consider the effects of sophisticated Customers who make the inferences by analyzing the situation of strategic interaction and react to their belief about doping.[23] Let us briefly elaborate on this alternative (hypothetical) model, which is obtained when reducing the payoff of Customers for outcome d-n-s sufficiently to let d-n-l be preferred. As we show in appendix 5.A3.1, there are no pure strategy SPNE in that model. Customers still leave after a scandal, but all other equilibrium choices are mixed actions. The equilibrium belief of customers is that if there is no scandal, then the probability of doping is exactly fifty percent such that Customers are indifferent between staying and leaving. The probability of doping in equilibrium is $p^* \approx 59\%$. This is smaller than in our model with naïve agents, in which the pure strategy

[22] The issue of imperfect test technology in doping tests is investigated by Kirstein (2012). He studies a game, in which the enforcing agency receives an informative but imperfect signal about whether an athlete is doped or not.

[23] We thank an anonymous referee for this suggestion.

equilibrium predicts that all athletes (who are calculating dopers) dope. However, compared with the benchmark case, where Customers stay unconditionally (which we also refer to as the inspection game), sophisticated Customers lead to an increase in the probability of doping from 25% to 59%. Thus, the qualitative result, that the introduction of Customers to the inspection game increases the level of doping, holds for both naïve and sophisticated Customers. The computed fractions of dopers, of course, depend on the absolute payoffs of the example and suggest that they have some cardinal interpretation.

In sum, it is a robust finding that the presence of customers who might withdraw their support accentuates the extent of doping. Let us now briefly discuss the interpretation and implications of this result.

5.3.3 Discussion of the Doping Equilibrium

The real world prediction of our simple model is that the number of dopers is large, while the probability of a doped athlete to be caught and punished is close to zero. The real extent of doping within professional sports is hard to assess and thus remains highly controversial. Theoretically, there are strong incentives to use performance-enhancing drugs. In particular, if our second prediction holds – that the probability of being detected and punished is small.[24] The explanation that our model provides for doping is that organizers do not want to uncover the full extent of doping because they anticipate that they would suffer losses in the case of scandals.

Our argument that Organizers lack the incentives for serious doping tests is in line with Eber (2002) who argues that Organizers have a low effort bias, which becomes stronger the more the authorities weight the economic stakes of professional sport.[25] Within his model, athletes form rational expectations about the effort of authorities to prevent doping, which leads to a credibility problem of the Organizers (Eber, 2002).[26]

The prediction that Organizers do not seriously test is also empirically difficult to assess. However, there are several pieces of evidence that support this view. For example, consider the anti-doping instrument called world anti-doping code (WADC). This is an international regulatory system that specifies test procedures, and lists of forbidden substances, and accredits doping labs. (The WADC is an instrument of the world anti-doping agency WADA and we assume for the moment that the WADA is free of incentives issues in the fight of doping.) Implementing the WADC in some discipline would contribute to establishing a strict anti-doping regime. As it turns out, however, the problem of the WADC is the lack of compliance on the part of the international sports associations (Emrich and Pierdzioch, 2013). For example, the

[24] In the absence of serious controls, athletes are in the classic (prisoner's) dilemma because they either can get a competitive advantage by doping or they have to assume that their rivals are doped (e.g. Breivik, 1992).

[25] Of course, there are also other reasons, why detected doping leads to losses. For example, a national sports association might have an interest that athletes from its country are successful in international competition.

[26] Concerning Customers' perceptions, one way to increase the public credibility of anti-doping activities might be to detect doping cases but very few of them. Indeed, we have not included the idea that the conviction of a few athletes enhances the credibility of clean sport. We have focused on the main effect, which is that the conviction of many athletes undermines the credibility of a sports event or even of a whole discipline. Importantly, we are not arguing that Organizers are unwilling to fight against doping, but simply that they have strong incentives not to fully uncover doping activities.

following prominent sports associations are reported to refuse the WADC: the International Football Association (FIFA), the International Tennis Federation (ITF), and the International Cycling Association (UCI) (Emrich and Pierdzioch, 2013).[27] Another indication that there need not be serious doping tests although many efforts in the fight against doping are claimed is the charter formulated by a movement called "change cycling now." The movement consists of sports journalists, former cycling officials, as well as of former cyclists, including a Tour de France winner. The charter strongly requests that the organization responsible for doping tests becomes independent and thus indirectly accuses the current institution as not being so. The charter expresses this as a principle to create doping-free cycling in the future: "The responsibility for deciding who is tested, when they are tested, and what drugs they are tested for, must reside in an independent entity that is beyond the control of the UCI."[28] Thus, even in cycling, where there is a long list of detected dopers, it seems that the probability of being detected when doped is not that high. We argue that in any discipline there are incentives to put insufficient effort into the detection of dopers.

Let us now return to our model and discuss efficiency.

5.3.4 Pareto Efficiency

Proposition 2 shows that the unique equilibrium outcome is d-n-s, which means that doping is prevalent. Whether this is a socially desirable outcome is not fully uncontroversial.[29] Let us discuss the assumptions that decide upon efficiency. In our model, the following assumption assures that d-n-s is indeed inefficient in the strong sense of being Pareto dominated.

Assumption 3. *For the preferences of the three players we assume the following:*

Ath: *c-t-s* $\succ^{Ath}$ *d-n-s* $\succ^{Ath}$ *d-t-s, i.e. Athletes prefer being tested and clean over being not tested when doped over being tested and doped.*

Org: *c-t-s* $\succ^{Org}$ *d-n-s, i.e. Organizers prefer the testing of clean Athletes over not testing doped Athletes.*

Cus: *c-t-s* $\succeq^{Cus}$ *d-n-s, i.e. Customers weakly prefer tested clean Athletes over not tested doped Athletes.*

Note that Assumptions A1, A2, and A3 are mutually consistent, e.g. the example payoffs of Figure 5.1 satisfy all three assumptions. The Assumptions A3 are plausible, but arguably much more controversial than A1 and A2. Athletes might dislike doping tests even if they are clean, because they have to be constantly available. However, we assume that Athletes are better off by being tested and clean than being doped, e.g. because doping would seriously affect their health. Organizers might have high costs of conducting doping tests and they might benefit from the performance of doped Athletes such that we had d-n-s $\succ^{Org}$ c-t-s. However, we take the view of

[27] The WADA does not have effective instruments to punish organizations that do not comply.

[28] The full charter can be found at
`www.changecyclingnow.org/wp-content/uploads/2012/12/Charter-of-the-Willing.pdf`, last access: July 30, 2014.

[29] Savulescu et al. (2004) discuss several arguments concerning the usefulness of anti-doping rules and conclude that performance-enhancing drugs should be legalized. Concerning fairness, they find that legalization is in line with the "spirit of sport" because it is still the aim to find the best athlete among all competitors.

benevolent Organizers who prefer to detect doped Athletes (as long as Customers stay) such that the relation is just the opposite. Finally, for Customers we keep the view that they cannot distinguish between the outcomes that do not include a scandal. Thus, d-n-s $\sim^{Cus}$ c-t-s $\sim^{Cus}$ c-n-s.

Clearly, under Assumptions A3, outcome d-n-s is Pareto dominated by outcome c-t-s. Thus, the unique equilibrium outcome in our model is not Pareto efficient. Outcome c-t-s, however, is not Pareto dominated by any other outcome as established by Proposition 3.

Proposition 3 (Pareto efficiency)**.** *Suppose Assumptions A1 and A3 hold. Then outcome d-n-s is not Pareto efficient, while outcome c-t-s is.*

In this subsection, we have shown that we are indeed in a social dilemma situation. The unique equilibrium outcome, which involves doping, is Pareto dominated by a doping-free outcome. The next question is how the institutions can be changed such that the Pareto efficient outcome c-t-s becomes an equilibrium outcome. If the controversial assumption A3 is not accepted, then the doping equilibrium need not be Pareto dominated. Still, however, it is of high interest to find conditions for a doping-free equilibrium.

5.4 Inducing a Doping-free Equilibrium

We first establish the results, then we discuss current policy suggestions in the light of the model.

5.4.1 Change of Customers' Information Structure

In order to induce an outcome without doping, we change the information structure in the game. In particular, we let the Customers be also informed about doping tests that turned out to be negative. Consider the extensive game tree illustrated in Figure 5.2.

As before, Organizers decide on testing the Athletes without observing whether there was doping or not. The Customers then decide upon staying a supporter or leaving. The information they have for this decision now consists of three information sets: one is after a scandal $(Dope, Test)$, one after a negative test $(Clean, Test)$, and one after no test, which consists of the two histories $(Dope, Notest)$ and $(Clean, Notest)$. This yields eight strategies for the Customers, which we denote by $\{SSS, SSL, SLS, SLL, LSS, LSL, LLS, LLL\}$, such that the first letter stands for the action after a scandal, the second letter for the action after a negative test, and the third letter for the action if there were no tests. (The example payoffs in Figure 5.2 are as in Figure 5.1.)

In the game with less transparency (studied in the former section), under Assumptions A1 and A2 the unique equilibrium outcome involved doping. The following proposition shows that with more transparency there is a doping-free equilibrium, as well.

Proposition 4 (doping-free equilibrium)**.** *Under Assumptions A1 and A2 there are two SPNE in the game with finer information structure:* $\hat{s} := (Clean, Test, LSL)$ *and* $s^{**} := (Dope, Notest, LSS)$.

Proposition 4 shows that a change in the information structure in our model is sufficient to obtain a doping-free equilibrium. *Thus, the social dilemma can be overcome by establishing*

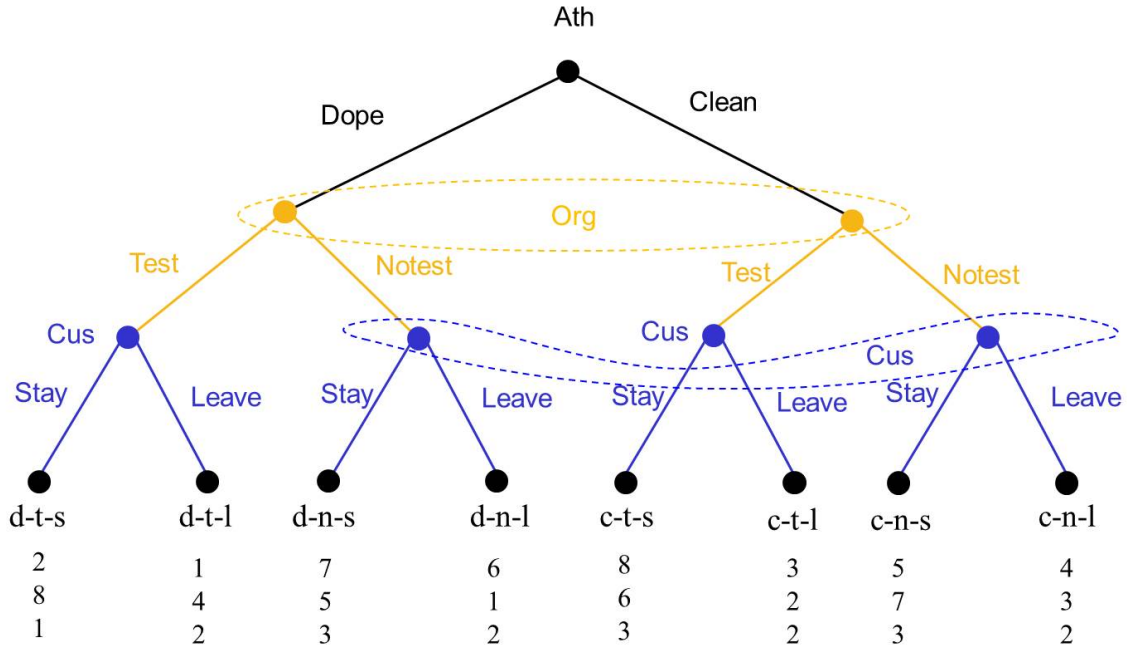


Figure 5.2: Structure of the game with well-informed Customers and an example for payoffs.

transparency. The intuition for this result can be gained from the interaction of Athletes and Organizers, given Customers who play LSL. This is represented in Matrix (5.3) using the example payoffs. Organizers do test, given that they lose Customers in the absence of tests.

		Org	
		Test	Notest
Ath	Dope	1,4	6,1
	Clean	**8,6**	4,3

(5.3)

Considering behavioral strategies there is a continuum of equilibria in which Athletes are clean, Organizers test, and Customers stay after no tests with probability $r^* \leq \frac{3}{4}$ for the example payoffs. Thus, there are doping-free equilibria although the probability that Customers leave in the absence of doping tests might be low.

However, the doping-free equilibria come with (at least) two caveats. First, they involve suboptimal behavior outside the equilibrium path. Indeed, after no test, the equilibrium strategy of the customers implies to leave (with positive probability), although this is not in line with Assumptions A1.[30] Second, there is still another equilibrium which involves doping.

The two issues would be solved at once, if Customers had different preferences. Suppose, hypothetically, that Customers were more skeptical about doping practices and therefore insisted on the proof of clean sports in order to stay supporters. With such Customers, the doping-free equilibrium $\hat{s}$ was unique, as we show in Appendix 5.A3.2. Moreover, there would be no more issue of suboptimal behavior outside the equilibrium path because the Customers' threat to leave after no tests would then be credible.

[30] Subgame perfection simply does not rule out this incredible threat. The notion of perfect Bayesian equilibrium would do so and only render s^{**} as an equilibrium.

It thus not only takes a better information level for the Customers but also a change in preferences: they would have to insist on doping tests in order to unambiguously induce incentives for a doping-free sport.

5.4.2 Implications for Anti-Doping Policies

In the literature on doping incentives various approaches are suggested to solve the doping issue. Many of them concern the change of incentives on part of the athletes. On the one hand, it is suggested to change the punishments or to increase the fines for being doped (e.g. Haugen, 2004). In the light of our model, however, this approach is not effective since in equilibrium athletes are not tested and thus do not get punished. On the other hand, the suggestion is to decrease the benefits of doping, e.g. by reducing the prize spread between different ranks or by reducing the number of competitions (Eber and Thépot, 1999). But also decreasing the benefits of doping only affects the behavior of athletes if it succeeds in making doping less attractive than not doping, (i.e. the payoff of doping must be reduced to such an extent that the ordinal preference that we assume in the model switches direction). This seems to be at least questionable.

Thus, for Athletes, which are calculating dopers, any anti-doping instrument has to make sure that the probability that doping is punished is sufficiently high. In this paper, we have identified the lack of the Organizers' incentives to really implement such a regime. A rather radical solution to these misguided incentives is to replace the actors that are responsible for doping tests. Indeed, it is currently debated in several countries (among them Germany) whether to establish a legislation that makes the state and its body responsible for the prosecution of dopers.[31] In some states, e.g. Belgium, this is already implemented. In principle, the proposed shift of responsibility is a solution to the lack of control since the police and the courts do not have the conflict of interest that NADAs and sports associations have. However, this approach is only fruitful if it is practically possible to fully circumvent the Organizers, i.e. if the collaboration of sports associations and NADAs is not crucial for the prosecution of doped athletes.

In subsection 5.4, we have elaborated on a different approach to fight doping. We show how Customers can contribute to doping-free sports if they are sufficiently well-informed. In particular, we require information about doping tests which admits Customers to condition their support for the sports event on the presence of doping tests (as illustrated in Figure 5.2). Whether or not Customers really insist on doping tests, then determines the extent of doping in equilibrium (Proposition 4 and 5.A3.2). Thus, a direct implication of our model is that *transparency* about the doping tests and their outcomes should be established.

This requirement is not satisfied in professional sports today. Most of the data that is publicly available only contains cases of detected doping but not information about the extent of testing. For example, the Internet Anti-Doping Database created by Norwegian sports journalist Trond Husø contains more than 5,000 cases, but mostly of detected dopers.[32] In the absence of doping scandals, this does not allow Customers to discriminate between clean sports and undetected doping (such as illustrated in Figure 5.1).

One type of actors who is in principle capable of establishing transparency are sports associations who we study as Organizers in our model. However, as argued above, such organizations

[31] The discussion caught new fire with the recent case of Lance Armstrong.
[32] Cf. `www.dopinglist.com`.

lack incentives to do so. Dilger and Tolsdorf (2004) and Striegel et al. (2010) assume that their lack of compliance is one reason why data on doping is so limited. In order to achieve more transparency, the WADA could open the access to their database called ADAMS. ADAMS was introduced to simplify the organization and realization of doping tests.[33] Currently, only certain actors of the immediate sports environment are allowed to use ADAMS. Opening the access to ADAMS seems to be a cheap way to establish transparency, while such a policy might involve several new issues, including the violation of privacy rights. Moreover, it can be difficult or costly to understand and interpret the data for Customers. A much simpler suggestion is that the WADA makes public to which extent sports associations and NADAs comply to anti-doping standards. This could be a simple rating which gives Customers a clear signal about which disciplines and events are credible in their fight against doping. Of course, this requires independence on part of the WADA, which is also doubted (cf. Eber, 2002; Preston and Szymanski, 2003), but, in principle, we conclude that there should be an independent rating or certifying agency that officially measures to which extent certain sports events have implemented an anti-doping regime. Whether or not doping prevails in the future is then dependent on the Customers' preferences.

5.5 Concluding Remarks

In this paper we have extended the inspection game (e.g. Avenhaus et al., 2002) by a third player: customers, who can withdraw their support. As it is shown in the application of doping in professional sports, the behavior of critical customers accentuates the fraudulent behavior.[34] Customers who are ready to leave after a doping scandal, undermine the organizers' incentives to test athletes on performance-enhancing drugs and to convict them on doping. As a consequence, athletes have stronger incentives to dope although this need not be in the best interest of any of the three types of players. Our analyzes substantially strengthens the argument already outlined by Eber (2002, p.95) who comes to the following conclusion: the institution responsible for doping controls "may have some temptations to slacken its antidoping effort when confronted with doping affairs to preserve the economic value of the shows (e.g. the Olympic Games organized by the IOC [International Olympic Committee]). Knowing that, athletes may rationally not believe in strong antidoping policies and may then continue to choose high levels of doping." Our analyzes of incentives suggests that the few spectacular cases of convicted dopers are not delinquent exceptions, but rather unlucky cheaters or scapegoats, because the probability of being detected when doped is low (cf. Preston and Szymanski, 2003). To elaborate on potential solutions for the doping dilemma, we show that a change in the information structure in our model serves to obtain a doping-free equilibrium (Proposition 4). The crucial change is to

[33] ADAMS has four main tasks: First, athletes are required to enter their actual wherabouts and other users will be informed about actual infringements against reporting standards (Athlete's Wherabouts). Second, it is also possible to manage medical exceptional permissions (Therapeutic Use Exemptions Management). Third, ADAMS informs about doping tests, infringements, and sentences (Information Clearing House). Finally, ADAMS is supposed to ease the scheduling of doping tests and the preparation of doping profiles (Doping Control Platform).

[34] Other counter-intuitive results of the inspection game are already known (Holler, 1993; Andreozzi, 2004; Friehe, 2008). They concern the indifference of the mixed strategy Nash equilibrium, which implies that a change of payoffs for one player does not affect the equilibrium behavior of this player, but only its opponent's. Maximin strategies are used to address this issue (cf. Aumann and Maschler, 1972; Holler, 1990).

establish transparency in the sense that customers know whether there were negative tests or there were no serious tests (cf. Figure 5.2 versus Figure 5.1). This allows customers and other stakeholders to condition their support on the presence of serious anti-doping tests. Practically, the required transparency could be established by a certificate or rating that shows which sports events have established a strict anti-doping regime.

However, our model is not restricted to doping and professional sports. In many different industries, e.g. textile or food, customers do not know very well the production process of the goods that they consume. In particular, it is hidden whether the producing companies complied to all standards and ethical norms – except if there is a scandal in the news. Scandals make public, e.g. the use of child labor in the production of clothes, as well as the violation of hygienic standards in the food industry. After the detection of such fraudulent activities in some organization, there is a loss of reputation and critical customers may react with a boycott. There are not few contexts, where the agent that is able to detect the potential fraudulent activities is also affected by such a scandal. Consider a company in the role of the Organizer, who has business relations with another firm (Athlete) that does potentially not comply to certain ethical standards. Detecting norm violations would also undermine the reputation of the company itself. Customers who react with a boycott substantially increase the loss of the company and thereby undermine its incentives to uncover (potential) scandals.[35] When there is no other agent who is capable of detecting the fraud without the help of the company, the number of fraudulent activities might even increase. As our model shows, this outcome can be altered if customers are informed about control activities of all companies by some independent institution. Thus, transparency is necessary in order to overcome this type of social dilemma.

5.A1 Some Evidence on the Importance of Customers

Customers of a sports event do not only expect high performances from the athletes but also their compliance to the rules. During the Olympic Games in Barcelona 1992, for example, 91% of 475 interviewed spectators answered that they want to see high performances at the Olympic Games, but a majority of them (58%) considers doping as a *threat* to the Games (Messing and Müller, 1996). For the Olympic Games of Sydney 2000 and Athens 2004 the number of people who agrees that doping is a threat has even increased to 69% and 82% (Messing et al., 2008) and doping is considered as the most severe threat for Olympic Games, ranking above terrorism and corruption (Messing et al., 2004). This view is not restricted to spectators, but it is also predominantly shared by athletes, students of sports science, and media representatives (Tröger, 2006). But what is the actual "threat" that starts out from doping in sports? Probably such a scenario can be best studied in an event for which it is known that doping is widespread – such as the world's most famous cycling tour, the Tour de France.

The recent exposure of the doping affair concerning the seven-times Tour de France winner (Lance Armstrong) is just one very spectacular case in a long list of disclosures. In 1998 a whole cycling team (Festina) was excluded from the Tour de France after a large amount of

[35] There is empirical evidence on a similar issue in the context of juridical judgments: An increase in the defined punishment, e.g. from prison sentence to capital punishment, can lead to a reduction of the number of convictions.

performance-enhancing drugs was found in a team car. In the 2006 Tour de France, an affair centered on a physician (Eufemiano Fuentes) led to the expulsion of several participants and some days after the Tour de France 2006, it was detected that the winner (Floyd Landis) was positively tested on performance-enhancing drugs. The fact that, in this case, as well as in many other prominent cases, doping delicti became public after the Tour de France, implies that the customers' reaction to the scandal cannot be simply measured by a change of the audience ratings during one Tour (Van Reeth, 2013). One year after the Fuentes affair, the German public-sector TV channel quit the live-broadcast of that actual Tour de France when a German cyclist (Patrik Sinkewitz) was convicted on doping. Although this TV channel reported from the Tour de France again in the years 2008 until 2011, they finally quit in 2012. The reason for that was a sharp decline in the audience ratings from one year to the next. (While the market share amounted to 13 percent in year 2008, there was a decline to approximately 9 percent in 2009.) Not only TV channels, also sponsors reacted with exit. For example, the cycling team-sponsor Phonak quit, after their team leader (Floyd Landis) was convicted of doping, and a German cycling team-sponsor (Gerolsteiner) quit after two German cyclists (Stefan Schumacher, Bernhard Kohl) were found guilty. A majority of fans supports such reactions of sponsors and TV broadcasters (Solberg et al., 2010). In sum, the recent history of cycling demonstrates that the reaction to the disclosures of systematic doping practices is the *withdrawal of support* from several stakeholders. This is true for media companies, sponsors, and – last but not least – customers (spectators). It is a notable fact that there are customers who still support the Tour de France despite (or maybe even because of) the doping scandals. However, it seems undeniable that the organizers of the event have suffered substantial losses due to the withdrawal of support of many customers, sponsors, and media companies.

Similar scenarios of withdrawal of support have not happened in most of the other disciplines. As the Olympic Games in London show, the interest in sports and, particularly, in track and field athletics is huge. This does not mean that track and field athletics is free of doping. For example, the US sprinter Justin Gatlin who sprinted to his personal best in London has a background on doping offenses. Further, the two nearest rivals (Tyson Gay and Asafa Powell) of the star in track and field athletics, Usain Bolt, were convicted on doping in 2013. These cases are not that exceptional: among 64 world class sprinters on the 100 meters track Dilger and Tolsdorf (2004) found that 16, i.e. 25%, have been convicted on doping somewhen in the period from 1997 until 2002. Also, the U.S. sports leagues for American Football and Baseball (NFL, MLB) have to deal with some doping scandals. For example, baseball star Alex Rodriguez was suspended for 211 matches until the end of season 2014 because of the suspicion that he consumed banned drugs.

It seems that, despite such cases, the public perception in many disciplines is that most of the athletes do not use performance-enhancing substances. For example, in the year 1988 the most prominent 100 meters track star Ben Johnson was convicted on doping, while during the next Olympic Games (in Barcelona 1992) only every fourth or fifth spectator (22%) agreed that doping and manipulation are determining factors of the Olympic performances (Messing and Müller, 1996). In professional tennis or soccer doping is rarely a topic at all.

Concerning the Tour de France, in contrast, most of the TV spectators (89%) in a survey

assumed that doping is a common practice.[36]

If the public perception of clean sport is critical for customers and other stakeholders to keep their support, then organizers have strong incentives to avoid a list of scandals comparable to the one of the Tour de France. Hence, the critical role of customers lies in their potential to withdraw support. This is exactly the aspect of customers that is incorporated in our model.

5.A2 Proofs

5.A2.1 Proof of Proposition 1

Proof. The only proper subgame of our game starts at node "scandal" $(Dope, Test)$. L is a Nash equilibrium (NE) in this trivial subgame. The second subgame is the game itself. Suppose Customers play LS in this game.

For the decisions of the Athletes and the Organizers LS induces the following matrix (5.4).

		Org	
		Test	Notest
Ath	Dope	d-t-l	**d-n-s**
	Clean	c-t-s	c-n-s

(5.4)

Now, it can be immediately observed that by applying A1, Athletes dope and Organizers do not test are mutual best responses since d-n-s $\succ^{Ath}$ c-n-s and d-n-s $\succ^{Org}$ d-t-l. Moreover, LS is a best response to $(Dope, Notest)$ because by A1 it holds that d-n-s $\succ^{Cus}$ d-n-l. □

5.A2.2 Proof of Proposition 2

Proof. $s^* = (Dope, Notest, LS)$ is a SPNE by Proposition 1. We show uniqueness of s^* by excluding all other strategy profiles from being an equilibrium. In the subgame that starts with the scandal, Customers choose leave in equilibrium (by A1). Thus, there are no SPNE where Customers play SS or SL. Given Customers play LL, there is no mutual best response for Organizers and Athletes because this is an inspection game situation (A2). Thus, only the four strategy profiles with Customers choosing LS remain. A1 excludes $(Dope, Test, LS)$ by d-n-s $\succ^{Org}$ d-t-l and $(Clean, Notest, LS)$ by d-n-s $\succ^{Ath}$ c-n-s from being an equilibrium. Finally, $(Clean, Test, LS)$ is not an equilibrium because c-n-s $\succ^{Org}$ c-t-s (A2). □

5.A2.3 Proof of Proposition 3

Proof. The implication that d-n-s is Pareto dominated by c-t-s is immediate from A3.

To establish that c-t-s is Pareto efficient, let us show that for any other outcome d-t-s, d-n-s,..., c-t-l, c-n-l there is at least one player who strictly prefers outcome c-t-s. From A3 we get: c-t-s $\succ^{Ath}$ d-n-s $\succ^{Ath}$ d-t-s. A3 and A1 imply that c-t-s $\succ^{Ath}$ d-n-s $\succ^{Ath}$ c-n-s. From A1 we get: c-t-s $\succ^{Ath}$ d-t-l. From A3 and A1 we get: c-t-s $\succ^{Cus}$ d-n-s $\succ^{Cus}$ d-n-l. From A1 we get: c-t-s $\succ^{Cus}$ c-t-l. Finally, from A1 we get: c-t-s $\succ^{Org}$ c-n-l. □

5.A2.4 Proof of Proposition 4

Proof. The game has two proper subgames: one starts at node $(Dope, Test)$ and one starts at node $(Clean, Test)$, cf. Figure 5.2. In both subgames only Customers act and by assumption A1 they will choose *Leave* in the first one and *Stay* in the second one. Thus, in each SPNE

[36] These figures are reported by a German newspaper and can be found at `www.zeit.de/online/2007/28/tour-de-france-medienkritik`, last access: July 30, 2014.

the Customers' strategy is either LSL or LSS. The following matrices show the decisions of Organizers and Athletes given that Customers choose LSL (Matrix 5.5) or LSS (Matrix 5.6):

		Org	
		Test	Notest
Ath	Dope	d-t-l	d-n-l
	Clean	**c-t-s**	c-n-l

(5.5)

		Org	
		Test	Notest
Ath	Dope	d-t-l	**d-n-s**
	Clean	c-t-s	c-n-s

(5.6)

Matrix (5.5) leads to mutual best replies $(Clean, Test)$. LSL is also a best reply to $(Clean, Test)$ because c-t-s $\succ^{Cus}$ c-t-l by A1 such that $\hat{s}$ is a SPNE. There are no other equilibria with LSL because A2 yields deviations from outcomes d-n-l and c-n-l, while d-t-l is not a candidate because, again, c-t-s $\succ^{Ath}$ d-t-l by A1.

Matrix (5.6) leads to mutual best replies $(Dope, Notest)$ (as already shown in proof of Proposition 2). LSS is also a best reply to $(Dope, Notest)$ because d-n-s $\succ^{Cus}$ d-n-l by A1. There are no other equilibria with LSS because A2 yields deviations from outcomes c-t-s and c-n-s, while d-t-l is not a candidate because c-t-s $\succ^{Ath}$ d-t-l by A1. □

5.A3 Model Variations

5.A3.1 Sophisticated Customers

Let us briefly elaborate on the alternative model discussed in Subsection 5.3.2, in which Customers are sophisticated and infer the level of doping from the situation of strategic interaction. Figure 5.3 shows this variation of our model. The difference to the payoffs of the initial example presented in Figure 5.1 is only the Customers' payoff at d-n-s, which turned from 3 to 1. Proposition 5.A3.1 shows that the Probability of doping in equilibrium is substantial, while the probability of being tested is much smaller.

Proposition 5.A3.1. *For the game depicted in Figure 5.3, in the unique SPNE Athletes dope with probability* $p^* = \frac{\sqrt{209}+23}{64} \approx 0.59$*, Organizers test with probability* $q^* = \frac{2\sqrt{209}-18}{23+\sqrt{209}} \approx 0.29$*, and Customers leave after a scandal with certainty and stay after no scandal with probability* $r^* = \frac{\sqrt{209}-3}{20} \approx 0.57$*. This is also a PBE with the equilibrium belief* $\alpha^* = \frac{1}{2}$ *that customers are not doped if there was no scandal.*

Proof. To describe the behavioral strategies, let p be the probability that the Athletes dope, let q be the probability that the Organizers test, and let r be the probability that Customers stay after no scandal, as depicted in Figure 5.3. After a scandal Customers leave with probability one in any SPNE. Note first that $p = q = 1$ cannot be part of an equilibrium since Athletes prefer not to dope if tested. Thus, there is a positive probability for no scandal in equilibrium. From the expected utility of the pure strategies it is directly derived that Athletes weakly prefer to dope (i.e. $p = 1$) if and only if

$$q \leq \frac{2}{5r+4}, \tag{5.7}$$

Organizers weakly prefer to test (i.e. $q = 1$) if and only if

$$p \geq \frac{1}{4-4r}, \tag{5.8}$$

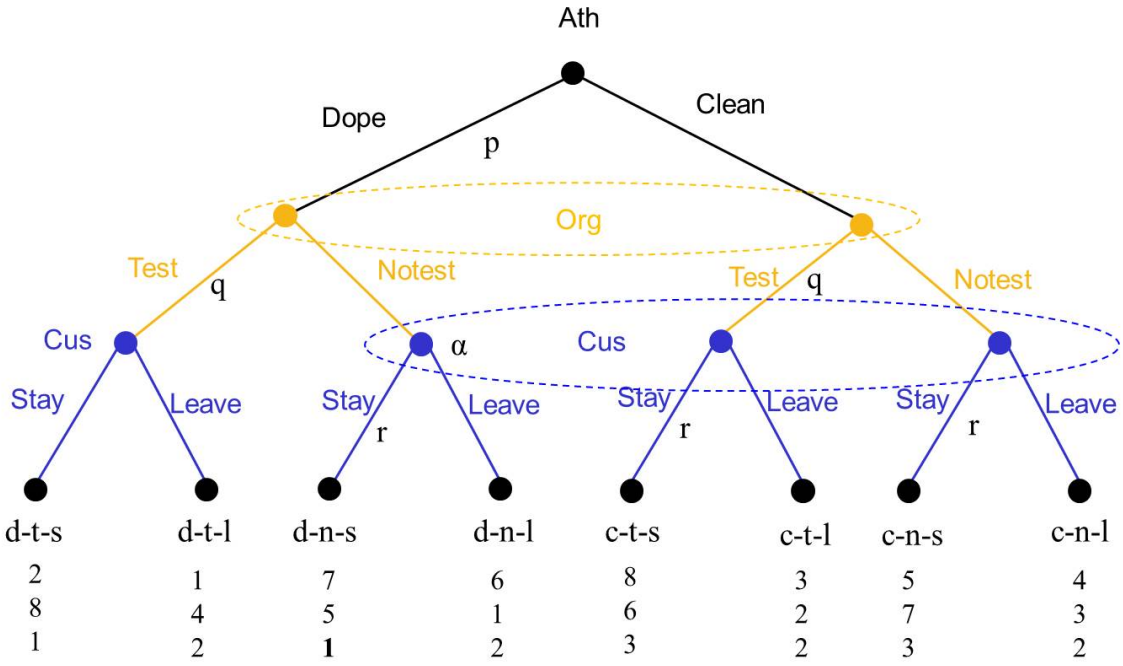


Figure 5.3: Structure of the game and an example for payoffs when Customers observe doping ex post.

and Customers weakly prefer to stay (i.e. $r = 1$) after no scandal if and only if

$$p \leq \frac{1}{2-q}. \tag{5.9}$$

Suppose, $p = 1$. This implies $r = 0$ by (5.9), which together with (5.8) implies $q = 1$. Now, (5.7) yields $p = 0$, a contradiction. Alternatively, suppose $p = 0$. This implies $r = 1$ by (5.9), which implies $q = 0$. Now, (5.7) yields $p = 1$, a contradiction. We conclude that in equilibrium, $p \in (0, 1)$. Hence, Athletes must be indifferent, i.e. $q = \frac{2}{5r+4}$ by (5.7). Thus, Organizers are also indifferent, i.e. $p = \frac{1}{4-4r}$ by (5.8). Since $p \leq 1$, $r \leq \frac{3}{4}$. Thus, Customers must either choose $r = 0$ or be indifferent.

If $r = 0$, then $q = \frac{1}{2}$ by (5.7) and $p = \frac{1}{4}$ by (5.8). This leads to a contradiction since (5.9) implies $r = 1$.

Thus, in equilibrium Customers are indifferent as well, i.e. $p = \frac{1}{2-q}$ by (5.9). Solving the system of equations in which (5.7), (5.8), and (5.9) hold with equality yields p^*, q^*, r^*, i.e. the first part of the proposition (we used the quadratic formula for the exact expressions). By plugging in p^* and q^* into $\alpha = \frac{p(1-q)}{p(1-q)+1-p}$, we get $\alpha^* = \frac{1}{2}$, which makes Customers indifferent between staying and leaving.

□

5.A3.2 More Critical Customers

Let us briefly elaborate on the model variation mentioned in Subsection 5.4.1, in which Customers are more critical than in our model. Thus, suppose that d-n-l $\succ^{Cus}$ d-n-s and c-n-l $\succ^{Cus}$ c-n-s, i.e. Customers preferred to withdraw their support if there are no doping tests. This is in contradiction to Assumptions A1. Let us hence change A1 to A1' such that these two orderings have changed, while all other binary comparisons are left unchanged. (This change of prefer-

ence reflects that Customers are here assumed to be more skeptical about doping practices and therefore insisted on the proof of clean sports in order to stay a supporter.) Under transparency, these alternative preferences rule out all doping equilibria as the following proposition shows.

Proposition 5.A3.2. *Under Assumptions A1' and A2* $\hat{s} = (Clean, Test, LSL)$ *is the unique SPNE in the game with well-informed Customers (cf. Figure 5.2).*

Proof. The beginning of the proof of this proposition is fully analogous to the first and second part of the proof of Proposition 4 because no assumption of A1' is used that does not coincide with the assumptions in A1. That is, we can restrict attention to two strategies of the Customers LSL and LSS, while for LSL, cf. Matrix (5.5), we find the mutual best response as $(Clean, Test)$ such that $(Clean, Test, LSL)$ is a SPNE. For the case of LSS, cf. Matrix (5.6), now the difference between A1' and A1 becomes relevant. Matrix (5.6), again, leads to best replies $(Dope, Notest)$. However, LSS is not a best reply to $(Dope, Notest)$ because d-n-l $\succ^{Cus}$ d-n-s by A1'. There are no other equilibria with LSS because A2 yields deviations from outcomes c-t-s and c-n-s, while d-t-l is not a candidate because c-t-s $\succ^{Ath}$ d-t-l by A1'. □

Bibliography

Abowd, J., Kramarz, F., and Margolis, D. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Adams, R., Almeida, H., and Ferreira, D. (2005). Powerful CEOs and their impact on corporate performance. *Review of Financial Studies*, 18(4):1403–1432.

Andreozzi, L. (2004). Rewarding policemen increases crime. Another surprising result from the inspection game. *Public Choice*, 121(1):69–82.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Antonovics, K. and Knight, B. G. (2009). A new look at racial profiling: Evidence from the boston police department. *The Review of Economics and Statistics*, 91(1):163–177.

Anwar, S. and Fang, H. (2005). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *NBER Working Paper*, (No. 11264).

Aoyagi, M. (2010). Information feedback in a dynamic tournament. *Games and Economic Behavior*, 70(2):242–260.

Artium, M. and Rimkus, N. (2001). Schiri–was pfeifst Du denn da...?! *Unpublished term paper. University of Dortmund*.

Aumann, R. J. and Maschler, M. (1972). Some thoughts on the minimax principle. *Management Science*, 18(5-Part-2):54–63.

Avenhaus, R., Von Stengel, B., and Zamir, S. (2002). Inspection games. *Handbook of game theory with economic applications*, 3:1947–1987.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7):435–452.

Bach, N., Gürtler, O., and Prinz, J. (2009). Incentive Effects in Tournaments with Heterogeneous Competitors – an Analysis of the Olympic Rowing Regatta in Sydney 2000. *Management Revue*, 20(3):239–253.

Backes-Gellner, U. and Pull, K. (2013). Tournament compensation systems, employee heterogeneity, and firm performance. *Human Resource Management*, 52(3):375–398.

Becker, G. (1957). *The Economics of Discrimination*. University of Chicago Press, Chicago.

Bennedsen, M., Pérez-González, F., and Wolfenzon, D. (2007). Do CEOs matter? *Unpublished Working Paper. Copenhagen Business School*.

Berentsen, A. (2002). The economics of doping. *European Journal of Political Economy*, 18(1):109–127.

Berentsen, A., Bruegger, E., and Loertscher, S. (2008). On cheating, doping and whistleblowing. *European Journal of Political Economy*, 24(2):415–436.

Berentsen, A. and Lengwiler, Y. (2004). Fraudulent accounting and other doping games. *Journal of Institutional and Theoretical Economics JITE*, 160(3):402–415.

Berger, J. and Nieken, P. (2014). Heterogeneous Contestants and the Intensity of Tournaments. An Empirical Investigation. *Journal of Sports Economics*, Published on June 17, 2014:Doi: 10.1177/1527002514538639.

Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics*, 118(4):1169–1208.

BFV (2001). Schiedsrichterbeobachter - Partner des Schiedsrichters -. *Bayrischer Fußball Verband*.

Bird, E. and Wagner, G. (1997). Sport as a common property resource. *Journal of Conflict Resolution*, 41(6):749–766.

Boeri, T. and Severgnini, B. (2008). The italian job: Match rigging, career concerns and media concentration in 'Serie A'. *IZA Discussion Paper*, (No. 3745).

Borland, J. (1992). Career concerns: Incentives and endogenous learning in labour markets. *Journal of Economic Surveys*, 6(3):251–270.

Boyko, R., Boyko, A., and Boyko, M. (2007). Referee bias contributes to home advantage in English Premiership football. *Journal of Sports Sciences*, 25(11):1185–1194.

Breivik, G. (1992). Doping games: A game theoretical exploration of doping. *International Review for the Sociology of Sport*, 27(3):235–253.

Brown, J. (2011). Quitters never win: The (adverse) incentive effects of competing with superstars. *Journal of Political Economy*, 119(5):982–1013.

Bull, C., Schotter, A., and Weigelt, K. (1987). Tournaments and piece rates: An experimental study. *The Journal of Political Economy*, 95(1):1–33.

Buraimo, B., Forrest, D., and Simmons, R. (2007). The Twelfth Man? Refereeing Bias in English and German Soccer. *IASE Working Paper*, (No. 07).

Casas-Arce, P. and Martínez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science*, 55(8):1306–1320.

Clarke, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *The Statistician*, 44(4):509–521.

Courneya, K. and Carron, A. (1992). The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology*, 14(1):13–27.

Dawson, P. (2012). Experience, social pressure and performance: The case of soccer officials. *Applied Economics Letters*, 19(9):883–886.

Deutscher, C., Frick, B., Gürtler, O., and Prinz, J. (2013). Sabotage in Tournaments with Heterogeneous Contestants: Empirical Evidence from the Soccer Pitch. *The Scandinavian Journal of Economics*, 115(4):1138–1157.

DFB (2007a). *DFB Schiedsrichter-Zeitung*, (5).

DFB (2007b). *DFB Schiedsrichter-Zeitung*, (6).

DFB (2009). *DFB Schiedsrichter Informationen*.

Dharmapala, D. and Ross, S. L. (2004). Racial bias in motor vehicle searches: Additional theory and evidence. *Contributions in Economic Analysis & Policy*, 3(1):Article 12.

Dilger, A. and Tolsdorf, F. (2004). *Doping als Wettkampfphänomen*. Events im Sport: Marketing, Management, Finanzierung, Köln.

Dohmen, T. (2003). In Support of the Supporters? Do social forces shape decisions of the impartial? *IZA Discussion Paper*, (No. 755).

Dohmen, T. (2008). The Influence of Social Forces: Evidence from the Behavior of Football Referees. *Economic Inquiry*, 46(3):411–424.

Dresher, M. (1962). A Sampling Inspection Problem in Arms Control Agreements: A Game – Theoretic Analysis. *Memorandum No. RM–2972–ARPA, RAND Corporation, Santa Monica, California*.

Eber, N. (2002). Credibility and independence of the World Anti-Doping Agency. *Journal of Sports Economics*, 3(1):90–96.

Eber, N. (2006). Doping. In Andreff, W. and Szymanski, S., editors, *Handbook on the Economics of Sport*, pages 773–783. Edward Elgar.

Eber, N. (2008). The performance-enhancing drug game reconsidered - A fair play approach. *Journal of Sports Economics*, 9(3):318–327.

Eber, N. and Thépot, J. (1999). Doping in sport and competition design. In *Recherches Économiques de Louvain/Louvain Economic Review*, pages 435–446.

Ebersberger, H., Malka, J., and Pohler, R. (1989). *Schiedsrichter im Fußball*. Limpert.

Ebersberger, H. and Pohler, R. (1997). Anmerkungen zum neuen Beobachtungsbogen. *DFB Schiedsrichter-Zeitung*, 3.

Ederer, F. (2010). Feedback and motivation in dynamic tournaments. *Journal of Economics & Management Strategy*, 19(3):733–769.

Emrich, E. and Pierdzioch, C. (2013). A note on the international coordination of antidoping policies. *Journal of Sports Economics*, Published on February 28, 2013:Doi: 10.1177/1527002513479802.

Emrich, E., Pierdzioch, C., and Pitsch, W. (2014). Die Marke Olympia und ihre Vertrauenseigenschaften - Eine Geschichte von Markt, Macht und Moral. *Unpublished Working Paper*.

Fee, C., Hadlock, C., and Pierce, J. (2010). Managers who lack style: Evidence from exogenous CEO changes. *Michigan State University, Working Paper*.

Franck, E. and Nüesch, S. (2010). The effect of talent disparity on team productivity in soccer. *Journal of Economic Psychology*, 31(2):218–229.

Franck, E. and Nüesch, S. (2011). The effect of wage dispersion on team outcome and the way team outcome is produced. *Applied Economics*, 43(23):3037–3049.

Frank, M. and Goyal, V. (2007). Corporate leverage: How much do managers really matter. *University of Minnesota and Hong Kong University of Science and Technology, Working Paper*.

Frick, B. (2011). Performance, salaries and contract length: Empirical evidence from German soccer. *International Journal of Sport Finance*, 6(2):87–118.

Frick, B., Gurtler, O., and Prinz, J. (2008). Anreize in Turnieren mit heterogenen Teilnehmern–Eine empirische Untersuchung mit Daten aus der Fussball–Bundesliga. *ZFBF: Schmalenbachs Zeitschrift für Betriebswirtschaftliche Forschung*, 60:385–405.

Frick, B. and Prinz, J. (2007). Pay and Performance in Professional Road Running: The Case of City Marathons. *International Journal of Sport Finance*, 2(1):25–35.

Friehe, T. (2008). Correlated payoffs in the inspection game: Some theory and an application to corruption. *Public Choice*, 137(1):127–143.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, 87(2):208–216.

Genakos, C. and Pagliero, M. (2012). Interim Rank, Risk Taking, and Performance in Dynamic Tournaments. *Journal of Political Economy*, 120(4):782–813.

Gershkov, A. and Perry, M. (2009). Tournaments with midterm reviews. *Games and Economic Behavior*, 66(1):162–190.

Gibbons, R. and Murphy, K. (1992). Optimal incentive contracts in the presence of career concerns: Theory and evidence. *Journal of Political Economy*, 100(3):468–505.

Gibbons, R. and Waldman, M. (1999). Careers in organizations: Theory and evidence. *Handbook of labor economics*, 3:2373–2437.

Graham, J., Li, S., and Qiu, J. (2012). Managerial attributes and executive compensation. *Review of Financial Studies*, 25(1):144–186.

Gürtler, O. and Harbring, C. (2010). Feedback in tournaments under commitment problems: experimental evidence. *Journal of Economics & Management Strategy*, 19(3):771–810.

Haugen, K. (2004). The performance-enhancing drug game. *Journal of Sports Economics*, 5(1):67–86.

Hentschel, S., Mühlheußer, G., and Sliwka, D. (2012). The Impact of Managerial Change on Performance. The Role of Team Heterogeneity. *IZA Discussion Paper*, (No. 6884).

Holder, R. L. and Nevill, A. M. (1997). Modelling performance at international tennis and golf tournaments: is there a home advantage? *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(4):551–559.

Holler, M. J. (1990). The unprofitability of mixed-strategy equilibria in two-person games: A second folk-theorem. *Economics Letters*, 32(4):319–323.

Holler, M. J. (1993). Fighting pollution when decisions are strategic. *Public Choice*, 76(4):347–356.

Holmström, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

Irlenbusch, B. and Sliwka, D. (2006). Career concerns in a simple experimental labour market. *European Economic Review*, 50(1):147–170.

Kahn, L. M. (2000). The sports business as a labor market laboratory. *The Journal of Economic Perspectives*, 14(3):75–94.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–291.

Kirstein, R. (2012). Doping, the Inspection Game, and Bayesian Enforcement. *Journal of Sports Economics*, Doi: 10.1177/1527002512461358.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Koch, A., Morgenstern, A., and Raab, P. (2009). Career concerns incentives: An experimental test. *Journal of Economic Behavior and Organization*, 72(1):571–588.

Kräkel, M. (2007). Doping and cheating in contest-like situations. *European Journal of Political Economy*, 23(4):988–1006.

Lallemand, T., Plasman, R., and Rycx, F. (2008). Women and competition in elimination tournaments evidence from professional tennis data. *Journal of sports economics*, 9(1):3–19.

Lazear, E. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy*, 89(5):841–864.

Levy, G. (2005). Careerist judges and the appeals process. *RAND Journal of Economics*, 36(2):275–297.

Lucey, B. and Power, D. (2004). Do soccer referees display home team favouritism. *Trinity College Dublin, Working Paper*.

Ludwig, S. and Lünser, G. K. (2012). Observing your competitor – The role of effort information in two-stage tournaments. *Journal of Economic Psychology*, 33(1):166–182.

Lynch, J. G. (2005). The effort effects of prizes in the second half of tournaments. *Journal of Economic Behavior & Organization*, 57(1):115–129.

Maennig, W. (2002). On the economics of doping and corruption in international sports. *Journal of Sports Economics*, 3(1):61–89.

Malmendier, U. and Tate, G. (2005). CEO overconfidence and corporate investment. *The Journal of Finance*, 60(6):2661–2700.

Malmendier, U. and Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *Journal of Financial Economics*, 89(1):20–43.

Maschler, M. (1967). The inspector's non-constant-sum game: Its dependence on a system of detectors. *Naval Research Logistics Quarterly*, 14(3):275–290.

McLaughlin, K. J. (1988). Aspects of tournament models: A survey. *Research in labor economics*, 9(1):225–256.

Mechtel, M., Bäker, A., Brändle, T., and Vetter, K. (2011). Red cards not such bad news for penalized guest teams. *Journal of Sports Economics*, 12(6):621–646.

Messing, M. and Müller, N. (1996). *Veranstaltungsbesuch und sportpolitische Polarisation deutscher Olympia-Touristen in Barcelona 1992*. Auf der Suche nach der Olympischen Idee. Kassel.

Messing, M., Müller, N., and Schorman, K. (2004). Local Visitors and Tourists at the Modern Pentathlon in Sydney 2000 – a Contribution on the Internal Differentiation of an Olympic Spectator. In Messing, M.; Müller, N. P. H., editor, *Olympischer Dreiklang. Werte–Geschichte–Zeitgeist*, pages 365–406. Kassel: Agon Sportverlag.

Messing, M., Müller, N., and Schormann, K. (2008). *Zuschauer beim antiken Agon und bei den Olympischen Spielen in Athen 2004 - anthropologische Grundmuster und geschichtliche Figurationen*. Antike Lebenswelten, Konstanz, Wandel, Wirkungsmacht. Festschrift für Ingomar Weiler zum 70. Geburtstag. Wiesbaden.

Miceli, T. and Coşgel, M. (1994). Reputation and judicial decision-making. *Journal of Economic Behavior & Organization*, 23(1):31–51.

Neale, W. C. (1964). The peculiar economics of professional sports: A contribution to the theory of the firm in sporting competition and in market competition. *The Quarterly Journal of Economics*, 78(1):1–14.

Nevill, A., Balmer, N., and Williams, A. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport & Exercise*, 3(4):261–272.

Nevill, A. M. and Holder, R. L. (1999). Home advantage in sport. *Sports Medicine*, 28(4):221–236.

Nevill, A. M., Newell, S. M., and Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sciences*, 14(2):181–186.

Nuesch, S. (2009). Are demographic diversity effects spurious? *Economic analysis and policy*, 39(3):379–388.

Page, K. and Page, L. (2010). Alone against the crowd: Individual differences in referees' ability to cope under pressure. *Journal of Economic Psychology*, 31(2):192–199.

Pitsch, W. and Emrich, E. (2011). The frequency of doping in elite sport: Results of a replication study. *International Review for the Sociology of Sport*, 47(5):559–580.

Pitsch, W., Emrich, E., and Klein, M. (2007). Doping in elite sports in Germany: Results of a www survey. *European Journal for Sport and Society*, 4(2):89–102.

Pitsch, W., Frenger, M., and Emrich, E. (2010). The impact of anti-doping legislation in Europe - outlines for the development of model-based hypotheses. In Emrich, E. and Pitsch, W., editors, *Sport and Doping. The Analysis of an Antagonistic Symbiosis*. Peter Lang: Frankfurt/M.

Pope, D. G. and Schweitzer, M. E. (2011). Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. *The American Economic Review*, 101(1):129–157.

Preston, I. and Szymanski, S. (2003). Cheating in contests. *Oxford Review of Economic Policy*, 19(4):612–624.

Reilly, B. and Witt, R. (2013). Red cards, referee home bias and social pressure: Evidence from English Premiership Soccer. *Applied Economics Letters*, 20(7):710–714.

Rickman, N. and Witt, R. (2008). Favoritism and financial incentives: A natural experiment. *Economica*, 75(298):296–309.

Rottenberg, S. (1956). The baseball players' labor market. *The Journal of Political Economy*, 64(3):242–258.

Savulescu, J., Foddy, B., and Clayton, M. (2004). Why we should allow performance enhancing drugs in sport. *British Journal of Sports Medicine*, 38(6):666–670.

Schotter, A. and Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *The Quarterly Journal of Economics*, 107(2):511–539.

Scoppa, V. (2008). Are subjective evaluations biased by social factors or connections? An econometric analysis of soccer referee decisions. *Empirical Economics*, 35(1):123–140.

Solberg, H. A., Hanstad, D. V., and Thoring, T. A. (2010). Doping in elite sport - do the fans care? Public opinion on the consequences of doping scandals. *International Journal of Sports Marketing & Sponsorship*, 11(3):185–199.

Striegel, H., Ulrich, R., and Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and alcohol dependence*, 106(2):230–232.

Strigel, E. (1999). Der lange Weg zum FIFA-Schiedsrichter. In *Dikty*, pages 20–23.

Sunde, U. (2009). Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data. *Applied Economics*, 41(25):3199–3208.

Sutter, M. and Kocher, M. (2004). Favoritism of agents - The case of referees' home bias. *Journal of Economic Psychology*, 25(4):461–469.

Teipel, D., Kemper, R., and Heinemann, D. (1999). *Beanspruchung von Schiedsrichtern und Schiedsrichterinnen im Sport*. Köln: Strauß.

Tröger, C. (2006). Olympia - Im Spannungsfeld von Mythos und Marke. *Dissertation Saarland University*.

Tsebelis, G. (1989). The abuse of probability in political analysis: The Robinson Crusoe fallacy. *The American Political Science Review*, 83(1):77–91.

Van Reeth, D. (2013). Television demand for the Tour de France: the importance of outcome uncertainty, patriotism and doping. *International Journal of Sport Finance*, 8:39–60.

Wicker, P., Prinz, J., Weimar, D., Deutscher, C., and Upmann, T. (2013). No Pain, No Gain? Effort and Productivity in Professional Soccer. *International Journal of Sport Finance*, 8(2):124–139.

Chapter 6
Appendix

Summary of results

I. *The impact of referees on match outcome in professional sports: Evidence from the German Football Bundesliga*

Within this paper, we attempt a new empirical strategy to estimate whether referees in professional football are biased although it is expected that referees behave impartial. Therefore we use an OLS approach to estimate whether referees have significant impact on match outcome. Running different regressions lead us to empirical evidence that referees have significant effects on *match outcome* (result, goal difference) and further *referee decisions* (e.g. yellow cards, awarded goals and penalties etc.). Additionally we find that these individual referee effects differ between home and away teams. Lastly, we figure out whether observable referee characteristics such as age, experience or profession can explain the individual referee effects. Here we find only limited support that referees follow career concerns. Further we cannot confirm our second assumption on referee behavior: there is no significant evidence that referees worry about their reputation. On the contrary, if we look at referee's profession we find further explanations for significant individual effects. Yet all these results hold true for chosen referee effects for home and away teams. In the end, we have to recognize that these observable characteristics cannot explain the significant referee effects. Thus we have to assume that other unobservable qualities affect the referee's individual effect on match outcome.

Diese Arbeit verwendet einen neuen empirischen Ansatz, um zu überprüfen, ob die Entscheidungen von Schiedsrichtern im professionellen Fußball beeinflusst sind, obwohl erwartet wird, dass sie sich unparteiisch verhalten. Hierfür wird ein OLS-Ansatz verwendet, der empirisch testen soll, ob Schiedsrichter einen signifikanten Einfluss auf den Spielausgang haben. Mittels verschiedener Regressionen können wir empirische Evidenz dafür finden, dass Schiedsrichter sowohl einen signifikanten Einfluss auf den *Spielausgang* (Ergebnis, Tordifferenz) als auch auf *Schiedsrichterentscheidungen* (z.B. Gelbe Karte, Tore, Elfmeter) haben. Darüber hinaus zeigen die Ergebnisse, dass sich dieser individuelle Einfluss von Schiedsrichtern zwischen Heim- und Gastmannschaften unterscheidet. Anschließend haben wir mittels beobachtbarer Eigenschaften wie z.B. Alter, Erfahrung oder Beruf versucht die gefundenen individuellen Effekte der Schiedsrichter zu erklären. Jedoch können wir die Annahme, dass die Karriere eines Schiedsrichters seine individuellen Effekte beeinflusst nur in begrenztem Umfang bestätigen. Des Weiteren finden sich keine signifikanten Ergebnisse für die Annahme, dass Schiedsrichter sich um ihre Reputation sorgen. Im Gegensatz dazu können die Berufe der Schiedsrichter weitere Erklärungen für die individuellen Effekte liefern. Gleichwohl lassen sich diese Resultate nur für einige

wenige Schiedsrichtereffekte für Heim- und Gastmannschaften bestätigen. Daher lässt sich vermuten, dass eher die unbeobachbaren Charakteristika der Schiedsrichter diese unterschiedlichen individuellen Effekte hervorrufen.

II. *Are football referees really neutral or do they have prejudices?*

This paper has examined whether referees have prejudices regarding home and away teams in German 1^{st} Bundesliga matches. Basis for this analyzes is a game-theoretic model that is usually applied in a racial-discrimination setting. The equilibrium of this model predicts that if referees are neutral, the fraction of wrong decisions over total referee decisions ("fail rate") will not differ significantly across home and away teams. We also test this hypothesis in additional subgroups such as *crucial*, *local derby* or *racetrack*. First we compare the average ratios of wrong referee decisions and find a lot of substantial differences across all our subgroups. Next we use a Pearson χ^2 test to verify if these differences are statistically significant and therefore an evidence of prejudiced referees. In summary, we find significant results for referee decisions on goals, penalties and total number of referee decisions for all our subgroups. Particularly for our most interesting subgroup concerning home and away teams, we find various significant differences between the fail rates. This permits us to conclude that referees judge home and away teams differently.

Diese Studie untersucht, ob Schiedsrichter der 1. Fußball Bundesliga Vorurteile gegenüber Heim- und Gastmannschaften haben. Die Grundlage für diese Analyse bildet ein spieltheoretisches Modell, das ursprünglich in wissenschaftlichen Arbeiten über rassistische Diskriminierung Verwendung findet. Das Gleichgewicht dieses Models besagt, dass der Anteil der falschen Schiedsrichterentscheidungen an den Gesamtentscheidungen ("Fehlerquote") sich nicht signifikant zwischen Heim- und Gastmannschaften unterscheiden wird, wenn der Schiedsrichter neutral ist. Diese Hypothese testen wir noch in weiteren Untergruppen. Hierzu zählen z.B. *wichtige und unwichtige Spiele*, *"Lokalderbies"* und Spiele in *Stadien mit und ohne Laufbahn*. In einem ersten Schritt vergleichen wir die durchschnittlichen Fehlentscheidungen innerhalb der Untergruppen. Hier zeigen sich maßgebliche Unterschiede innerhalb der Untergruppen. Anschließend testen wir mittels einem Pearson χ^2 Test, ob diese Unterschiede statistisch signifikant sind und somit einen Beweis für Vorurteile durch Schiedsrichter liefern. Abschließend können wir festhalten, dass wir signifikante Ergebnisse für die Schiedsrichterentscheidungen Tore, Elfmeter sowie für die Summe aller Schiedsrichterentscheidungen finden können. Insbesondere für die Untergruppe Heim- und Gastmannschaften können wir mehrere signifikante Unterschiede zwischen den jeweiligen Fehlerquoten finden. Daraus lässt sich schließen, dass Heim- und Gastmannschaften von den Schiedsrichtern unterschiedlich bewertet werden.

III. *The Impact of Intermediate Information on Effort Provision in Soccer*

This study contributes work on analyzing individual drivers of effort and examines how intermediate information affects individual effort. Therefore we analyze substitute football players from 1^{st} German Bundesliga and measure effort with number of *runs*, *sprints* and running *distance*. As an indicator for intermediate information we use *goal difference* at

the time of substitution. Using OLS regression technique, we find that football players show higher effort when the respective team is leading by one goal. Further, effort is higher in matches where a teams leads by one goal compared to matches with a tied score. This result is in line with prospect theory stating that individuals value potential losses higher than gains. Lastly, effort declines if the score indicates the game to be decided. These results also have implications for any other contests. This implies that it would be irrational to provide information to contestants if a competition is already decided because both competitors would decrease effort. On the other hand if a contest is very closed the leading competitor should be informed to increase effort, but the trailing competitor should not receive any information concerning the intermediate score.

Die vorliegende Studie liefert einen Beitrag für die Frage inwieweit das Anstrengungsniveau von Individuen beeinflusst werden kann und im Besonderen wie Informationen über den aktuellen Zwischenstand in einem Wettbewerb die individuellen Bemühungen beeinflussen können. Untersuchungsgegenstand sind hier Einwechselspieler in Spielen der 1. Fußball Bundesliga. Indikatoren für das Anstrengungsniveau sind die Anzahl der *Läufe* und *Sprints* sowie die zurückgelegte *Distanz* in einem Fußballspiel. Als Indikator für intermediäre Informationen nutzen wir die *Tordifferenz* zum Zeitpunkt der Einwechslung. Mittels OLS-Regressionen zeigt sich, dass Fußballspieler sich mehr anstrengen, wenn die eigene Mannschaft mit einem Tor führt. Zusätzlich zeigt sich in Spielen in denen eine Mannschaft mit einem Tor führt, ein höheres Anstrengungsniveau verglichen mit dem Anstrengungsniveau in ausgeglichenen Spielen. Dieses Resultat stimmt mit den Aussagen der "Prospect Theorie" überein. Diese besagt, dass Individuen potentiellen Verlusten ein höheres Gewicht beimessen als potentiellen Gewinnen. Zu guter Letzt kann noch festgehalten werden, dass das Anstrengungsniveau sehr stark sinkt sobald ein Spiel entschieden ist. Die hier gefundenen Resultate haben auch Implikationen für Wettbewerbe jedweder Art. Dies bedeutet zum einen, dass es unvernünftig wäre in einem längst entschiedenen Wettbewerb die Teilnehmer über diesen Zwischenstand zu informieren. Dies hätte zur Folge, dass beide Wettbewerber ihr Anstrengungsniveau senken würde. Andererseits sollte in einem sehr ausgeglichenen Wettbewerb der Führende über den aktuellen Zwischenstand informiert werden, um seine Bemühungen zu steigern. Dies gilt jedoch nicht für den zurückliegenden Wettbewerber.

IV. *Nobody's Innocent – The Role of Customers in the Doping Dilemma*

In this analyzes we have extended the inspection game by a third player, namely customers who can withdraw their support. We adapt this inspection game to doping in professional sports and find out that customers who react critical and withdraw their support after a doping scandal, emphasizes fraudulent activities from athletes. This is because a customer's withdraw leads to lower incentives for organizers of sporting contests to test athletes on performance-enhancing drugs and convict them on doping. As a result, athletes have stronger incentives to use doping substances. Nevertheless, our model show that a change in the information structure would induce to a doping-free equilibrium. Therefore it is essential to establish transparency which means that customers have to be informed

whether there were negative tests or there were no serious tests. This transparency enables customers and other stakeholders to decide about their support on basis of serious anti-doping tests. Finally, our model is not restricted to doping and professional sports. There exist many industries (e.g. textiles or food) where customers have no full information on the production process of the goods they consume or rather it is not known whether the producing companies fulfill all requirements.

Diese Arbeit erweitert das "Inspection Game" um einen dritten Spieler, nämlich den Zuschauer (bzw. Kunden) der seine Unterstützung für ein Produkt bzw. Sportart o.ä. zurückziehen kann. Als praktische Anwendung für dieses Spiel wurde hier Doping im Profisport gewählt. Ein Ergebnis des Models ist, dass kritisches Verhalten der Zuschauer als Reaktion auf einen Doping-Skandal dazu führt, dass betrügerisches Verhalten durch die Athleten weiter verstärkt wird. Dies ergibt sich daraus, dass solch ein Rückzug der Unterstützung zu geringeren Anreizen für die Organisatoren solcher Wettbewerbe führt, die Athleten auf unerlaubte leistungssteigernde Mittel zu testen und damit Doping aufzudecken. Als eine Folge daraus haben die Athleten wieder einen stärkeren Anreiz Dopingmittel einzunehmen. Jedoch kann mittels einer Änderung der Informationsstruktur in diesem Spiel ein dopingfreies Gleichgewicht hergeleitet werden. Zu diesem Zweck ist es jedoch notwendig für Transparenz zu sorgen. Das bedeutet, dass Zuschauer und andere Interessenvertreter umfassend über die einzelnen Dopingtests informiert werden müssen, damit sie auf Basis dieser Testergebnisse darüber entscheiden können, ob sie den Sport weiterhin unterstützen möchten. Letztlich ist dieses Modell nicht nur auf Doping und Profisport anwendbar. In vielen weiteren Industrien (z.B. Textil oder Nahrungsmittel) haben die Kunden keine ausreichenden Informationen über den Produktionsprozess der Güter, die sie konsumieren. Mit anderen Worten, ihnen ist nicht bekannt, ob die Produzenten dieser Güter alle Anforderungen an ein "sauberes" Produkt erfüllen.

Published work

Büchel, B., Emrich, E. and Pohlkamp, S. (2014). Nobody's Innocent: The Role of Customers in the Doping Dilemma. *Journal of Sports Economics* (published online 7 October), Doi: 10.1177/1527002514551475

Eidesstattliche Versicherung

Ich, Stefanie Pohlkamp, versichere an Eides statt, dass ich die Dissertation mit dem Titel

Essays in Sports Economics

selbst und bei einer Zusammenarbeit mit anderen Wissenschaftlerinnen oder Wissenschaftlern gemäß den beigefügten Darlegungen nach §6 Abs. 3 der Promotionsordnung der Fakultät Wirtschafts- und Sozialwissenschaften vom 24. August 2010 verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht benutzt.

Hamburg, den 29.10.2014