UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Fakultät für Wirtschaftswissenschaften
Lehrstuhl für Organisations-, Sport- und Medienökonomie

THE ECONOMICS OF INDIVIDUAL BEHAVIOR IN COMPETITIVE ENVIRONMENTS: EMPIRICAL EVIDENCE FROM REAL-LIFE TOURNAMENTS

Der Fakultät für Wirtschaftswissenschaften der Universität Paderborn

zur Erlangung des akademischen Grades Doktor der Wirtschaftswissenschaften

- Doctor rerum politicarum -

vorgelegte Dissertation

von

Friedrich Scheel

geboren am 7. Juni 1984 in Erfurt

2014

VORWORT

An dieser Stelle möchte ich die Gelegenheit nutzen, mich bei all denjenigen zu bedanken, die mich während meines Promotionsprojekts begleitet und mit ihrer Unterstützung und ihrem Zuspruch sowie zahlreichen hilfreichen Diskussionen maßgeblich zu diesem für mich so wichtigen Erfolg beigetragen haben.

Allen voran möchte ich meinem Doktorvater Prof. Dr. Bernd Frick für die optimale Betreuung, sein Vertrauen sowie für seine „ansteckende Begeisterung“ für die sport- und personalökonomische Forschung danken. Er hat mich stets ermutigt, die z. T. noch im Reifeprozess befindlichen Forschungsaufsätze auf internationalen Konferenzen vorzustellen, um dort wertvolle Anregungen, Lob und Kritik von einem breiten Fachpublikum zu erhalten. Mein besonderer Dank gilt Prof. Dr. Martin Schneider, Prof. Dr. Stefan Betz und Prof. Dr. Claus-Jochen Haake für ihre Bereitschaft, als Zweitgutachter bzw. Promotionskommissionsmitglieder mitzuwirken.

Einen mindestens genauso großen Anteil an dieser Arbeit haben meine (ehemaligen) Kolleginnen und Kollegen, Dr. Andre Kolle, Dr. Linda Kurze, Dr. Marcel Battré, Prof. Dr. Christian Deutscher, Anica Rose, Tobias Neuhann, Sina Kolaschewski, Filiz Gülal und PD Dr. Benjamin Balsmeier, mit denen die Zusammenarbeit weit über die berufliche Ebene hinaus ging und aus der viele sehr gute Freundschaften hervorgegangen sind. Ich kann mich wirklich glücklich schätzen, Teil eines so gut harmonierenden Teams gewesen zu sein und freue mich umso mehr, den Großteil der „alten Besatzung“ bei meinem neuen Arbeitgeber wieder anzutreffen. Ein herzlicher Dank gebührt auch den studentischen Hilfskräften, die mich insbesondere bei der Erstellung der Datensätze tatkräftig unterstützt haben.

Abschließend möchte ich mich bei meinen Eltern, meiner Familie, meinen Freunden und meiner Freundin Tine für die liebevolle Unterstützung und die nötige Ablenkung bedanken. Sie haben mich stets motiviert und waren für mich da, wenn ich sie brauchte.

Friedrich Scheel Paderborn, im Oktober 2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAP	Anonymous Application Procedures
ACB	Analysis of Competitive Balance
ANOVA	Analysis of Variance
ATP	Association of Tennis Professionals
CHF	Swiss Francs
CV	Coefficient of Variation
DFB	Deutscher Fußball-Bund (German Football Association)
FIFA	Fédération Internationale de Football Association (International Federation of Association Football)
FIS	Fédération Internationale de Ski (International Ski Federation)
GRE	Graduate Record Examinations
HHI	Herfindahl-Hirschmann-Index
IAAF	International Association of Athletics Federation
LAPG	League's Average per Game Attendance
LPGA	Ladies Professional Golf Association
LPM	Linear Probability Model
MIT	Massachusetts Institute of Technology
MLB	Major League Baseball
MLS	Major League Soccer
NBA	National Basketball Association
NFL	National Football League
NHL	National Hockey League
OECD	Organisation for Economic Co-operation and Development
OLS	Ordinary Least Squares
PPS	Percentage Points
SD	Standard Deviation
GSOEP	German Socio-Economic Panel
TOV	Thrill of Victory
UEFA	Union of European Football Associations
UOH	Uncertainty of Outcome Hypothesis
WPCT	Winning Percentage

1 INTRODUCTION

"*Incentives are the essence of economics*" (Lazear 1987: 744) and can, to a large extent, explain individual behavior. Among others, (intrinsic as well as extrinsic) incentives induce individuals to invest in education, competencies and social skills, as specified by neoclassical human capital theory (see e.g. Mincer 1958, 1970, 1974; Schultz 1961; Becker 1962, 2009). Peak performances and major achievements among artists, actors, musicians, athletes (and other individuals in the focus of public attention) are by and large evoked by incentives. Thereby, incentives do not necessarily have to be of pecuniary nature. As a hypothetical example, despite the expected profits, medical scientists might have social and/or moral incentives to withhold the development of a new anti-cancer medication whose anticipated side-effects are likely to outweigh the benefits. On the other hand, incentives can also have adverse effects. If, say, the expected payoff from criminal activity is substantially larger than the expected income from legal work, utility-maximizing individuals have an incentive to engage in illegal activities, as argued by Becker (1968). From a purely rent-seeking perspective, it might – under certain conditions – be economically rational for individuals to join a terrorist organization (see e.g. Krueger and Maleckova 2003; Berrebi 2007), to deliberately lose fights in Japanese sumo tournaments (see e.g. Duggan and Levitt 2002; Dietl et al. 2010) or to use illicit performance-enhancing drugs in a sporting contest (see e.g. Bird and Wagner 1997; Berentsen 2002; Dilger et al. 2007; Kräkel 2007), to name but a few examples.

In the context of institutions, incentives serve as a means to align the interests of both parties in principal-agent relationships. In particular, firm owners face the challenge of motivating workers to forgo leisure time and instead increase individual effort, or more accurately, labor. Thus, the provision of incentives is a key element in order to overcome – or at least reduce – agency problems in firms arising from information asymmetries. Building on Jensen and Meckling's (1976) seminal work, the hypotheses derived from agency theory have been (and continue to be) extensively researched by, inter alia, Hölmstrom (1979, 1982), Fama (1980), Milgrom and Roberts (1988) and Hölmstrom and Milgrom (1991, 1994). For a thorough review of the earlier literature one can refer to Prendergast (1999).

Although commonly intended to motivate agents to increase effort, monetary incentives can also trigger the opposite effect. Depending on the institutional setting, (inappropriately designed) *extrinsic* incentives might crowd out *intrinsic* motivations and thus lead to nega-

tive externalities on behalf of the principal (see e.g. Gneezy and Rustichini 2000; Gneezy et al. 2011). This so-called "crowding-out effect" is predominantly observed in environments which do typically not involve pecuniary incentives, such as the voluntary sector (e.g. Frey and Goette 1999; Osterloh and Frey 2002), nonprofit, charitable organizations (e.g. Andreoni and Payne 2011) or other altruistically motivated actions such as blood donations (e.g. Mellström and Johannesson 2008).

Incentive contracts in employer-employee relationships can take various forms. Standard economic theory posits that workers should be paid a piece rate according to their marginal productivity.[1] Contrary to this, Lazear and Rosen (1981) were among the first to propose compensation schemes which pay according to an individual's rank in an organization rather than the absolute output level.[2] Presupposing that workers are relatively homogeneous with respect to their abilities, such rank-order tournaments can be quite efficient in that they provide strong incentives for all individuals in a firm. However, tournaments can have adverse effects, too. If the difference between the winner's prize (e.g. the promotion to a higher level in the firm's hierarchy) and the second best alternative is too large, individuals might engage in sabotage activities and, thus, no longer act in the principal's interest, as emphasized by Dilger et al. (2007). Hence, it is the role of the principal to design incentive and reward structures that are optimal in (i) inducing high effort levels among all contestants and, at the same time, (ii) minimizing the likelihood of illegal activities. This, in turn, causes direct and indirect (or "hidden") costs of control (Falk and Kosfeld 2006) that must be compensated by productivity gains.

Apart from corporate tournaments, in which employees might compete for a pay raise, a promotion or the "employee-of-the-year-award", rank-order tournaments can be found in all kinds of institutional environments: Students could vie for a limited number of scholarships; turbine manufacturers could compete with one another for a major contract with one of the leading aircraft manufacturers; (more or less talented) singers in a television casting show might battle for a recording contract, while cities or even entire nations might compete against each other in a bid to organize and host major sporting events such as the Olympic Games or the FIFA Football World Cup. In general, tournament theory offers a whole set of testable hypotheses and helps explaining (individual) behavior in competitive

[1] See Gibbons (1987) for a critical discussion of piece-rate compensation schemes.
[2] See Nalebuff and Stiglitz (1983) as well as Rosen (1986) for notable extensions to Lazear and Rosen's (1981) initial theoretical framework. For a comprehensive review of the tournament theoretical literature over the past 30 years one can refer to Connelly et al. (2014).

environments. Given these merits, it is not surprising that empirical research using data from real-life tournaments is abundant, but yet far from exhaustive.

A relatively new field of research that has recently gained momentum particularly among personnel economists and that frequently addresses (corporate) tournaments is "insider econometrics". Combining rich and often highly sensitive firm-specific data with knowledge from industry experts, the insider econometric approach is a powerful tool when analyzing the impact of HRM practices on productivity (e.g. Bloom and Van Reenen 2011). This might include the analysis of incentive effects of different pay regimes, institutional changes or organizational differences (e.g. team work versus autonomous work) on employees in companies. Pioneering works in this field include, inter alia, Ichniowski et al. (1997) on productivity effects of innovative HRM practices in steel companies[3] and Lazear (2000) analyzing the (positive) incentive effects of a switch from hourly wages to piece rates on the productivity of workers in a large US-based auto glass company. In a similar vein (and with similar results), Shearer (2004) examines the impact of hourly pay and piece rates on the performance of tree planters in British Columbia, while Bandiera et al. (2005, 2007, 2009) look at the productivity of workers on a UK fruit picking farm under different pay systems. Exploring peer effects among cashiers in a US supermarket chain, Mas and Moretti (2009) find strong support for positive productivity spillovers. Their results suggest that the mere presence of highly productive personnel in a particular shift reduces free-riding and thus increases the productivity of coworkers. This, however, only applies to workers who can see their productive peer, but not to workers who do not see him. While the majority of these studies addresses HRM issues, the insider econometric approach can be applied to other areas such as mergers and acquisitions or social networks, too (see Frick and Fabel 2013 for a brief literature review). For a comprehensive overview of insider econometric studies published over the past one and a half decades one can refer to Ichniowski and Shaw (2009).

Although the insider econometric approach certainly offers many advantages, it also entails a few limitations: First, the access to private and highly sensitive corporate data is of course strictly limited (if not completely denied to "outsiders"). Second, the results of these studies often reflect idiosyncratic (i.e. firm- or industry-specific) characteristics that cannot

[3] The authors find that production lines using a cluster of innovative work practices (e.g. pay for performance, team work, flexible job assignments and on-the-job training) are significantly more productive than lines using rather traditional work practices such as hourly pay, narrow job definitions and close supervision.

be generalized to other firms or industries per se. Third, counterfactual evidence about what would have happened if certain measures had not been adopted in a firm is typically non-existent. One way of circumventing *some of* these limitations is to turn to the (professional) sports industry. Many sports contests are designed as rank-order tournaments and thus provide a fruitful ground for empirical testing. Detailed performance and compensation data of production workers (e.g. players or even referees) are basically available in the form of by-products of today's comprehensive media coverage of (live) sporting events. At the same time, it is possible to observe individuals throughout their entire careers and often in employment contracts with different employers.[4]

Despite the already large and still growing sports economics literature, many fields still remain untapped and numerous questions remain unanswered. The vast majority of empirical works has focused on the North American Major Leagues for American Football (NFL), Basketball (NBA), Baseball (MLB) and Hockey (NHL), as well as on European football (soccer). Other, less popular sports have received less attention or have not been considered at all. Yet, it seems particularly worthwhile to examine individual behavior in competitive environments that are perhaps distinctly different from "common" athletic contests. Given, for example, the presumably less dramatic exit barriers for professional athletes who are active in a niche sport (i.e., with less money involved in the sport itself, the outside options should become relatively more lucrative), individuals might have significantly shorter careers than professional athletes in other (major) sports. Women's sports, typically receiving less attention from the media and, thus, providing lower incentives for females to pursue a (professional) sports career, is an interesting and so far neglected field of research.[5] Moreover, while most of the existing literature has focused on the determinants and consequences of individual performance of the athletes *themselves*, little research has been done regarding the determinants and consequences of the monitors' (i.e., referees') performance. Exploiting some of the as yet untapped research potential, the present work is set out to empirically analyze individual behavior in competitive, tournament-like environments.

[4] See Alchian (1988), Kahn (2000), Rosen and Sanderson (2001) and Frick (2004) for further arguments in favor of sports data.

[5] This does of course not apply to all women's sports. Women's tennis, for example, receives similar attention as the men's professional circuit while prize money levels and breakdowns have been equalized for all so-called Grand Slam tournaments.

This dissertation, which consists of five separate studies to be published in peer-reviewed academic journals in the field of personnel and sports economics, can basically be divided into two parts. The first two studies examine individual performance and career outcomes. In this part, an analysis of the determinants of individual career duration is provided while a particular focus is placed on how individual performance fosters or impedes career progress. The remaining three studies analyze gender differences in competitive environments. As will be discussed in more detail in the respective chapters, a large part of the literature suggests that women are, inter alia, less competitive and more risk averse than – equally endowed – men and that these differences can, to some extent, explain the observed gender wage gap as well as the underrepresentation of females in leading positions. Hence, the second part of this work aims to shed light on this topic by examining the (socio-) economic causes and consequences of gender differences. Aside from that, the development of the gender gap in competitiveness is examined over a longer time period. This, in turn, gives some indication of how institutional changes or changing socio-cultural conditions might lead to a reduction of the gender gap over time. While the first four studies use data from various sporting competitions where individuals are assumed to be rational and utility-maximizing, the last study is based on data from a high-stakes TV quiz show where candidates show signs of bounded rationality in their decision making. All datasets used in the analyses have been carefully compiled by the authors of the studies and represent unique and hitherto unavailable data bases for empirical testing. The econometric methodologies applied have been tailored to both the peculiarities of the respective datasets and the specific research questions and will be explained in more detail in the respective analyses.[6]

The remainder of this work is organized as follows: Chapter 2 presents joint work with Bernd Frick where we investigate the impact of individual performance and of competitive pressure on the duration of ski jumpers' careers. Our dataset includes all athletes who managed to win World Cup points in at least one competition in the seasons 1979/80-2010/11. It appears that individual careers are of a rather short duration (four years on average). Almost 50 percent of the athletes have left the circus again after two seasons and only about 10 percent manage to survive for ten years and more. First and not surprisingly, individual performance has the expected impact on career duration (the more World Cup points an athlete accumulates during a season, the less likely he is to terminate his career).

[6] In order to avoid redundancies across the separate studies, recurring econometric methods are only introduced once.

Second, the degree of competition in the respective national associations also has a statistically significant impact on individual career duration: Poorly performing athletes have a rather high survival probability if they cannot be replaced (i.e., if the jumpers with whom they compete for the limited number of spots in their national team perform even worse). Third, we find that "superstars" – (former) World and Olympic Champions – have significantly longer careers and that their nominations seem to be justified more by their past success than by their current performance. This, in turn, points at potential labor market inefficiencies that need to be investigated more closely.

In another joint paper with Bernd Frick, chapter 3 focuses on the labor market of (hitherto semi-professional) football referees. Using performance evaluations on all 89 referees who officiated in the 3,968 matches played in the first three divisions of German professional football (soccer) in the seasons 2008/09-2011/12, this chapter seeks to identify the impact of individual performance on career progress (and career impediments) of German football referees. We contribute to the literature on favoritism in organizations by explicitly examining the consequences (rather than the causes) of biased assessments. Given the abundant information available, the labor market for football referees is a particularly suitable setting for such an analysis. Since until the start of the 2012/13 season referees were paid on a match basis only, it is reasonable to assume that (i) rational individuals try to maximize the number and importance (i.e., profitability) of matches they are assigned to, and that (ii) relatively better performing referees are nominated more frequently and have better career prospects than poorly performing referees. Moreover, it is conceivable that particularly poor performances entail immediate sanctions in the form of compulsory breaks or a temporary demotion to a lower division. Perhaps surprisingly, we find no evidence for any short-term performance effects: Neither a referee's waiting time nor the quality of his subsequent nomination is determined by the performance in the previous match. Yet in the long term, referee performance has a statistically significant and positive impact on the probability of being promoted to a higher division, thus supporting tournament theoretical predictions.

Turning to the research of gender differences in competitive environments, chapter 4 begins with a long-term analysis of performance differences between highly self-selected male and female long-distance endurance athletes. Using repeated cross-sectional data from three prestigious long-distance races, the empirical evidence presented in this chapter suggests that although the performance gender gap between the most talented male and

female athletes seems to have stabilized (i.e., among the top athletes, men are consistently about 10% faster than women), the vast majority of women competing in culturally diverse environments has become more competitive over time, causing gender differences beyond the top positions to decline. This might, to some extent, reflect changing socio-cultural conditions that reduce discrimination of females in accessing leisure time and enable ambitious women to train as intensively as equally talented men. As expected, gender differences between culturally homogeneous athletes from mostly gender-equal societies (i.e., Scandinavian countries) show significantly smaller changes over time. Perhaps surprisingly, here the performance gender gap seems to have slightly widened across years. A possible explanation is that saturation and substitution effects matter in the sense that women today are less eager (than men) to engage in traditional (and rather male-dominated) sports and leisure activities, but instead prefer more "trendy" pastimes. This remains to be tested in future research.

Chapter 5, co-authored with Bernd Frick and Wiebke Held, addresses a peculiarity of the professional team sport industry that has been emphasized by some of the founding fathers of sports economics research (e.g. Rottenberg 1956; Neale 1964; Sloane 1971): Unlike purely profit maximizing firms in a "traditional" market, firms (i.e., teams) in the sports industry to some extent depend on the performance of their competitors (while the survival of the latter is of pivotal importance for their own survival). Thus, in leagues with win maximizing or, more generally, utility maximizing teams, some degree of outcome uncertainty is said to be needed in order to sustain fan interest in the form of stadium attendance and TV ratings. Drawing on repeated cross-sectional data covering a period of 21 years (seasons 1990/91 – 2010/11), chapter 5 analyzes the long-term development of competitive balance in selected European team sport leagues. While most research in this strand of the sports economics literature has focused on men's professional team sports, the purpose of this study is to provide a comparative analysis of men's and women's leagues. Thereby, we aim to measure competitive balance levels of the following men's and women's leagues, respectively: English "Premier League" and "Football League" (i.e., first and second division English football), German football "Bundesliga" as well as German "Handball Bundesliga". We find statistically significant gender differences in all three football leagues. That is, men's leagues appear to be more "balanced" than women's leagues. This, in turn, is indicative of a more equal distribution of talent among men as compared to women. The results of the Handball Bundesliga are somewhat different. Here, the gender gap seems to

have diminished over time to the effect that overall (almost) no statistically significant differences in competitive balance are found.

As already mentioned, chapter 6 differs from the previous chapters in that the data used in the analysis are not gathered from athletic contests but from a TV quiz show. In this joint paper with Julia Nagelschneider, we analyze gender differences in high-stakes decision making under risk. Drawing on data from the UK TV quiz show "The Million Pound Drop" (as well as the German and Swiss equivalents, "Rette die Million" and "Die Millionenfalle"), we test, first, if male teams are less risk averse than female teams, and second, if male candidates are overconfident particularly when playing with a female team partner. The results suggest that team composition has a significant influence on the teams' risk behavior which, in turn, affects "survival" chances of teams. Men are found to be less risk averse, since they bet larger shares of their (remaining) budget on the answer option which they believe to be correct. Women, on the other hand, prefer to diversify their bets. Moreover, men are found to be overconfident when playing with a female partner. That is, in two-player mixed teams men act as sole decision-makers significantly more often than in men's teams, even though this behavior is not associated with a higher expected payoff. These results are in line with the literature so far. Yet, whereas previous research and in particular laboratory experiments were often criticized for their lack of external validity, we are able to observe individual decision making in a real-life setting. Moreover, the exceptionally high stakes in this game show are assumed to reveal the candidates' true preferences. From an economic perspective, these findings have important implications for management and can, to some extent, explain the underrepresentation of women in top management positions. Finally, chapter 7 summarizes the main results and concludes with some general implications as well as a plea for future research.

2 FLY LIKE AN EAGLE: CAREER DURATION IN THE FIS SKI JUMPING WORLD CUP

2.1 INTRODUCTION

Although still a "niche sport" in many parts of the world ski jumping is quite popular in Europe as well as Japan. Invented by Norwegian soldiers more than 200 years ago, ski jumping has been part of the Olympic program since the inaugural Olympic Winter Games in Chamonix in 1924. Apart from the competitions held during the Olympics and the official World Championships (starting in 1924, too), the sport features a well-established and largely televised professional circuit, with the season's peak event being the "4-hills-tournament" taking place around the turn of the year. Since the 1979/80 season, all major events (i.e., the so-called World Cup events) offer prize money to a limited number of athletes.[7] Although prize money levels are considerably lower than in other (major) sports, professional ski jumpers usually earn far more than they could earn in their second best alternative. Since many of the athletes have started to train professionally at a rather early age, few – if any – of them have invested in education and training, implying that the opportunity costs of quitting the career are quite high. Therefore, the majority of the athletes make an effort to stay in the sports business for as long as possible (including potential engagements as a manager or head coach after retiring from an athletic career). However, the permanent risk of injury on the one hand and intense competition from other athletes on the other hand make "survival" difficult. In particular the competition with athletes from the same country can represent a considerable threat for athletes from particularly strong "ski nations": In every World Cup competition, each national federation is guaranteed a limited number of slots only (currently a maximum of 7 per nation). Nations with a "tradition" in ski jumping (e.g. Norway and Austria) usually dispose of a large pool of world-class athletes of whom only the top contenders can be selected.[8] On the other hand,

[7] Prize money levels in FIS ski jumping are substantially lower than those of other (major) sports. The winner of a World Cup event receives 10,000 CHF, while 8,000 are awarded to the 2nd place, 6,000 to the 3rd, 5,000 to the 4th, 4,500 to the 5th, 4,000 to the 6th, 3,600 to the 7th, […], 300 to the 28th, 200 to the 29th and 100 CHF to the 30th. In the season 2010/11, Thomas Morgenstern topped the annual earnings list with 213,200 CHF, followed by his Austrian teammate Andreas Kofler (150,300 CHF) and four-time Olympic gold-medalist Simon Ammann from Switzerland (136,400 CHF). Endorsement contracts are usually not disclosed, but are said to be far more lucrative than prize money.

[8] Hence, similar to e.g. the US and Jamaican sprint trials or the domestic qualifying events for marathon runners in Ethiopia and Kenya, a nomination for the Austrian or Norwegian World Cup ski jumping team is presumably more difficult to achieve than a spot among the Top 30 in a given World Cup competition.

equally talented ski jumpers from rather "weak" nations (who are barely threatened by the performance of their compatriots) might be able to survive in the circus for quite a while.

Given the detailed data available (which can be accessed via www.fisskijumping.com) and the idiosyncrasies with regard to the restrictive nomination criteria, professional ski jumping is a particularly interesting setting to empirically test the (potential) determinants of the length of professional athletes' careers. Notwithstanding these characteristics, the career length of professional ski jumpers has, to the best of our knowledge, not yet been empirically analyzed. We try to close this gap in the literature by analyzing the impact of individual performance and of competitive pressure on the duration of ski jumpers' careers. For this purpose, we use data including all athletes who managed to win World Cup points in at least one competition in the seasons 1979/80-2010/11 (n = 766 different competitions and 698 athletes with 2,646 athlete-year-observations).

The remainder of the paper is organized as follows: Section 2.2 provides a (selective) review of the literature on the determinants of career lengths of professional athletes. In section 2.3 we describe the data used, explain our methodology and display some descriptive results. The econometric evidence is presented in section 2.4 while some implications for further research are presented in the concluding section.

2.2 LITERATURE REVIEW

The statistical analysis of survival data was initiated by bio-medical scientists computing so-called "life tables" to illustrate the distribution of survival times of, e.g., cancer patients (see, inter alia, Berkson and Gage 1950; Cutler and Ederer 1958). In the following years, the initial "survival analysis" was extended to include more complex econometric methods while at the same time the scope of applications increased. In labor economics, for example, duration data have been used to analyze the length of employment spells (see Flinn and Heckman 1982a) as well as the duration of unemployment (see Kiefer and Neumann 1979; Lancaster 1979; Nickell 1979; Flinn and Heckman 1982b). For sports economists, survival analyses are a useful tool when examining the career length of professional athletes. Given the high salaries of NFL, NBA or European soccer players, the individual opportunity costs of quitting are almost prohibitively high. Hence, from a microeconomic (i.e., athlete's) perspective the analysis of the determinants of individual career length is of considerable economic relevance.

In a now seminal paper Atkinson and Tschirhart (1986) use information on 260 players from the National Football League (NFL) who were active between 1971 and 1980 (with an average career length of 4.5 years) to estimate a series of hazard models. They find that rookies experience an increasing hazard during their first years whereas those who survive the middle "shakeout" years benefit from a falling hazard rate for the remainder of their careers.[9] As expected, both team and individual performance have a positive effect on a player's career length. Quarterbacks survive longer than either wide receivers or running backs. Moreover, the authors find that in terms of career length racial discrimination was not an issue in football as black players without a college degree tend to survive longer than white players without a degree, while black players with a college degree have, on average, shorter careers than observationally similar white players.

Spurr and Barber (1994) analyze the careers of 608 baseball pitchers who began their careers in the minor leagues during the years 1975-1977 of whom only 94 (15%) eventually signed a contract with a major league team between 1975 and 1988. Their results suggest that the more a player's performance diverges from the mean in either direction, the less time is required to make a decision on a player's transition (i.e. a promotion to a team playing in a higher division or a demotion to a team in a lower league). Decisions on the fate of marginal players, on the other hand, take much longer. Moreover, they find that once a threshold amount of information about a player's quality is available, the marginal gains from additional information decrease rapidly. Accordingly, a player's probability of exiting his current state – be it by promotion, demotion or termination – increases with time.

Ohkusa (2001) analyzes the effect of income and productivity on the length of player careers in Japanese professional baseball (n = 595 batters and 350 pitchers). As expected, a higher income is associated with a lower exit probability for both batters and pitchers. Perhaps surprisingly, a higher productivity reduces the exit probability for batters, but significantly increases the exit probability for pitchers. This rather unexpected result is attributed to a finding presented by Ohkusa in another paper (1999) documenting that the estimated learning curve of productivity is a decreasing function of experience for pitchers (i.e., pitchers perform best in their youth), whereas it is an increasing function for batters (whose productivity increases with experience).

[9] Professional athletic careers in general seem to be skewed towards early exit (see e.g. Witnauer et al. 2007).

Staw and Hoang (1995) examine the survival probabilities of NBA players picked in the first two rounds of the 1980-1986 drafts. Using a sample of 275 players who played at least one year in the NBA (of whom 184 were cut from the league within the period of observation until the season 1990/91)[10] they find that scoring performance and "toughness" (measured by the number of rebounds and blocks) have a statistically significant and positive impact on career length, whereas draft number and the number of times a player has been traded have a significantly negative impact on career duration. Using the same sample of NBA drafts active between 1980 and 1991, Hoang and Rascher (1999) present robust evidence for several forms of discrimination against black players. First, they find that black players earn significantly lower salaries than similarly endowed white players. Second, black players are found to have a 36% higher risk of being cut than white players, which translates into a significantly shorter expected career length of 5.5 seasons (as opposed to 7.5 seasons for white players). Third, while the net present value of wage discrimination amounts to $329,000, the career earnings effect of exit discrimination is considerably larger ($808,000) and is attributed by the authors to customer racial discrimination.[11]

In an attempt to shed additional light on the much debated question of whether men and women differ in their competitive orientations Coate and Robbins (2001) study the career length of male and female tennis professionals. Their sample consists of 236 male and 216 female tennis players who made it into the top 50 in the singles ranking at least once in their respective careers between 1979 and 1994. The results suggest that despite significantly lower real earnings[12] women pursue an active career for as many years as men and compete just as intensely (measured by the annual number of tournaments played).

[10] Drafted players survived, on average, six seasons. However, second round draft picks' careers were 3.3 years shorter than those of first round draft picks.

[11] Using two large unbalanced panels from the NBA and MLB in more recent years, Groothuis and Hill (2004, 2008 and 2013) fail to find evidence of salary or exit discrimination in either of the two leagues. Their results suggest that, in line with Becker's (1971) theoretical work on discrimination, market competition (fostered by objective and easily accessible performance indicators) induced discrimination to disappear.

[12] We emphasize that a large part of the earnings gap can be explained by differences in prize money levels for male and female tennis players at that time. Meanwhile, a number of events – particularly the four "Grand Slams" – have implemented identical prize money levels and distributions for men and women. Nevertheless, the gender pay gap in professional tennis seems to persist. As of November 2012, the annual top 100 female earners had accumulated $83 million in total and hence 20% less than their male counterparts who totaled $104 million (www.tennisdigital.com, www.wtatennis.com and own calculations). We admit, however, that by looking at the top 100 male and female earners we ignore a potential bias: If female tennis professionals were more heterogeneous than men in terms of their ability, the lion's share of the tour's prize money would be taken by few women, whereas on the potentially more homogeneous men's tour the "cake" would be divided among a larger number of players. Hence it appears worthwhile to look at the earnings gap between the top 20, top 100 and top 500 male and female tennis players in order to better understand their respective level of competitiveness.

Using a dataset that includes every single player who appeared in at least one match in the top-tier of German professional soccer ("Bundesliga") in the seasons 1963/64 to 2002/03 (n = 4,116), Frick et al. (2007, 2009) find that defenders, midfielders and forwards have significantly shorter careers than goalkeepers. Moreover, the number of games played and goals scored per season reduce the hazard significantly, whereas disciplinary sanctions (i.e., yellow and red cards per season) seem to have no impact on career length. As expected, a team's demotion to the second division increases the hazard dramatically as only few players of relegated teams manage to sign with another first division club. Foreign players (especially those from Eastern and Western Europe as well as from South America) have a significantly higher exit probability. This finding may, to some extent, be attributed to discrimination in the sense that managers, "co-workers" or spectators prefer players of German origin. On the other hand, higher hazard rates do not necessarily imply discrimination. A large part of the effect is presumably driven by more lucrative outside options especially for South American and Western European players who leave the Bundesliga voluntarily to sign contracts with teams in England, France, Spain and Italy.[13]

In a similar vein, Boyden and Carey (2010) examine the length of player careers in Major League Soccer (n = 1,100 players; 1,166 spells; 3,435 player-year-observations between the years 1996 and 2007) finding that MLS careers are rather short (2.4 years on average). Again, the data do not allow distinguishing between players who are forced to exit the league (i.e. did not receive a new contract) and those who are lured away by foreign clubs. Anecdotal evidence suggests that the latter effect dominates. Since especially the Top 5 European soccer leagues pay far higher salaries than MLS (average annual salary in the German Bundesliga currently is 1.5 million €; see Frick 2011), it is no surprise that MLS has even greater difficulties than the German Bundesliga to retain foreign talent in the league.

In this paper we apply parametric as well as semi-parametric duration analysis to a unique dataset covering a hitherto unexplored individual sport – ski-jumping – where the forces of competition are completely different from other sports labor markets.

[13] Players from Eastern Europe are less popular among supporters, as shown in Kalter (1999) who finds that replica shirts with the names of South American players are bestsellers, while those with the names of observationally similar (in terms of experience, goals, games played, etc.) Eastern European players do not sell well.

2.3 DATA, METHODOLOGY AND DESCRIPTIVE RESULTS

In the following sections, we present the data, explain the methodology used in the analyses and display some preliminary – and rather descriptive – results. Moreover, we discuss the peculiarities of the labor market for professional ski jumpers.

2.3.1 DATA COLLECTION

Our empirical analyses are based on an unbalanced panel from the FIS Ski Jumping World Cup covering all athletes who managed to win World Cup points in at least one competition during the seasons 1979/80-2010/11.[14] The final dataset includes 698 individual athletes competing in 766 different competitions, yielding 2,646 athlete-year-observations. Since the introduction of the World Cup in 1979/80, an average of 80 different athletes have managed to win World Cup points in approximately 24 World Cup events per season. It should be noted, however, that both the number of athletes finishing "in the points" as well as the number of World Cup events per season increased considerably following a modification of the points regime before the start of the 1993/94 season (the number of "in the points ranks" was increased from 15 to 30). Figure 2-1 below illustrates the substantial increase in the number of World Cup events starting in the 1993/94 season (the volatility in the number of events is due to weather conditions; sometimes scheduled events have to be cancelled on rather short notice). In the seasons 1979/80-1992/93, the FIS hosted on average between 22 and 23 ski jumping World Cup events per season. Due to the increasing popularity of the sport in recent years, the average number of World Cup events per season was increased to about 25 (min. 19, max. 30). This increase, along with the already mentioned modification of the points regime, constitute two important institutional changes that positively affected the number of athletes receiving World Cup points (and, not to forget, the number of athletes receiving financial rewards).

[14] The data were obtained from www.fisskijumping.com and http://de.wikipedia.org/wiki/Skisprung-Weltcup.

Figure 2-1: Number of World Cup events per season

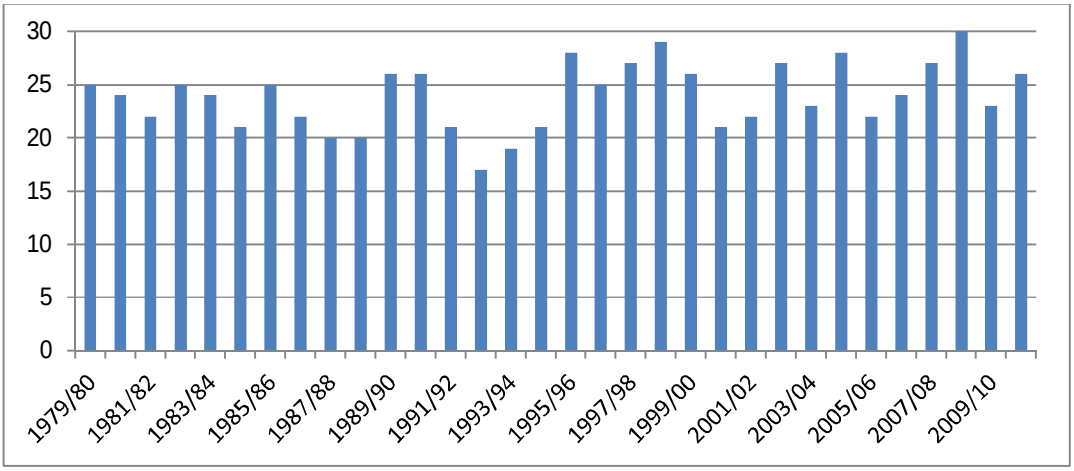
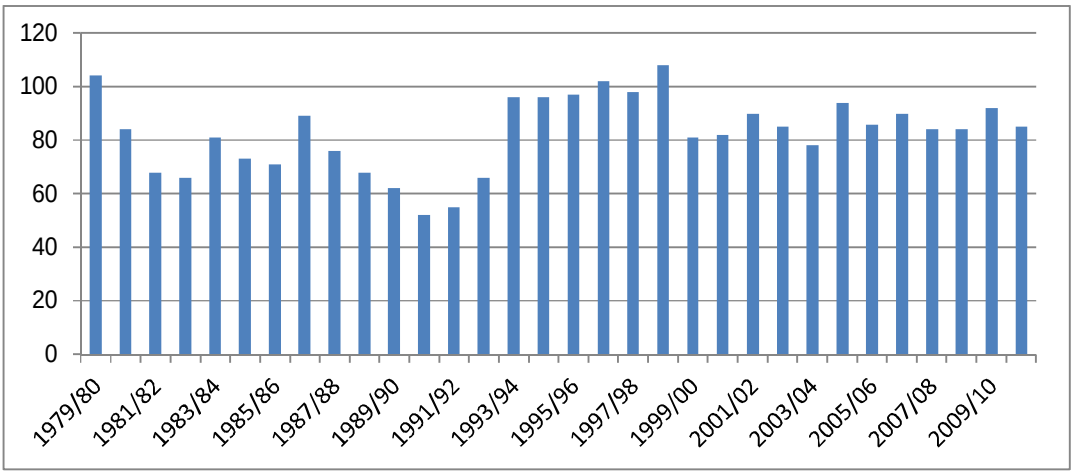


Figure 2-2: Number of athletes winning World Cup points



It appears from Figure 2-2 above that the modification of the points regime before the start of the 1993/94 season had, as expected, an immediate and positive effect on the number of athletes receiving World Cup points. Moreover, the continuously increasing number of World Cup competitions in the following seasons led to a further increase in the pool of athletes achieving a top 30 performance in at least one event in a particular season. In the last decade, however, it seems that the number of athletes receiving World Cup points has leveled at about 80 – irrespective of the annual number of World Cup events – suggesting that a relatively large number of athletes remain unsuccessful. In other words: Despite the large number of approximately 25 World Cup competitions per season, many professional ski jumpers fail to finish “in the points” / “in the money” at least once in a given year. This

development, in turn, may be attributed to the changing nature of the qualification system that we will now describe in more detail.

2.3.2 THE PECULIAR NATURE OF COMPETITION IN PROFESSIONAL SKI JUMPING

The average career length of all athletes in our sample is 4.0 seasons, with a minimum career duration of 1 season (n = 228 athletes) and a maximum career duration of 20 seasons (n = 1 athlete; now 41 year-old Noriaki Kasai from Japan). Perhaps surprisingly, about 30 percent of all careers lasting for two or more seasons consist of more than one spell, i.e. athletes disappear from the circus and return after one or more seasons. Clearly, few of these athletes in fact interrupt their careers – be it for personal reasons or due to serious injury (Norwegian Anders Jacobsen is a prominent recent example for an athlete who took a year off and returned to become even more successful than he used to be before his "temporary retirement"). Apart from these rare voluntary exits we interpret exits (and the subsequent re-entries) as products of the peculiar nature of competition in FIS ski jumping, where athletes constantly face a number of "competitive threats": First, the tournament structure of a World Cup event exerts considerable competitive pressure particularly on the less talented athletes: Every competition is designed as a (knockout) tournament consisting of three rounds where apart from the top 10 in the World Cup ranking (they are guaranteed entry to the first (main) round) the remaining athletes have to qualify for entry into the competition to which 50 athletes are admitted (since 10 athletes are seeded, the rest compete for 40 vacant sports).

To identify the top 30 athletes who advance to the final round, two alternative formats are currently being used: a knockout-tournament where athletes compete pairwise with the 25 winners plus the five best performing "lucky losers" advancing to the second round, and a rank-order tournament where the 30 best ski jumpers reach the second round.[15] Hence only about 30 percent of the athletes competing in each World Cup event eventually vie for World Cup points and financial rewards (see footnote 7 above for the exact prize money level and distribution in a World Cup event). Second, apart from the competitive pressure to succeed in the tournament and to secure a spot among the top 30, weaker athletes face a high risk of being relegated to the less attractive second division, the "Continental Cup",

[15] These qualification criteria were introduced in the season 1990/91. Prior to that season, all athletes competing in the first round (often more than 100) were admitted to the second round. The main reason for the new format was to make the sport more attractive for TV stations by reducing the duration of the competition.

which was introduced before the start of the 1993/94 season. Third, since each national federation is guaranteed a limited number of slots only (currently a maximum of 7 per nation), athletes from strong federations (e.g. Austria and Norway) face a substantially higher risk of exit (or relegation) than athletes from weak federations (e.g. France and Italy). Thus, an athlete's career length is not only affected by his individual performance, but also by the performance of his compatriots. Strong athletes from nations with a particular tradition in ski jumping are, therefore, threatened by relegation while even rather weak athletes from weak nations may be able to survive in the circus for quite a while.[16]

2.3.3 Parametric and Semi-Parametric Estimation of Career and Spell Duration

In an attempt to examine the impact of individual performance and of competitive pressure on the duration of professional ski jumpers' careers we estimate two semi-parametric proportional hazard models (see Cox 1972) as well as two log-normal regression models using career duration and spell duration respectively as the dependent variable. Descriptive statistics are displayed in Table 2-1.

[16] A prominent example is Michael Edwards, better known as „Eddie the Eagle": As the first and, at that time, only active British ski jumper he qualified for the 1988 Winter Olympics in Calgary despite being handicapped by his weight (he was 9 kg heavier than the next heaviest competitor) and his extreme farsightedness which forced him to wear his (often fogged) glasses at all times. In both Olympic competitions that he participated in (the event from the normal and the large hill) he finished last. Irrespective of his permanent lack of sporting success he managed to capitalize on his increasing popularity and appeared in a number of well-paid advertising campaigns. In 1990, the International Olympic Committee reacted to the "Edwards phenomenon" by introducing stricter qualification norms for the Olympic ski jumping competitions.

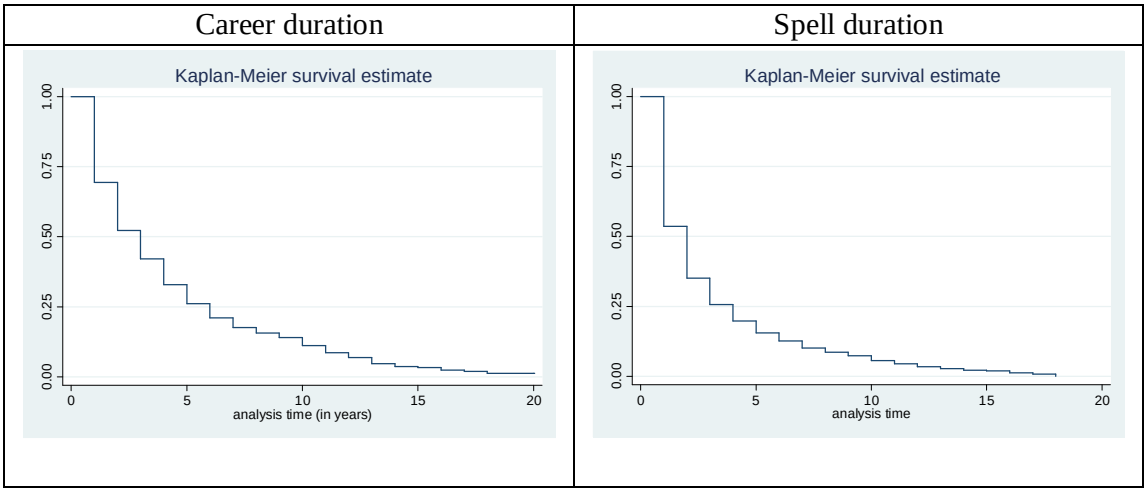
Table 2-1: Descriptive statistics of the determinants of professional ski jumpers' careers

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLES						
Career end	Dummy: Equals 1 if athlete terminates his career, 0 otherwise	2,646	0.23	-	0	1
Spell end	Dummy: Equals 1 if athlete terminates or interrupts his career, 0 otherwise	2,646	0.33	-	0	1
INDEPENDENT VARIABLES						
Performance (-related) Indicators						
World Cup points	Number of individual World Cup points in $season_j$	2,646	136.6	257.6	1	2,083
N_Points	Standardized number of individual World Cup points in $season_j$	2,646	0.00	1.47	-1.00	8.94
N_Points2	Square value of n_points	2,646	2.15	6.31	0	79.98
N_Points3	Cubic value of n_points	2,646	7.62	42.15	-0.99	715.25
Olympic Champion	Dummy: Equals 1 if athlete has won an Olympic gold medal in an individual event to date, 0 otherwise	2,646	0.04	-	0	1
World Champion	Dummy: Equals 1 if athlete has won a World Champi-onship title in an individual event to date, 0 otherwise	2,646	0.09	-	0	1
CONTROL VARIABLES						
Athlete ID (career duration)	Individual identification number by athlete	2,646	-	-	1	698
Athlete ID (spell duration)	Individual identification number by (uninterrupted) spell	2,646	-	-	1	965
Time trend	Linear time trend (1 corresponds to season 1979/80 etc.)	2,646	17.04	9.27	1	32
Nation's share of points	Relative share of cumulated individual World Cup points of $nation_k$ in $season_j$	2,646	10,83	8.10	0	34.18
Points regime	Dummy indicating the period before (0) and after (1) the modification of the points regime	2,646	0.62	-	0	1
No. of World Cup events	Number of World Cup events in $season_j$	2,646	24.00	2.92	17	29

Before discussing our set of explanatory variables, we present descriptive evidence based on a nonparametric estimation of the survivor function which is commonly referred to as the product-limit or Kaplan-Meier (1958) method. The Kaplan-Meier estimates yield the percentage of all athletes who are still in the sample after analysis time *t*. In our case, about 30 percent of all careers end after one season and only about half of the athletes manage to survive on the tour for more than two seasons. 65 athletes (about 10 percent) survive for ten or more seasons. When looking at the length of uninterrupted careers (i.e., spell duration) it appears that almost 50 percent of all spells last for only one season while only one third of all spells are longer than two seasons. Of the 65 athletes who manage to win World

Cup points in ten or more different seasons, 43 jumpers have uninterrupted spells of at least 10 seasons.

Figure 2-3: Kaplan Meier survival estimates



Overall, average career length is 4.0 years (std dev = 3.3; n = 698 athletes) while average spell length is 3.3 years (std dev = 3.0; n = 965 spells). These figures are in line with the evidence presented in Frick et al. (2007, 2009) who report almost identical career and spell durations for soccer players in the German "Bundesliga" (4.0 years (std dev = 3.3); 3.4 years (std dev = 2.9)). Interestingly, a large part of the career duration literature finds that the career lengths of professional athletes in individual as well as team sports range between three and seven years (see e.g. Atkinson and Tschirhart 1986; Staw and Hoang 1995; Hoang and Rascher 1999; Ohkusa 2001; Groothuis and Hill 2004, 2008, 2013; Witnauer et al. 2007; Boyden and Carey 2010; Gibbs et al. 2012). Thus, irrespective of the particular design of competition, of injury risks and of the (potential) financial rewards, individual careers appear to be relatively short across different sports.

2.3.4 COVARIATES

The goal of our empirical analysis is to identify the covariates that have a statistically significant effect on the career length of professional ski jumpers. As a first measure of individual performance we use the number of World Cup points accumulated by $athlete_i$ in $season_j$. To control for the observable variation in the number of World Cup competitions per season (see Figure 2-1 above) as well as for the already discussed changes in the points regime (see Figure 2-2 above) we calculate for each athlete the relative number of World Cup points as the deviation of athlete i's points in season j from the average number of

points accumulated by all athletes in season j. As we expect a non-linear relationship between individual performance and career length we also include in our estimations the squared and the cubic term of relative number of points.

Second, we include two dummy variables indicating whether an athlete has ever won an Olympic gold medal or a title as World Champion in an individual ski jumping competition (we deliberately exclude team competitions). The interaction of these dummies with an athlete's individual performance enables us to test for potential **superstar effects**. More specifically, we analyze whether (former) superstars' nominations into the (usually rather small) national squads are justified by their current performance or rather driven by their past success (perhaps accompanied by the team manager's hope for a positive outlier in performance).

Third, we are interested in the impact of **competitive pressure** on the duration of ski jumpers' careers. In order to test our hypothesis that athletes from rather weak nations tend to "survive" longer than equally talented athletes from strong nations (because the latter are permanently challenged by their compatriots), we calculated two different measures reflecting the different nations' level of competitiveness: (*i*) a concentration ratio measuring the dispersion of individual World Cup points within $nation_k$ in $season_j$, and (*ii*) the relative share of cumulated individual World Cup points of $nation_k$ in $season_j$. Although plausible, option *(i)* does not adequately reflect the level of a specific nation's competitiveness, because concentration measures such as the Gini coefficient are likely to yield biased results.[17]

Assume, for example, a strong nation with a large pool of talented athletes. Here the head coach is likely to rotate some of the athletes, while in a weak nation with a small pool of talented athletes every athlete has a chance to be nominated for every single event. Consequently, a weak nation sending the same athletes to all World Cup events in a particular season appears to have a more homogeneous distribution of talent than a strong nation in which e.g. three athletes are guaranteed their nomination while the remaining three or four spots are permanently reallocated among equally talented athletes from a large pool (see Appendix A for an illustrative example).

[17] The Gini coefficient has been (and continues to be) widely used to measure, inter alia, the degree of competition in individual industries (e.g. Simon and Bonini 1958), price dispersion in certain markets (see e.g. Borenstein and Rose 1994 for evidence from the airline industry or, with completely different results, Gerardi and Shapiro 2009) and income inequality using micro- (e.g. Andrews and Leigh 2009) as well as macro-data (e.g. Solt 2009 and Malul et al. 2013).

Supposing that "strong" (i.e. successful) nations have, ceteris paribus, a larger pool of (equally able) athletes than weaker nations, we use each nation's relative share of World Cup points per season as a proxy for competitive pressure within a country's team. Moreover, we include a **time trend** that controls for the (presumably increasing) opportunity costs of retiring early (the availability of prize money as well as endorsement contracts make professional careers even more attractive than they used to be in the past).[18]

2.4 EMPIRICAL RESULTS

The Cox model is a now well-established econometric tool to analyze right-censored and time-dependent data with two distinctive advantages compared to other proportional hazard models (Kiefer 1988). First, it allows including the information of right-censored data – a feature that is of particular importance in our context as many of the athletes in our dataset are still active at the end of the observation period, i.e. the end of the season 2010/11. For these athletes the event to be explained by the covariates – exit from the World Cup tour – has not yet occurred and the observation is, therefore, right-censored. Second, since it is the most general of the available regression models, the Cox model does not make any assumptions about the nature or the shape of the underlying survival distribution[19] and can integrate time-dependent variables. Moreover, semi-parametric models like the Cox model can also handle left truncation, also known as delayed entries. Left truncation occurs when individuals included in the sample arrived in the "risk pool" already prior to the start of the observation period. These subjects (e.g. ski jumpers who had already been active before the introduction of the World Cup) can be included in the study. However, their failure time must go into the econometric model without truncation (see Cox and Oakes 1984 and Cleves et al. 2008 for detailed analyses on the effects of including left-censored data).[20]

[18] We admit that, in principle, we should also control for a number of additional rule changes that have been enacted by FIS in the recent past, such as the implementation of a minimum body weight, the regulation of the size of the ski jumping suits and the length of the skis, because all these changes may have affected career length in one way or the other. The problem here is that some athletes may have benefited from these rule changes at the expense of others. Unfortunately, information on e.g. the body weight of the athletes is not available, implying that we cannot control for the impact of these rule changes on the length of the individuals' careers.

[19] Since the shape of the survival distribution is left unspecified, only a functional form for the influence of the covariates can be specified (Blossfeld and Rohwer 2002: 228).

[20] In our sample, the first year of the observation period coincides with the inaugural season of the FIS Ski Jumping World Cup. Although a certain percentage of the athletes in our sample had been active before this season (in e.g. Olympic Games and/or World Championships), none of them had ever competed in a World Cup competition. Hence our data are less susceptible to left-truncation bias.

The advantages of the Cox semi-parametric proportional hazard models notwithstanding, we continue estimating two additional parametric models. As suggested by the *Akaike Information Criterion* (Akaike 1974) we estimate a log-normal model (i.e., we assume that the logarithm of the durations follows a normal distribution) which in our case yields more robust and reliable results than, for example, a log-logistic or Weibull (1951) model. Both models are of the following functional form:

$$\begin{aligned} \text{Exit} = \quad & \beta_0 + \beta_1 \text{N\_Points} + \beta_2 \text{N\_Points2} + \beta_3 \text{N\_Points3} + \beta_4 \text{Olympic Champion} + \beta_5 \text{World Champion} + \beta_6 \text{OC} * \text{N\_Points} + \beta_7 \text{WC} * \text{N\_Points} \\ & + \beta_8 \text{time trend} + \beta_9 \text{nation's share of points} + \beta_{10} \text{nation's SoP} * \text{N\_Points} + \varepsilon_i, \end{aligned}$$

where exit either stands for an individual's career end or the end of an uninterrupted spell, depending on the model specification. In addition to the explanatory variables that were specified in Table 2-1 above, we include a set of interaction terms to disentangle and quantify potential superstar effects and to measure the impact of competitive pressure on individual career outcomes.

In Table 2-2 we present the results of our estimations. We report both hazard ratios and regression coefficients that have to be interpreted differently. The hazard ratios from the Cox models indicate an individual's probability of being eliminated from the World Cup tour during a specific time interval, conditional on having been in the tournament until the beginning of that interval. In this case, the event to be explained is the exit from the tour. A value between 0 and 1 implies a reduction of the exit probability. The closer the hazard ratio is to zero, the less likely is the exit. In the first model, for example, the hazard ratio of 0.466 of the *World Champion* dummy implies a ceteris paribus 53.4 percent higher chance of survival in t_1 for all athletes who have previously won a World Championship title compared to the reference category, i.e. athletes who have not managed to win that title until t_0. Given the magnitude of the impact, this difference is not only statistically significant but also economically relevant. Hazard ratios can be interpreted in a similar way if the explanatory variables are continuous. Looking at our first model, it appears that a one unit increase in the standardized individual World Cup points in t_0 increases the survival probability in t_1 by more than 70 percent. Hazard rates larger than 1, on the other hand, reduce the probability of survival. Thus, according to our first model, a one percent increase of a *nation's share of points* decreases the probability of survival for a particular individual by

two percent. That is, the more successful a national team is, the larger the competitive pressure in that particular squad and, hence, the less likely an individual athlete from that particular nation is to survive in the World Cup circus.

Table 2-2: Career and spell duration determinants of professional ski jumpers

	Cox Proportional-Hazard Model (1.1)	Lognormal Model (1.2)	Cox Proportional-Hazard Model (2.1)	Lognormal Model (2.2)
	Career Duration		Spell Duration	
Variables	Hazard Ratio	Coefficient	Hazard Ratio	Coefficient
N_Points	0.279***	0.912***	0.318***	0.845***
	(0.0518)	(0.102)	(0.0378)	(0.0704)
N_Points2	1.423***	-0.281***	1.290***	-0.202***
	(0.157)	(0.0705)	(0.0948)	(0.0443)
N_Points3	0.916**	0.069***	0.958**	0.033***
	(0.0340)	(0.0221)	(0.0165)	(0.00980)
Olympic Champion	0.682	0.456**	0.672*	0.518**
	(0.181)	(0.203)	(0.142)	(0.206)
World Champion	0.466***	0.779***	0.461***	0.990***
	(0.0905)	(0.126)	(0.0769)	(0.120)
OC * N_Points	1.414	-0.315**	1.093	-0.089
	(0.336)	(0.144)	(0.251)	(0.124)
WC * N_Points	1.393*	-0.201**	1.266	-0.234***
	(0.248)	(0.0914)	(0.186)	(0.0641)
Time trend	0.966***	0.028***	0.975***	0.020***
	(0.00392)	(0.00313)	(0.00274)	(0.00233)
Nation's share of points	1.020**	-0.013**	1.014**	-0.008**
	(0.00835)	(0.00532)	(0.00612)	(0.00406)
Nation's SoP * N_Points	1.018*	-0.016**	1.015**	-0.011**
	(0.0100)	(0.00610)	(0.00720)	(0.00435)
Constant		1.366***		1.035***
		(0.0984)		(0.0738)
No. of Subjects	698	698	965	965
No. of Failures	613	613	881	881
Time at Risk / No. Of Obs.	2,646	2,646	2,646	2,646
Wald Chi^2	282.87***	476.21***	381.23***	724.16***

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The results of the lognormal models (Models 1.2 and 2.2) can, in turn, be interpreted like the coefficients estimated in a simple OLS model. Using the same variables as above to illustrate the interpretation we see that a *World Champion* title increases the probability of survival in the career duration model (Model 1.2) by almost 80 percent while a one percent

increase in the *nation's share of points* reduces an athlete's survival probability by 1.3 percent.

Irrespective of the concrete model specification, the effects of most of the independent variables go in the same direction and are highly robust in a statistical sense. For ease of interpretation, we therefore mainly refer to the Cox estimates (Models 1.1 and 2.1). As expected, performance (i.e., the number of standardized World Cup points accumulated by an individual athlete) has a statistically significant and non-linear impact on the probability of survival. That is, the better the performance in a given season the higher the probability of a World Cup appearance in the subsequent season. The hazard ratios of the squared (cubic) terms of World Cup points are larger (smaller) than one, suggesting that the marginal returns to effort are increasing, but at a decreasing rate for the less talented athletes and at an increasing rate for the top contenders. Athletes who have won an Olympic gold medal until t_0 are on average about 30% less likely to terminate or interrupt their career in t_1 than observationally similar athletes without a gold medal. Note, however, that these effects are not (or barely) significant in the Cox estimations but appear to be robust in the lognormal models.[21]

The alleged "superstar effect" is far stronger in the case of World Champions who – as already discussed above – have a substantially higher survival probability than the rest of the athletes. The interaction of standardized individual World Cup points and the World Champion dummy yields mixed results with respect to the level of statistical significance but the hazard ratios and coefficients are consistent and suggest that the status of a World Champion is more important in terms of career length than an individual's current performance in the World Cup. In other words, athletes who have previously won the World Championship have a significantly higher survival probability even if their current performance deteriorates. Since athlete survival is to a large extent determined by the national federations' nominations into their (limited) squads this suggests that (former) superstars are preferred to currently better performing compatriots who lack the glory of past Championship titles. Whether this head coach behavior is due to favoritism, the acknowledgement of previous achievements and/or the anticipation of positive outliers in performance remains to be tested.

[21] We are tempted to attribute these inconsistencies to the relatively small number of observations. Since the Olympic Games take place only every four years and some of the titles have been repeatedly won by the same athletes, only 10 different athletes in our sample have won an Olympic gold medal while 23 different individuals have won the title of a World Champion.

Moreover, the inclusion of a linear time trend in the estimations reveals that survival probabilities – and thus career and spell lengths – have considerably increased over time. Both the hazard ratios and the respective coefficients are statistically significant (at the one percent level) and suggest that from one season to the next individual survival probabilities increase by 2-3 percent. Although we cannot separately control for the many rule changes and modifications in the tournament design that have occurred over time, this suggests that in the aggregate the institutional changes – along with changes in nutrition, medical support and training methods as well as the increasing opportunity costs of exiting – have all worked in the athletes' direction.

Finally, competitive pressure has a statistically significant and negative impact on athlete survival in all model specifications. More specifically, a one percent increase of a nation's share of overall World Cup points per season reduces the survival probability of an individual athlete of that particular squad by 1-2 percent. Thus, the stronger a national team, the greater is the domestic competitive pressure and the less likely its individual athletes are to survive. Along the same lines, the interaction of an individual's standardized World Cup points with his national squad's share of points reveals that equally talented athletes from different nations have significantly different survival probabilities, depending on the level of competitive pressure in the respective national team. That is, athletes from strong nations – such as e.g. Norway or Austria – face a ceteris paribus higher risk of being cut than athletes from e.g. Italy or France who – due to the lack of domestic competitors – are less likely to be replaced by other athletes.

Figure 2-4: Survival curves at mean of covariates

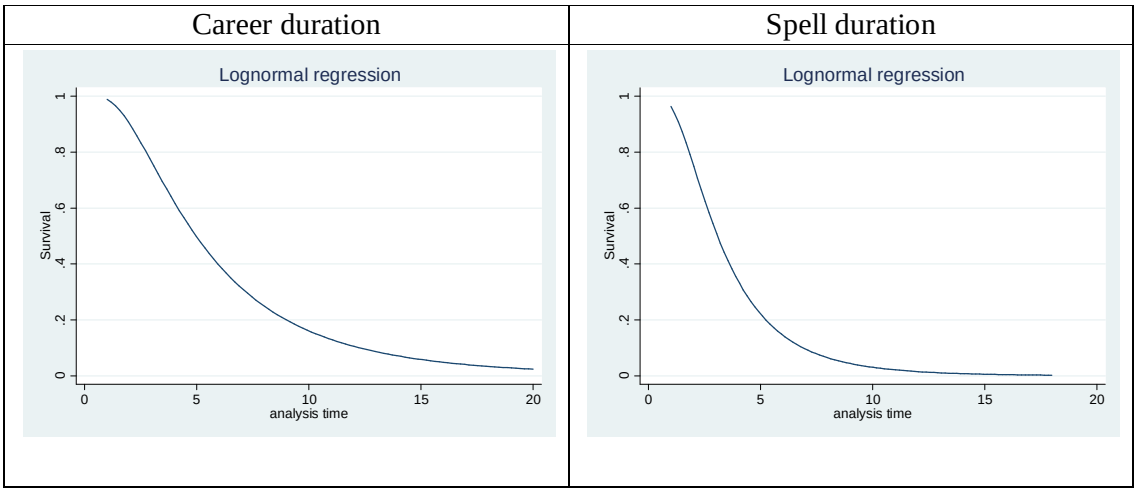


Figure 2-4 displays the survival curves based on the mean of the covariates in the lognormal regression separately for career length as well as for spell length (taking into account interrupted careers). Corroborating the results of the Kaplan-Meier survival estimates (see Figure 2-3 above), it appears that individual careers are of rather short duration. Although the smoothed estimates seem to underestimate the number of early exits, the intuition is that the majority of athletes leave the circuit prematurely while only few individuals are able to survive for ten seasons or longer.

2.5 CONCLUDING REMARKS

Using a hitherto unavailable dataset including almost 700 professional ski jumpers who were active in the FIS Ski Jumping World Cup between 1979/80 and 2010/11, we use parametric as well as semi-parametric duration analyses to identify the (main) determinants of athlete "survival" (and thus individual career length and earning opportunities).

First, and not surprisingly, individual performance has a statistically positive impact on career duration. That is, the more World Cup points an athlete accumulates during a particular season, the more likely he is to return in the subsequent season and hence the longer is his expected career duration. This effect is getting stronger over time as careers last longer today than they used to in the past. Whether this is due to institutional changes affecting the design of the competition or by improvements in nutrition, health and training or due to the increasing opportunity costs of exiting cannot yet be tested due to data limitations.

Second, superstars – (former) World Champions – experience significantly lower hazard rates and thus have on average longer careers (and spells) than athletes failing to win a gold medal in a World Championship tournament. While this result is obviously driven by selection effects (superstars are on average more talented than the rest of the athlete population), we include in our estimations an interaction term yielding a striking result: Athletes who have won a World Champion title face less competitive pressure from their compatriots, as their nomination seems to be driven by their past success rather than by their current performance. Even if their performance deteriorates, (former) superstars are systematically preferred to their (recently better performing) compatriots.

Third, athletes from strong nations with a tradition in ski jumping have significantly shorter careers (and spells) than equally talented athletes from weak nations (who are barely

challenged by the performance of their compatriots). Hence, the nationality of an athlete has a considerable impact on individual survival probabilities.

This competitive pressure, in turn, may simply be the result of the superior conditions that athletes from strong nations can enjoy: They have access to better training facilities, better medical services, receive larger funding and support and can – given the presumably higher domestic demand for ski jumping events on TV – generate significantly higher earnings through endorsement contracts compared to athletes from countries where ski jumping is still a peripheral sport. This can be tested in future research using for instance TV viewing figures (which are available for some countries for a small fee), information on country-specific club memberships (indicating the domestic popularity of the sport) and, if available, information on the earnings of the top athletes.

2.6 APPENDIX A

The following example illustrates that a concentration ratio – in this case the Gini coefficient – is an inappropriate measure to determine a national team's level of competitiveness in professional ski jumping. For this purpose we look at individual World Cup points accumulated by all athletes from Norway and the Czech Republic in the season 2010/11. Norway, the birthplace of ski jumping and traditionally one of the most successful nations in professional ski jumping, came second in the annual nations ranking. In total, 11 different athletes managed to win World Cup points in that particular season. The Czech Republic disposes of a considerably smaller pool of talented athletes and usually finishes well below the podium places in the nations ranking. In the season 2010/11, 6 athletes were able to win World Cup points, securing their country a top 10 spot in the nations ranking.

Table 2-3: Gini coefficients of concentration of points in the FIS ski jumping World Cup

Norway

i	x_i	h_i	(2i - n - 1) / n	h_i* ((2i - n -1) / n)
1	2	0.0007	-0.9091	-0.0006
2	7	0.0023	-0.7273	-0.0017
3	32	0.0106	-0.5455	-0.0058
4	56	0.0185	-0.3636	-0.0067
5	106	0.0349	-0.1818	-0.0064
6	155	0.0511	0.0000	0.0000
7	344	0.1134	0.1818	0.0206
8	364	0.1200	0.3636	0.0436
9	419	0.1381	0.5455	0.0754
10	645	0.2127	0.7273	0.1547
11	903	0.2977	0.9091	0.2707
∑	3033	1.0000	0.0000	0.5438

Czech Republic

i	x_i	h_i	(2i - n - 1) / n	h_i* ((2i - n -1) / n)
1	1	0.0013	-0.8333	-0.0011
2	53	0.0670	-0.5000	-0.0335
3	56	0.0708	-0.1667	-0.0118
4	119	0.1504	0.1667	0.0251
5	180	0.2276	0.5000	0.1138
6	382	0.4829	0.8333	0.4024
∑	791	1.0000	0.0000	0.4949

Table 2-3 lists the individual World Cup points (x_i) accumulated by all athletes (i) from Norway (n = 11) and (ii) the Czech Republic (n = 6) in the season 2010/11 in ascending order. Each athlete's relative share of points (h_i) is then obtained by the following formula:

$$h_i = x_i / \sum_{i=1}^{n} x_i. \quad (1)$$

The remaining computations are displayed in the upper line of Table 2-3 and yield the respective Gini coefficients (which can be found in the highlighted lower right corners). The Gini coefficient can take values between 0 (all World Cup points are evenly distributed among all athletes) and 1 (maximum heterogeneity, i.e. one athlete wins all the points for

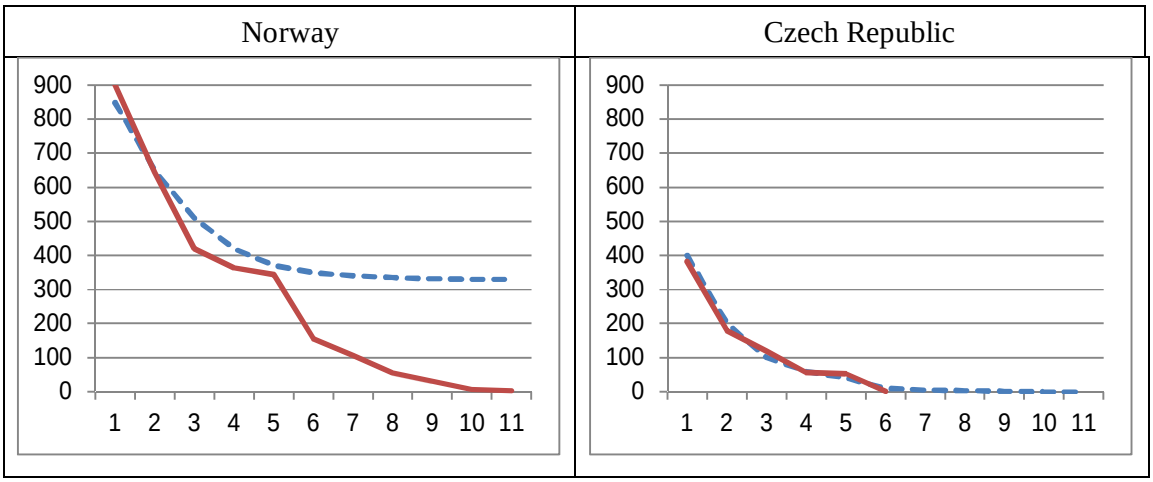
his national team while all other athletes fail to win a single point). It appears from Table 2-3 that the Czech ski jumpers were more homogeneous and thus had to face greater competitive pressure from their compatriots than the Norwegian athletes. However, these results suffer from two obvious "distortions". First, the number of athletes accumulating World Cup points varies considerably between the two nations (and the remaining nations in the sample). Second, given the limited number of World Cup slots per nation, most of the Norwegian athletes experienced some form of rotation throughout the season and hence participated in a smaller number of events. The Czech athletes, on the other hand, have been active in all (or most) of the World Cup events in that season. Due to this bias it is not possible to simply calculate the Gini coefficient with a reduced (and comparable) number of Norwegian athletes. Table 2-4 below shows an adjusted Gini coefficient for Norway, where we add up all World Cup points of the least performing (or less often nominated) individuals and allocate the sum of points to a hypothetical single sixth athlete who is assumed to have participated in all events.

Table 2-4: Adjusted Gini coefficient of a reduced number of athletes

I	x_i	h_i	(2i - n - 1) / n	h_i* ((2i - n -1) / n)
1	344	0.1134	-0.8333	-0.0945
2	358	0.1180	-0.5000	-0.0590
3	364	0.1200	-0.1667	-0.0200
4	419	0.1381	0.1667	0.0230
5	645	0.2127	0.5000	0.1063
6	903	0.2977	0.8333	0.2481
∑	3033	1.0000	0.0000	0.2039

The actual distributions of individual World Cup points as well as the estimates of the distributions of talent in the two nations in our example are presented in Figure 2-5 below. The hypothetical talent pool is represented by the dotted line and can be seen as a proxy for individual World Cup points each athlete could have achieved had he been active in all World Cup competitions in that season. We assume here that the Norwegian ski jumping federation has access to a number of equally talented athletes from the right tail of the talent distribution whereas none of the less talented Czech athletes has the potential to win World Cup points.

Figure 2-5: Distribution of talent in selected countries in the World Cup season 2010/11



Note: Number of athletes on the x-axis, cumulative individual World Cup points on the y-axis.

Although the distribution of talent in both countries is based on rough estimates – which are far from statistical accuracy – Figure 2-5 illustrates that a largely homogeneous pool of athletes might induce regular changes in the compositions of the team which, in turn, can result in misleading concentration ratios. Therefore, it seems sensible to use each national team's relative performance (as expressed by the share of World Cup points per season) as a proxy for competitive pressure in a particular squad.

3 THE PERFORMANCE OF GERMAN FOOTBALL REFEREES: ARE THERE SANCTIONS FOR POOR OFFICIATING?

3.1 INTRODUCTION

In their seminal paper on the economic theory of tournaments, Lazear and Rosen (1981) posit that (promotion) tournaments serve as a means of identifying (and rewarding) the most productive workers in a firm. In the sports context, too, rank-order tournaments – in combination with appropriately designed reward structures – are widely used to align the interests of principals (tournament organizers) and agents (athletes). However, a large body of literature has documented the existence of inefficiencies in particular labor markets that are evoked by subjective and biased performance assessments, including favoritism. Prendergast (1999) provides a thorough *theoretical* analysis of this phenomenon. To *empirically* test the theory of favoritism in principal-agent relationships, the labor market for professional football players is a particularly suitable setting. First, as pointed out by Kahn (2000: 75), "*[t]here is no research setting other than sports where we know the name, face, and life history of every production worker and supervisor in the industry. Total compensation packages and performance statistics for each individual are widely available, and we have a complete data set of worker-employer matches over the career of each production worker and supervisor in the industry.*" Second, (major) professional team sports in general and European football in particular experienced tremendous economic growth in the recent past, representing a labor market of a considerable size[22] that merits closer inspection. Third, professional football entails the advantage of being a widely televised sport, where independent experts can easily evaluate player- and – in this context more importantly – referee performance ex post. These performance data are open to the public and can be accessed online.[23] Thus, from a researcher's perspective, it is possible to monitor the monitor (i.e., the umpire) at almost no cost.

Given these merits, it is not surprising that an already large and still growing strand of literature exploits football data to unveil and explain in-group favoritism. Following Akerlof's (1980) theoretical framework of the interaction between social customs and agents' choices, the commonly observed home bias, for instance, is predominantly attributed to

[22] According to the latest financial report provided by Deutsche Fußball-Liga (2014), the 18 first division teams in the German Bundesliga generated aggregate revenues of almost 2.2 billion € in the season 2012/13.
[23] The data used in this study come from various special editions as well as the website of "Kicker" (www.kicker.de), a renowned football magazine.

referees' unintended adherence to social norms (see e.g. Buraimo et al. 2012b). Such behavior may certainly distort competition and lead to dramatic sporting consequences as well as economic losses for the disadvantaged teams. Yet, little research has been done so far concerning the consequences of poor – be it biased or simply inadequate – officiating on the side of the referees.[24] Put differently, the question of whether poor officiating leads to sanctions in the form of fewer nominations for the referees, longer waiting times and/or long-term impediments to their career has, to the best of our knowledge, not been empirically answered to date.[25] This paper aims at closing this gap in the literature.

In the next section, we present a selective review of the literature on favoritism in organizations, with a particular focus on empirical studies using football data. We then proceed in section 3.3 with a detailed description of the institutional framework, including the nomination procedure and ranking system of referees in German football, which serves as a basis for the subsequent formulation of hypotheses. In section 3.4, we briefly explain how referees are incentivized. The data, methodology and some descriptive statistics are presented in section 3.5 while the empirical findings are discussed in section 3.6. We then conclude with a plea for further research in this field.

3.2 Favoritism in Organizations – Theoretical Considerations and Empirical Evidence from Professional Sports

A growing empirical literature has unmasked what passionate football fans always knew: Referees in professional sports are (unfortunately) not impartial. While some referees are even found to be corrupt (see Hill 2009; Distaso et al. 2012 for the "Calciopoli" scandal in Italian football; or Buraimo et al. 2012a for an analysis of consumer demand following this scandal), the majority of the studies finds that referees somehow favor the home team. Irrespective of the alleged referee bias, home advantage is a common phenomenon in sports.[26]

[24] Using a dataset of more than 3,000 Bundesliga matches from the seasons 1995/96-2004/05 Frick et al. (2009a, b) find that referees with an average performance in a particular match have to wait one week longer until their next nomination than referees with a particularly good performance. Since the dataset is restricted to first division matches, temporary demotions (i.e., nominations to lower division matches) had to be omitted. In a similar vein, Frick (2012) focuses explicitly on the career duration of football referees active in the German Bundesliga. The results suggest that individual survival is mainly determined by the number of nominations, which may serve as a proxy for individual performance.

[25] A notable exception is presented by Boeri and Severgnini (2011) who look at the assignments of referees to important matches in the Italian football league dependent on their performance in the previous match. The scope of the study, however, is consciously restricted to a short time period in which some Italian referees were heavily engaged in match rigging.

[26] Courneya and Carron (1992: 14) define home advantage as "the consistent finding that home teams in sport competitions win over 50 % of games played under a balanced home and away schedule".

The reasons for this home advantage are manifold. Nevill and Holder (1999) provide a literature review of earlier studies that have identified four factors thought to mainly contribute to home bias, namely learning and rule factors as well as travel and crowd factors. The accumulated evidence suggests that learning (becoming increasingly familiar with the conditions when playing at home) and rule factors (e.g. batting last in baseball) play only a minor role in causing home advantage. Travel factors were found to significantly contribute to home advantage when teams or individual athletes had to cross a number of time zones. However, in European football travel distances are relatively small in most national leagues and even in (the majority of matches played in) pan-European cup competitions. Thus, travel factors are considered a minor cause of home advantage. The most prominent causes of home advantage appear to stem from crowd effects. Nevill and Holder (1999) propose two possible mechanisms to explain the observed effects: The crowd is able to either (i) raise the performance of the home team (i.e., social support), or (ii) influence the umpires to subconsciously favor the home team (i.e., social pressure). Although disentangling these effects often proves to be a cumbersome task for the econometrician, the vast majority of the literature suggests that the latter is the driving force behind the home effect. This is in line with the assumptions predicted by the theory of conformity[27] (Bernheim 1994; Becker and Murphy 2000).

The following literature review summarizes and classifies the more recent empirical evidence on referee home bias in professional sports. Notwithstanding the fact that there are other manifestations of bias (see, inter alia, Plessner and Betsch 2001 for sequential effects in awarding penalties in football; Jones et al. 2002 for reputation bias among association football referees; Price and Wolfers 2010 for racial bias by referees in the NBA; or, with similar findings in the MLB, Parsons et al. 2011), the following studies explicitly focus on differential treatment caused by some form of home bias. The first (and main) part of the literature review captures studies using data from professional football. In the second part the scope is broadened to empirical findings from selected individual sports.

[27] The theory of conformity posits that individuals who care about status are willing to conform to a certain standard of behavior, anticipating that even small deviations from the social norm could seriously impair their reputation.

3.2.1 HOME TEAM FAVORITISM IN ASSOCIATION FOOTBALL

In association football, referees represent the major monitoring authority responsible for the enforcement of rules under the Laws of the Game.[28] Typical tasks of a referee include the determination of extra time, issuing sanctions in the form of yellow and red cards as well as allowing goals and penalties. Unlike in major US sports, where umpire decisions can be overruled by instant video proof, referee decisions in football are usually incontestable and can thus adversely and irrevocably affect the course of the game. As pointed out by Buraimo et al. (2010), "*[c]ritical refereeing decisions can be pivotal for a team's prospects of winning championships, qualifying for lucrative European competition or avoiding relegation*." The empirical literature examining home bias in football refereeing can be divided according to the above-mentioned typical tasks of a referee.

3.2.1.1 DETERMINATION OF EXTRA TIME

Referees are in charge of tracking the amount of time that is 'lost' due to injuries, substitutions, time wasting or other interruptions of the game. Although they receive official recommendations from the assistant referee, the so-called fourth official, the "men in black" take sole responsibility for the amount of time that is added at the end of each half of a game. A number of studies examining the amount of extra time find that referees add significantly more "injury time" when home teams are trailing as compared to when home teams are leading in score. Seminal evidence in this strand in the favoritism literature stems from Garicano et al. (2005) who use data from Spanish football. Their results suggest that home team favoritism in the form of extra time exists even after controlling for the number and the length of interruptions in the game. These results are confirmed by Lucey and Power (2004) for the Italian Serie A and US Major League soccer. Scoppa (2008) reports similar results for Italian football. Sutter and Kocher (2004) and Dohmen (2008) find that in the German Bundesliga, too, referees favor trailing home teams in terms of added time.

The following studies are unable to identify persistent bias but rather demonstrate how institutional changes can foster referee performance and impartiality. Rickman and Witt (2008) first confirm the above results for English football but then show that home team favoritism in the English Premier League disappears after the introduction of professional employment contracts. Referees who were paid on a match basis at the beginning of the

[28] The Laws of the Game gather the codified football rules that are defined by the Fédération Internationale de Football Association (FIFA) and approved by the International Football Association Board.

sample period were found to significantly favor the home team by adding more extra time when the home team was one goal behind. This effect disappeared after the modification of the payout regime from match fees to annual salaries. Looking at exactly this transition phase between the two payout regimes in English football, Bryson et al. (2011) find that referees who move onto salary contracts increase their performance relative to those who are still paid a match fee. These results are robust to referee fixed effects, thus ruling out potential ex ante ability and/ or sorting effects. Nevill et al. (2013) report a systematic decline in refereeing bias in top-tier English and Scottish football leagues which they attribute to improved training of referees post World War II that has facilitated greater resilience to crowd influence and, thus, more objective decision-making. Another example for the reduction in referee bias is presented by Rocha et al. (2013). Applying ordinary least squares and probability regressions to a dataset from the first division of the Brazilian Football Championship, the authors lend further support to the notion of favoritism among referees. However, when controlling for the importance of a match (as measured by the quality of both teams and whether a match was broadcasted or not) and the quality of the referee, the results are somewhat different: With increasing importance of a match the amount of extra time tends to be closer to an objectively justified amount. In other words, monitoring reduces referee bias. In a more general sense, these results suggest that career concerns play a major role in agents' choices in the presence of comprehensive monitoring.

3.2.1.2 SANCTIONS

In order to avoid illegal behavior by players, referees can issue cautions in the form of yellow and red cards, the latter implying the expulsion of a player from the field. The use of disciplinary sanctions (and in particular a red card) can weaken a team considerably[29] and thus represents another potential source of referee bias. In fact, a number of empirical researchers have added weight to the hypothesis that umpires favor home teams in terms of issuing sanctions in the form of cards. Analyzing referee patterns in the award of disciplinary sanctions in the English Premier League over a period of 7 years, Dawson et al. (2007) show that away teams receive more cards than home teams. Since this effect persists even after controlling for the quality of both teams, the evidence suggests that referees (presumably unintentionally) favor the home team. In line with these results, Buraimo et al. (2010), Dawson and Dobson (2010) and Reilly and Witt (2013) find that referees in Eng-

[29] See Ridder et al. (1994) for a quantification of the effect of a player's expulsion on the outcome of a football match; i.e., the final score.

lish and German football as well as in pan-European cup competitions issue fewer cards to the home team and that this effect cannot solely be attributed to away teams playing more aggressively but rather points at home team favoritism on behalf of the referees. Whether the observed home bias in awarding cards is constant across all referees or subject to some referees being more favorably inclined towards the home team than their colleagues (see Boyko et al. (2007) and Johnston (2008) for contradictory evidence), remains a contested issue and leaves room for future research.

3.2.1.3 GOALS AND PENALTIES

A few studies offer support for referee home bias in terms of penalty decisions. Nevill et al. (1996) find that home teams in the highest divisions of the English and Scottish football leagues are given significantly more penalties than away teams. Yet, Sutter and Kocher (2004) argue that concentrating on awarded penalties only produces spurious evidence for favorable treatment. Assuming that home teams typically play more offensively and are thus more likely to enter the penalty area, the "surplus" in penalties might simply reflect different styles of play. In order to circumvent any confounding factors of this nature, the authors look at the ratio of legitimate and rewarded penalties to legitimate but refused penalties in the German Bundesliga in the season 2000/2001. Their analyses reveal that in 81 percent of the cases the home team is awarded a legitimate penalty whereas visiting teams receive a legitimate penalty in only 51 percent of the cases, pointing at a considerable and statistically significant (χ^2=9.7; p<0.01) home bias in terms of awarding penalties. Investigating the impartiality of referees in 3,519 matches played in 12 consecutive seasons in the German Bundesliga, Dohmen (2008) corroborates these findings and additionally identifies refereeing home bias with regard to awarding disputable goals.

3.2.1.4 CROWD EFFECTS

As postulated at the beginning of this section, the majority of studies ascribe the observed referee home bias to social forces. That is, referees – who intend to act as neutral rule-enforcers – unintentionally conform to the crowd's preferences in order to avoid crowd displeasure. In an attempt to quantify this crowd effect, a number of studies have included information on the size and composition of the crowd as well as the proximity of football fans to the pitch and the related crowd noise (see, inter alia, Garicano et al. 2005; Boyko et al. 2007; Dohmen 2008; Buraimo et al. 2012; Goumas 2012). Although empirical evidence on the magnitude of specific crowd effects is mixed, the general tenor suggests that all of the above-mentioned crowd characteristics somehow have an impact on referee behavior.

The larger the crowd and the greater the share of home team supporters in the stadium the more inclined the referee is towards the home team. Interestingly, home bias seems to be lower in stadia where a running track separates the fans from the pitch. Skeptics would surely argue that this reflects team-specific effects. However, Buraimo et al. (2010) make use of a quasi-experiment in which they show that teams in the German Bundesliga that changed their ground structure with removal of a track – Schalke and Bayern Munich moved to new stadia in 2001 and 2005, respectively, while Hannover modernized its old stadium in 2003 – benefitted from a significant increase in home team favoritism on the part of the referees.

Additional experimental evidence in this field is presented by Nevill et al. (2002) who asked 40 qualified English referees to assess videotaped tackles that were recorded during an English Premier League match from the 1998/1999 season. 22 of the referees evaluated the video scenes with the crowd noise, while the remaining 18 referees watched the videos in silence. Applying binary logistic regression analyses to their dataset, Nevill et al. (2002) find that those referees assessing tackles in the noise condition awarded significantly fewer fouls (15.5 percent) against the home team than the referees watching the scenes in silence. Moreover, the statistically significant difference of 15.5 percent almost exactly reflects the reported percentage difference/advantage for home wins in football. In a similar vein, Pettersson-Lidbom and Priks (2010) investigate referee decisions in the absence of crowd noise by looking at games in the Italian Serie A that had to be played in empty stadia following a series of uproar and riots during games. Their results suggest that referees are more likely to penalize home team players in the absence of spectators whereas home team favoritism prevails in "normal" matches with a crowd involved. However, Buraimo et al. (2010) express criticism towards the reliability of the results for a number of reasons including, among others, the relatively small subsample (only 24 matches were played in empty stadia), a lack of control for within-game influences as well as a potential endogeneity problem regarding teams with a reputation of crowd trouble. Summarizing, these findings suggest that crowd noise is a salient feature in explaining the causes for refereeing home bias. Yet, when analyzing the impact of social pressure on referee behavior in football, player- and team-specific effects need to be addressed econometrically.

3.2.2 HOME BIAS IN SELECTED INDIVIDUAL SPORTS

Apart from the extensive literature addressing the issue of home team favoritism in professional team sports, a considerable strand in the literature identifies a similar phenomenon

in individual sports: Athletes competing on home soil fare significantly better than their opponents. The empirical evidence presented in the following suggests that "social support", i.e., the crowd's positive influence on the "local hero" (see section 3.2), cannot explain the home advantage alone, but rather points at some form of home bias on behalf of judges/referees.

Whereas home bias seems to be less of a problem in sports that involve relatively objective performance assessments – Bray and Carron (1993), Koning (2005) and Nevill et al. (1997), for instance, find little evidence of home advantage in alpine skiing, speed skating, professional tennis and golf – empirical evidence of home bias in more subjectively judged individual sports is abundant. Analyzing data from the Winter Olympics between 1908 and 1998, Balmer et al. (2001) find that home advantage was significantly larger in events that were subjectively assessed by judges (e.g. figure skating and freestyle skiing) than in more objectively evaluated competitions (e.g. Nordic skiing), which is indicative of judges scoring "home" contestants disproportionately higher than "away" contestants. In the same spirit, Balmer et al. (2003) observe home bias in the Summer Olympics held between 1896 and 1996 for subjectively assessed sports, i.e., boxing and gymnastics, while no home advantage is found for two rather objectively judged groups of sports, namely athletics and weightlifting. Balmer et al. (2005) confirm the prevalence of officiating bias for European championship boxing by showing that the expected probability of a "home win" is significantly larger when bouts are decided by points as opposed to fights ending in a (technical) knockout. Similar to the above-mentioned experimental study by Nevill et al. (2002), who asked football referees to assess videotaped tackles in two different conditions (i.e., with and without crowd noise), Myers et al. (2012) apply the same methodology to Muay Thai (Thai boxing) judges. Their results are in line with Nevill et al. (2002), suggesting that Muay Thai judges, too, apparently succumb to crowd noise and (inadvertently) favor the "home" boxer.

Although experimental evidence of this nature is characterized by a high level of internal validity (since only the crowd noise condition is allowed to vary), Myers and Balmer (2012) argue that experiments lack external validity, as particularly referee decisions in the laboratory bear little resemblance to "real-life" decisions in live sports settings. In an attempt to achieve external validity for the above results, Myers and Balmer (2012) conduct a controlled experiment to determine the impact of crowd noise on officials in a live sporting event including "home" and "away" contestants. In an international Thai boxing tour-

nament, judges were randomly assigned to a ”(crowd) noise” or “no crowd noise” group, with the latter receiving noise cancellation headphones. The results corroborate the prevailing evidence on officiating bias and contribute to the existing literature by providing first experimental evidence on the impact of crowd noise on officials in a live tournament setting.

It should be noted, however, that crowd noise cannot always explain partiality on the part of judges or referees. Some scholars provide evidence for systematic referee bias irrespective of any crowd effects. Zitzewitz (2006), for example, examines the (voting) behavior of ski jumping and figure skating judges who were active in events before, during and after the 2002 Winter Olympics. One of the key results of the study suggests that figure skating judges and – to a lesser extent – ski jumping judges exhibit nationalistic bias. That is, judges deliberately award higher scores to athletes from their own country than other judges do. This finding stands in stark contrast to the evidence presented so far in as much as the observed favoritism by agents is due to (deliberate) strategic decision-making rather than the result of (subconscious) crowd influence.[30]

3.2.3 SUMMARY OF THE LITERATURE AND AVENUES FOR FURTHER RESEARCH

Summarizing, the results of the above studies lend strong support to the notion that referee home bias is a widespread phenomenon (and presumably the main influencing factor of home advantage) in sports. Moreover, officiating bias appears to be most pronounced in sports which involve subjective decision-making. However, what has as yet not been addressed in the literature is whether poor officiating (which, given the presumably more objective nature of independent performance evaluations of referees, is a necessary consequence of favoritism) entails any career-impeding sanctions on the side of the officials. The aim of this study is to close this gap in the literature. Before the data, methodology and empirical findings are presented, the following section describes how referees in the German Bundesliga are selected, evaluated and incentivized.

[30] Corroborating the general susceptibility to partiality of figure skating judges, Findlay and Ste-Marie (2004) show that judges award higher scores to figure skaters they know compared to those skaters they do not know. All else being equal, this finding points towards a significant reputation bias of judges.

3.3 INSTITUTIONAL FRAMEWORK, NOMINATION PROCEDURE AND HYPOTHESES

In order to ensure a high level of quality and consistency among referees, the German Football Association (*Deutscher Fußball-Bund*, henceforth DFB) has implemented a ranking system that is specifically designed to select and incentivize the most talented referees. Analogous to the classification of football teams into divisions, referees, too, are divided into divisions, with the three top divisions currently consisting of 19-22 referees, each.[31] With a few exceptions, only the highest ranked referees are allowed to officiate in first division games. This does not categorically rule out that first division referees are also appointed to second and third division matches. The ranking system does however exclude the possibility of a third division referee officiating in a first division game. In other words, referees are almost exclusively authorized to umpire games in divisions that are equal to or lower than their own qualification level (i.e., the division that has been officially appointed to them by the DFB). Moreover, top referees can achieve the ultimate status of becoming an official FIFA referee which legitimizes them to officiate in European Cup competitions as well as in international caps. Thus, there is a huge intrinsic (and, as will be elucidated in the next section, extrinsic) incentive for referees to climb up the career ladder to eventually qualify for the pool of first division referees. The referee classification into divisions and the small elite pool of FIFA referees (approximately 10 at a time) are updated annually, prior to the start of the season. The DFB has initially imposed a mandatory retirement for referees at an age of 50. Perhaps in response to the increasing dynamics of professional football and/ or a growing pool of promising referee aspirants the age threshold for retirement has been gradually reduced to the current level of 47, allowing more referees (typically between one and five per season) to get promoted to a higher division. At the same time, only few referees are demoted, suggesting that the DFB prefers an "up-or-out" principle (see e.g. Waldman 1990) to career progression.

Nevertheless, the question remains whether those referees who get promoted are in fact the top referees in their respective division or if other factors such as seniority, network effects

[31] In the early years of the German Bundesliga, which celebrated its inaugural season in the year 1963, at times more than 50 different referees per season were given the opportunity to umpire at least one first division match. The pool of referees was then constantly reduced to 37 in the season 1980/81, roughly 25 in the mid-1990s and approximately 20 officials per division in the post-millennial seasons (data can be accessed via http://www.weltfussball.de/schiedsrichter/bundesliga).

or pure luck determine success and failure.[32] Given the widely acknowledged positive incentive effects of rank-order tournaments, and assuming the DFB's tournament-like classification of referees to be effective (in identifying the most productive agents), we derive the following hypothesis:

> $H_{promotion}$: *The better a referee's average performance and the more consistent this performance throughout the season, the higher the likelihood of a permanent promotion to a higher division.*

Consequently, a referee's qualification level should serve as an indicator of individual ability and, thus, referee performance is assumed to improve with increasing level of qualification. That is, FIFA referees should, on average, perform better than "ordinary" first division referees, who, on the other hand, are supposed to deliver better performances than second or third division officials. Accordingly, the second hypothesis states:

> $H_{ability}$: *The higher a referee's qualification level (i.e., the higher the division he is appointed to prior to the season), the better the average performance.*

We deliberately exclude a consistency measure in our second hypothesis as we look at aggregate performance data where outliers in performance variation are likely to be smoothed out. While the first two hypotheses largely reflect a referee's long-term performance and its expected impact on career progression, it remains to be tested whether a referee's short-term performance has any (immediate) impact on individual career prospects. More specifically, we want to test if a referee's performance in one particular game somehow affects his subsequent nomination. It is possible that, for example, poorly performing referees have to wait significantly longer until their next nomination than their (better performing) colleagues. Another potential sanction on behalf of the DFB would be to appoint well qualified but poorly performing referees to matches played in a lower division. In this context, it should be mentioned that referees are nominated for a match on rather short notice. Typically, officials are informed about their nomination less than a week prior to the match day. The information on the exact fixture that a referee is assigned to is internally disclosed about 48 hours before the match, while the public is informed only 24 hours prior to kick-

[32] The performance of a referee is typically assessed by a delegate of the DFB who monitors the umpire directly in the stadium. The corresponding evaluation sheet, however, is handed over to the authorities only a few days after the game. In the meantime, evaluators can thus rely on other "experts" analyzing critical scenes that are televised in the aftermath of a game.

off. The DFB implemented these precautionary measures in order to reduce the likelihood of match rigging (see DFB 2013). Therefore, it is possible that particularly poor performances have an immediate effect on the subsequent nomination and, thus, the following two hypotheses can be formulated:

$H_{waiting\ time}$: *Referees who perform poorly in one particular match have to wait significantly longer than their (better performing) colleagues until they are nominated for the next match.*

$H_{demotion}$: *Referees who perform poorly in one particular match have a significantly higher likelihood of being temporarily demoted to a lower division with their subsequent assignment.*

In order to gain a better understanding of the financial implications associated with the different nomination outcomes, the following section intends to illustrate how top-class German football referees are incentivized.

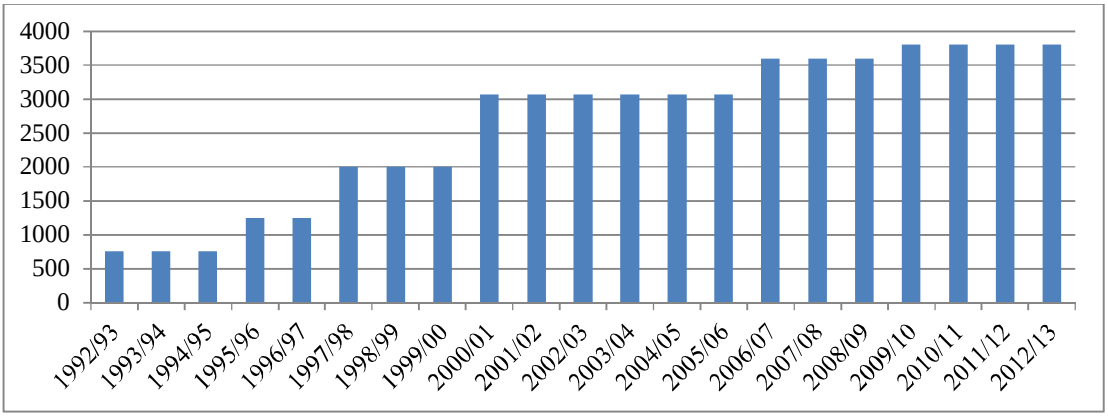
3.4 Remuneration of Referees in German Association Football

Notwithstanding the assumption that the labor market for football referees is likely to attract a far more passionate and intrinsically motivated workforce than, for example, the steel industry, the DFB relies on financial incentives to ensure high levels of effort among its top referees. Although until recently German referees were still considered amateurs who, unlike their colleagues in, e.g., England and Spain did not receive a fixed annual salary (see Frick et al. 2009a for a detailed breakdown of fix and variable pay of referees in the "Top 5" football leagues in Europe), they invest a lot of time and effort that calls for some form of financial compensation.[33] After all, the referees who are active in the three top divisions in Germany represent the very right tail of the talent distribution of a pool that currently consists of roughly 75,000 referees in Germany. On the other hand, more than 99 percent of German referees officiate in minor league matches on a mostly voluntary basis (usually, the regional federations pay travel costs and/ or a small allowance). Initially, even the top referees received only a small allowance of approximately 100 € on top of their travel expenses. Following a dramatic increase in league revenues in the early 1990s (see

[33] In addition to their primary task of officiating matches which is often associated with substantial time and travel expenses, referees in the highest divisions are required to take part in training courses and performance evaluations on a regular basis. Since many of the referees already pursue a "regular" profession, the time-consuming refereeing activity most likely impedes career progress and thus leads to foregone income.

Frick and Prinz 2006), the DFB introduced a match fee of 750 € that was paid to first division referees with the beginning of the season 1992/93. This fee was then gradually increased to the current level of 3,800 € (see Figure 3-1).

Figure 3-1: Match fees of football referees in the German Bundesliga, 1992/93-2012/13



Note: Match fees (see y-axis) prior to the introduction of the euro in Germany, on 1 January 2002, are converted and approximated to euro amounts. Match fees are not adjusted for inflation. Own illustration based on data provided by the DFB.

The two assistant referees (also named linesmen) receive half of what the referee is paid, while the recently introduced "fourth official", who monitors the activities off the pitch, receives a quarter of what the referee is paid. Referees and their assistants in the second and third division receive approximately 50 and 20 percent, respectively, of what the referee teams in the first division are paid (see Table 3-1).

Table 3-1: Current match fees of referees and assistants in the top 3 German football divisions (in €)

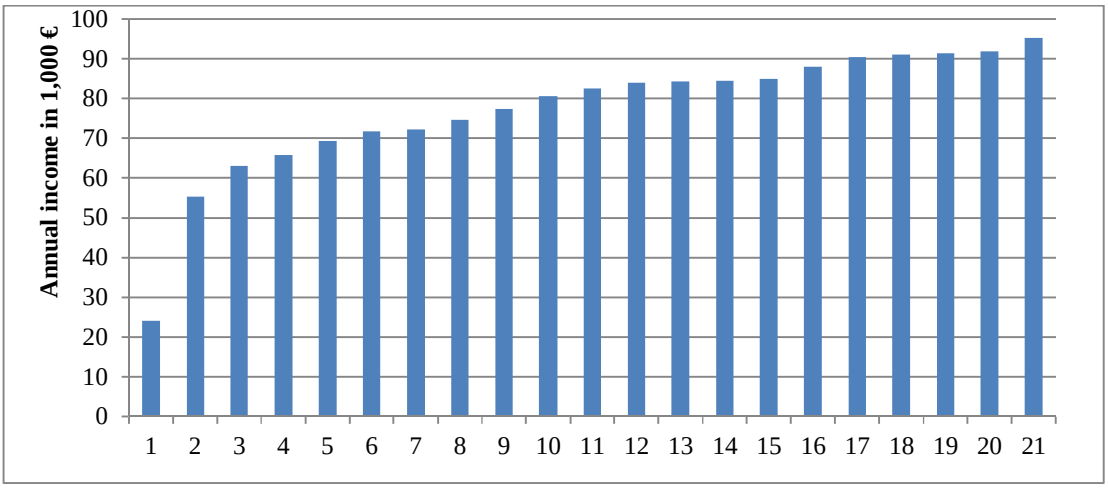
Division	Referee	Linesman	4th Official
Bundesliga	3,800	1,900	950
2nd Division	2,000	1,000	500
3rd Division	750	375	187.50

Source: Own illustration based on data from the DFB.

In addition to the match fees, the DFB covers all travel expenses, such as transportation, food and accommodation costs, amounting to approximately 200 € for a single referee. Moreover, referees who are assigned to a first division match receive an additional premi-

um of 500 € from the league sponsor. Since referees were exclusively paid on a match basis until the season 2011/12, we assume that rational individuals sought to maximize the number of nominations (and thus their income). Figure 3-2 below illustrates how much the 21 Bundesliga referees who were active in the season 2011/12 (and who are displayed on the x-axis) earned from match fees (y-axis).

Figure 3-2: Cumulative individual income of Bundesliga referees in the season 2011/12



Source: Own illustration and own calculations based on data from the DFB.

It appears that most of the first division referees accumulated between 70,000 and 90,000 € from match fees alone.[34] As already discussed above, top referees can generate additional income by officiating in European Cup competitions as well as in international caps, where match fees sometimes amount to 6,000 € and more. Hence, FIFA referees usually earn more than 100,000 € per year. Assuming that many of the top referees had to reduce (or completely refrain from) investments in vocational education and training, the opportunity costs of quitting the career are quite high. Thus, referees have a huge incentive to (i) climb up the career ladder and eventually become a Bundesliga or even FIFA referee, and (ii)

[34] Usually, referees can expect a sufficiently large number of nominations per season that guarantees them a satisfying minimum wage (currently about 50,000 € for first division referees) even in the absence of a fixed salary. The negative outlier above can be explained by a sad and unique case where an experienced first division referee, Babak Rafati, attempted suicide only hours before his expected opening whistle in the Bundesliga match between Cologne and Mainz on 19 November 2011. Symbolically enough, Rafati's debut as a Bundesliga referee in the year 2005 saw exactly the same two teams facing each other. Rafati, who was found in his hotel room by his assistant referees who provided first aid, later explained that he suffered from serious depression and that he could no longer cope with the enormous pressure to perform. Not having returned to the pitch ever since, he soon after officially quit his career and meanwhile gives seminars on performance pressure, mobbing and burn-out.

maximize the number and quality of nominations. Despite the relatively high income even top referees were, until recently, still considered amateurs in an otherwise highly professionalized environment. With the beginning of the season 2012/13, however, the DFB introduced a fixed salary that has since been paid to all first and second division referees. This base salary is paid irrespective of the number of nominations but depends on the individual qualification level and experience: FIFA referees with a Bundesliga experience of five years or more received an annual base salary of 40,000 € whereas equally experienced Bundesliga referees received a fixed sum of 30,000 € in the season 2012/13. Bundesliga referees with less than five years of experience and second division referees were paid 20,000 and 15,000 €, respectively, while linesmen received between 2,500 and 15,000 €, depending on their status and experience (see Table 3-2).

Table 3-2: Annual base salary of Bundesliga referees in the season 2012/13 (in €)

Status/Division/Experience	Referee	Linesman
FIFA, experience of 5 years or more	40,000	15,000
Bundesliga, experience of 5 years or more	30,000	15,000
Bundesliga, experience of 5 years or less	20,000	10,000
2nd Division	15,000	2,500

Source: Own illustration based on data from the DFB.

Only one year after the introduction, base salaries for first and second division referees were raised substantially while match fees remained unchanged. In the season 2013/14, FIFA referees received a fixed sum of 60,000 € while experienced (inexperienced) Bundesliga and second division referees earned 50,000 (40,000) and 25,000 €, respectively. In response to the increasing requirements for referees, Wolfgang Niersbach, current DFB president, recently announced a further pay raise that is intended to create optimal conditions for top referees in the long term. From the season 2016/17, base salaries will increase to 75,000 € for FIFA referees. Bundesliga referees can expect an annual fixed fee of 55,000 - 65,000 €, depending on their experience, while second division referees will then

receive a guaranteed annual base pay of 35,000 €. With respect to the earnings structure it is striking that the absolute pay gap between FIFA and Bundesliga referees has remained unchanged over the years while it has increased between the two latter and second division referees. The relative pay gap, on the other hand, has decreased dramatically over time.

Whether the ongoing adjustment of (fixed) salaries leads to reduced effort levels of referees – as predicted by tournament theory – remains to be tested and opens avenues for future research. In this paper, however, we explicitly focus on the time period before the introduction of fixed salaries for top division referees. In the following section, we present our data, explain the econometric models used in the analyses and display some descriptive results.

3.5 DATA, METHODOLOGY AND DESCRIPTIVE STATISTICS

Our empirical analyses are based on two distinct datasets. The first covers all matches played in the first three divisions in German association football in the four seasons between 2008/09 and 2011/12 (n = 3,968 matches) and includes a number of match-specific variables – such as attendance and a measure of team heterogeneity reflecting match uncertainty – as well as a whole set of referee-specific variables (n = 89 referees). The second dataset includes season-level variables of all 89 referees who were active in at least one of the four seasons and comprises information on each referee's qualification level (i.e., the division he is appointed to prior to the season), average referee performance and performance consistency throughout the season as well as career-related outcomes such as promotion, demotion or career end (n = 251 referee-season observations). The data were obtained from various special editions (and the online source) of "Kicker", a renowned football magazine, while averaged betting odds from the website www.betexplorer.com serve as a proxy to compute the teams' implicit winning probabilities which are then translated into a heterogeneity index (see Appendix B).

We deliberately restrict our datasets to the period of four (consecutive) seasons for two reasons: First, season 2008/09 marks the inaugural season of the fundamentally modified third division that hitherto consisted of several regional divisions (a three/four-track league until 1999/2000 and a two-track league until 2007/08) which were then merged into a nationwide single division.[35] Top referees, who are usually appointed to first division match-

[35] The restructuring of the third division was enacted by the DFB in order to foster the already increasing professionalization of higher division football in Germany. Since the modification of the league structure,

es, also officiate in second and third division fixtures on a regular basis. The same applies for designated second division referees who are often assigned to third division matches, too. Hence, it is possible to observe and compare the performances of referees with different qualification levels. Moreover, we are able to determine a referee's waiting time between two assignments in *any* of the top 3 divisions. Finally, the data allow us to observe temporary promotions and demotions (i.e., movements between divisions) of referees. We thus contribute to the literature and extend the already mentioned studies by Frick et al. (2009a, b) who look at first division matches only.[36] Second, season 2011/12 represents the last season where referees were paid on a match basis only. With the introduction of annual fixed salaries in the subsequent season, the incentives for referees are likely to have changed dramatically. In order to obtain a relatively balanced panel with a consistent league structure and almost identical incentive schemes across the years – as already shown in Figure 3-1, match fees in season 2008/09 were slightly lower than in the following three seasons – the two datasets almost necessarily need to be restricted to the current observation period.

We proceed as follows: In section 3.5.1, we present the match-level data of our first dataset and define and explain the variables used in the empirical analyses. Subsequently, we provide some descriptive evidence and elaborate on the econometric models applied. The season-level data and the corresponding estimation techniques which are used to test for long-term performance-effects on career progress are described in section 3.5.2.

3.5.1 MATCH-LEVEL DATA

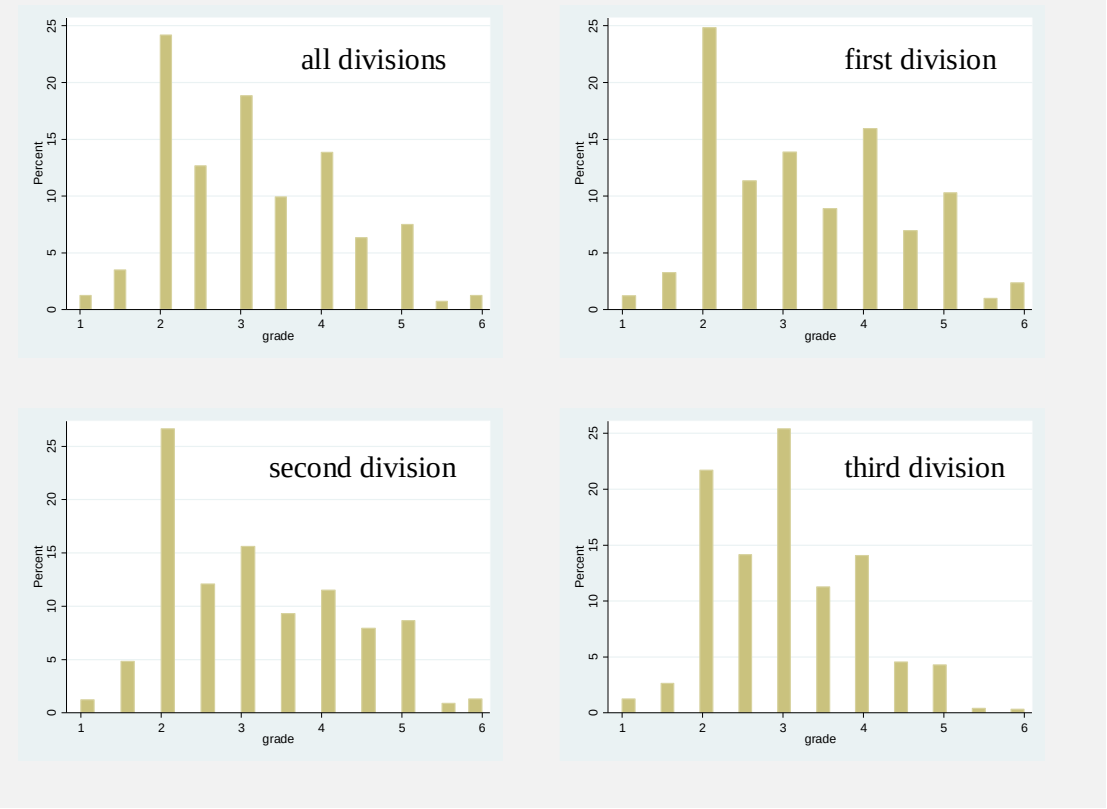
Recall that our first dataset includes all 3,968 matches that were played in the first three German football divisions in the period 2008/09 to 2011/12. Apart from abundant match-specific information, we have detailed information on the referee of each fixture. These include age, experience (i.e., the number of previous assignments in any of the top divisions), the current qualification level (i.e., first, second or third division referee) as well as

third division games have been largely televised, thus guaranteeing third division teams considerable earnings from TV broadcast rights which are an important source of revenues besides gate revenues and sponsorship payments. According to a recent DFB report (see DFB annual financial report 2012/13), the third German football league generates more revenues and attracts more fans to the stadia than any other team sport's first division in Germany.

[36] One of the shortcomings of earlier studies is that the status of a referee sometimes tended to be inaccurately reported: referees who appear to pause for a number of games might in fact be assigned to lower division matches. The nomination for a lower division match, in turn, could either imply a sanction on behalf of the league officials or manifest the necessity to assign an experienced and qualified umpire to an important lower league match.

a potential FIFA status. We also include in our dataset a dummy variable indicating whether a referee is temporarily promoted in the sense that he is assigned to a match in a division that is higher than his current qualification level. However, out of 89 different referees in our sample only seven were temporarily promoted (n = 37 matches). Among the covariates of interest, *grade* appears to be the most important one as it reflects the performance of a referee in a given match in form of a (German) school grade. The best possible grade is 1, which corresponds to an immaculate performance, whereas the lowest possible grade, 6, is awarded for a woefully insufficient referee performance. It appears from Figure 3-3 that the grades are non-normally distributed with the distributions varying across divisions.

Figure 3-3: Distribution of grades of all referees active in Germany's top 3 football divisions



The performance evaluations used in our analyses are conducted by independent monitors who observe the referee live at the stadium, while the corresponding grades are published by "Kicker" a few days after the match. These independent monitors are different from the official delegates of the DFB whose evaluations are usually not disclosed. Yet anecdotal evidence suggests that there is a rather high correlation between the official referee evalua-

tions that are carried out by the DFB and those conducted by independent "Kicker experts". Against the backdrop of the already mentioned delay between the actual performance assessment and the eventual publication of the score which gives evaluators enough time to consult TV images of (and other opinions on) critical scenes, it seems reasonable to assume that different evaluators come up with a similar grading.

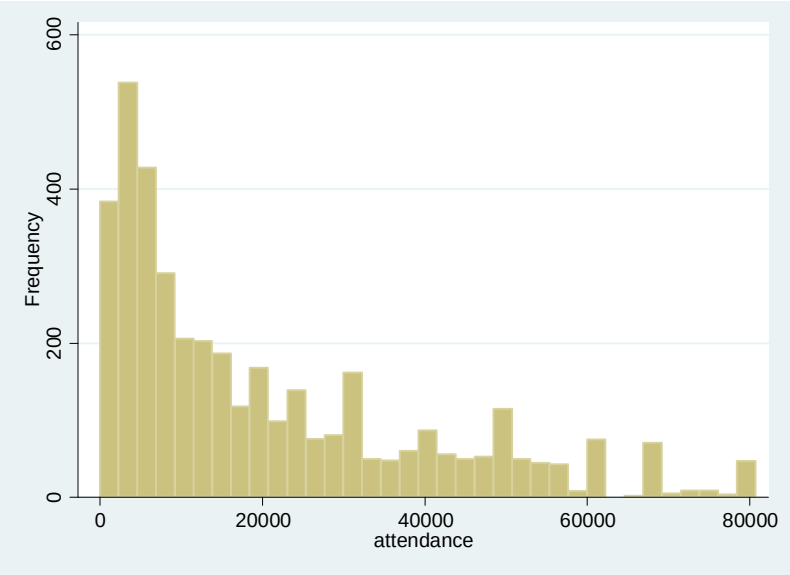
In addition to that, we include a number of match-specific variables in order to control for the division and season in which a match takes place, the actual match day as well as the attendance. Note that divisions 1 and 2 consist of 18 teams each, while division 3 is slightly larger with 20 teams, implying that proportionately more matches are played in the third division (380 per season as compared to 306 in the first or second division). *Match day* intends to capture the increasing importance of matches towards the end of the season. These matches are often decisive in terms of championship, qualification for the UEFA[37] Champions League and Europa League as well as promotion and relegation.[38] *Attendance* is another control variable that is thought to reflect the importance of a match. We use absolute attendance figures instead of capacity utilization to take account of the assumption that a near-capacity crowd of 60,000 or more fans requires a more qualified and experienced referee than a sold-out small stadium. Besides, utilization rates – at least in the first division – amount to approximately 90 percent and offer little variation. However, since *attendance* is non-normally distributed[39] across the three divisions but appears to be right-skewed (see Figure 3-4), we use *LogAttendance*.

[37] Union of European Football Associations.

[38] In order to differentiate between "decisive" and rather unimportant matches taking place on the (pen-) ultimate match day we include a dummy variable that is 1 if at least one team is still in contention for any of the above mentioned outcomes, and is 0 otherwise. Even though the share of "decisive" matches is almost 60 percent (133 out of 224), this variable has no statistically significant effect on the nomination outcome and is therefore omitted in the estimations. The results are, however, available upon request.

[39] According to the skewness and kurtosis test for normality by D'Agostino et al. (1990) and empirical adjustments to this test by Royston (1991c) we can reject the hypothesis that *attendance* is normally distributed.

Figure 3-4: Frequency distribution of attendance figures in Germany's top 3 football divisions



Moreover, we aim to quantify match uncertainty by introducing a heterogeneity measure that is based on betting odds. In line with Fama (1970), who argues that in efficient markets prices fully reflect all information available, it can be assumed that the betting market serves as a reliable source of match-relevant information. More specifically, odds issued by bookmakers appear to be the most appropriate tool in determining the *ex ante* relative strength of two opposing teams as all publicly available information is assumed to be reflected in betting odds. This has been empirically tested by Forrest et al. (2005) who employ a large sample of almost 10,000 English football matches and convincingly demonstrate that odds-setters are better forecasters than even the most sophisticated statistical models. We use averaged betting odds from the website www.betexplorer.com in order to compute the implicit winning probabilities for each team which are then translated into a heterogeneity index, *PosHET*, which takes values between 0 (complete balance) and 1 (maximum heterogeneity). A sample calculation following Deutscher et al. (2013) can be found in Appendix B. Both *LogAttendance* and *PosHet* can be interpreted as signals for match importance (and thus the necessity to assign a more or less qualified referee to this particular fixture) and are included in our estimations so as to reduce the risk of omitted variable bias. Detailed summary statistics can be found in Table 3-3.

Table 3-3: Summary statistics of the match-level analysis on referee performance and nomination outcomes

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLES						
Waiting time in games	Waiting time of a referee after an assignment, measured in games	3,879	2.41	1.60	1	24
Waiting time in days	Waiting time of a referee after an assignment, measured in days	3,879	20.10	23.54	1	484
Nomination in t_{+1}	Dummy: Equals 1 if referee is nominated for a match of the subsequent match day, 0 otherwise	3,940	0.32	-	0	1
Division in t_{+x}	Categorical variable indicating the division of a referee's next assignment in t_{+x}	3,879	2.06	-	1	3
INDEPENDENT VARIABLES Referee-Specific Information						
Grade	Referee's performance evaluation, 1 indicating an immaculate performance, while 6 represents the worst grade that is awarded for a very poor performance	3,968	3.11	1.08	1	6
Age	Referee's age	3,968	33.41	6.04	20	46
Experience	Referee's experience, expressed as the number of previous assignments in any of the top divisions	3,968	103.42	100.88	0	357
Appointed division	Referee's qualification level (i.e., the division appointed prior to the season)	3,968	1.75	-	1	3
FIFA referee	Dummy: Equals 1 if referee is an official FIFA referee, 0 otherwise	3,968	0.24	-	0	1
Temporary promotion	Dummy: Equals 1 if referee is assigned to a match of a division that is higher than his qualification level, 0 otherwise	3,968	0.01	-	0	1
CONTROL VARIABLES Match-Specific Information						
Division	Categorical variable indicating the division in which the match takes place	3,968	2.08	-	1	3
Season	Categorical variable indicating the season in which the match takes place (1 corresponds to season 2008/09 etc.)	3,968	2.50	-	1	4
Match day	Match day in a given season	3,968	18.27	10.32	1	38
Attendance	Stadium attendance[40]	3,968	20,174	19,329	1	80,720
LogAttendance	Log transformation of stadium attendance	3,968	9.33	1.24	0	11.30
Averaged Betting Odds and Heterogeneity (Match Uncertainty) Measures						
Odd home win	Bookmakers' average odd for a home win	3,967	2.25	0.87	1.1	17.5
Odd away win	Bookmakers' average odd for an away win	3,967	3.83	1.90	1.17	22.88
Odd draw	Bookmakers' average odd for a draw	3,967	3.49	0.48	3.03	8.97
WinProb home	Implicit probability of a home win	3,967	0.45	0.12	0.05	0.85

[40] In response to different violations of the rules three matches had to take place in the absence of spectators. In order to "keep" these observations after the log transformation the respective utilization rates were indicated as 1.

WinProb away	Implicit probability of an away win	3,967	0.28	0.11	0.04	0.81
Prob draw	Implicit probability of a draw	3,967	0.27	0.03	0.11	0.30
Payout ratio	Bookmakers' average payout ratio	3,967	0.92	0.01	0.90	0.95
HET	Heterogeneity (match uncertainty) measure	3,967	0.17	0.23	-0.76	0.81
PosHET	HET, taking only positive values (see Appendix B)	3,967	0.23	0.16	0	0.81

Note: The reduced number of observations for all dependent variables is due to the fact that i) for those referees who terminated their career during the observation period it is of course impossible to observe the waiting time until (or the quality/division of) the next assignment, while ii) for the remaining referees in our sample we could only observe the respective last assignment within the observation period but not a potential next assignment beyond the observation period. The same holds true for the 28 referees who were active during the last match day of season 2011/12 in any of the three divisions and for whom information on their subsequent match is certainly missing. Betting odds and, thus, match uncertainty measures are available for all but one game (i.e., the third division fixture between Erfurt and Dresden in the season 2008/09).

3.5.1.1 ESTIMATING POTENTIAL SHORT-TERM SANCTIONS USING OLS REGRESSION

As our empirical analyses are based on various regression functions, which will be explained in more detail below, we model several dependent variables. Testing the hypothesis $H_{waiting\ time}$ requires a response variable that measures a referee's waiting time between his current and his next assignment. In this context, two alternative measures arise, namely (i) *waiting time in days*, and (ii) *waiting time in games*. Although it appears to be more precise at first glance, measuring the waiting time in (calendar) days is prone to bias since match days in German top football leagues do not always follow a weekly rhythm. Quite the contrary, due to regular mid-week league games as well as national cup and international cap breaks, the timespan between two match days usually varies between three days and two weeks. Moreover, every season includes a winter break (usually lasting between three and six weeks), while an extended summer break of two to four months occurs in-between two seasons. Given the variation in time between two match days, it is almost impossible to adequately determine a referee's adjusted waiting time (in calendar days) between two assignments. On that account, it seems appropriate to measure the waiting time in match days. By doing so, we are able to circumvent the obvious bias, as illustrated in the following example: A referee who is active on the penultimate match day of $season_j$ and who is again nominated for the second match day of $season_{j+1}$ might have to pause for more than 100 days, making it difficult to determine the net waiting time in days. On the other hand, the waiting time measured in games (in this case amounting to three games, implying that a referee's next assignment occurs in the third succeeding match) requires no further adjustment.

In order to test whether a referee's performance in one particular match has an immediate effect on the waiting time (wt) until his next assignment, we estimate an OLS model of the following functional form:

$$wt = \beta_0 + \beta_1 grade + \beta_2 appointed\ division + \beta_3 FIFA\ referee + \beta_4 LogAttendance + \beta_5 age + \beta_6 experience + \beta_7 PosHET + \beta_8 controls + \varepsilon_i,$$

where β_0 denotes the intercept with the ordinate, 'controls' stands for season as well as division dummies that are included as control variables in the first model and ε_i is the unexplained random error term. Here and in the following, all models are estimated with clustered and robust standard errors, using *referee ID* as cluster variable. M*atch day* and *temporary promotion* turned out to have no statistically significant effect on the waiting time (nor on the alternative response variables that are explained in the following sections), while leaving the coefficients of the other covariates as well as the explained variance virtually unaffected. Therefore, these variables are omitted in the entire match-level analysis. First, we estimate a pooled model that covers all observations of our dataset. In a second step, we estimate separate regressions for each division to examine division-specific effects.

3.5.1.2 EXAMINATION OF IMMEDIATE SANCTIONS APPLYING PROBIT REGRESSION

Albeit the average waiting time is 2.4 games (standard deviation = 1.6) some referees have to wait significantly longer until their next assignment, thus representing potential outliers that could impact the results. In order to capture only immediate nomination effects we shall look at the nomination outcome in $t_{+1.}$ The corresponding dependent variable, *nomination in* t_{+1} (nom), is a dichotomous (or binary) outcome variable that equals 1 if a referee is nominated for the subsequent match, and is zero otherwise. To explain the correlation between a set of independent variables and a binary outcome variable, the most frequently applied estimation techniques are logit and probit models. Both are considered superior to the linear probability model (LPM) which in fact violates several critical assumptions (see Gujarati and Porter 2009: 543-546; Wooldridge 2013: 559-561). Although the coefficients of both models have to be interpreted differently, logit and probit estimations yield very

similar results.[41] For the sake of brevity, we restrict our analysis to probit regressions. The underlying model is of the following general form:

$$nom = \beta_0 + \beta 1 grade + \beta_2 appointed\ division + \beta_3 FIFA\ referee + \beta_4 LogAttendance + \beta_5 experience + \beta_6 PosHET + \beta_7 season\ controls + \varepsilon_i.$$

In anticipation of the results in section 3.6 it should be stated that depending on the division, we observe countervailing and to some extent counterintuitive effects for FIFA referees. To be more precise, FIFA referees who are active in a first division match in t_0 have the expected higher likelihood of being nominated in t_{+1} than "ordinary" first division referees. On the admittedly rare occasion that a FIFA referee is active in the third division (n = 71 matches, see Figure 3-6 below) the probability of being nominated for the subsequent match is 11 percentage points lower compared to non-FIFA referees who are active in the third division.

This rather surprising result may provide grounds for believing that even top referees are penalized for poor performances in the sense that they are degraded to a lower division match and/ or are forced into a small break. However, the actual cause appears to be somewhat different: The season in the third division usually starts two to four weeks earlier than in the first and second division. During that time, a disproportionately large number of third division matches is officiated by FIFA referees. The same holds true for "ordinary" first and second division referees. Thus, the pool of referees eligible for nomination for the first match days in the third division is relatively large as it includes first, second and third division referees. In contrast, due to the temporarily increased competitive pressure among referees (i.e., a pool of about 60 (instead of the usual 20) referees competes for an assignment in only 10 third division matches per match day), individual chances of being successively nominated are substantially lower. Therefore, the finding that qualified (FIFA) referees appear to be sanctioned following an assignment in the third division is most likely the result of a temporary oversupply of labor at the beginning of the season. In fact, if the

[41] The characteristic difference between both methods is the underlying distribution: logit models are based on a standard logistic distribution while probit models follow a standard normal distribution. Both distributions have a mean value of zero but their variances are different (i.e., 1 for the standard normal and $\pi^2/3$ for the standard logistic distribution). Consequently, the normal distribution has slightly thinner tails than the logistic distribution, leading to marginally different outcomes towards the tails (see Gujarati and Porter 2009: 572 for an illustrative example).

observations from the first two match days of each season are excluded from the estimations, the alleged "sanctioning effect" disappears, while other findings remain unaffected.[42]

3.5.1.3 TESTING SHORT-TERM PROMOTIONS AND DEMOTIONS WITH ORDINAL PROBIT AND POISSON REGRESSION

Notwithstanding the preceding limitations, it seems worthwhile to explicitly test $H_{demotion}$, i.e., whether referees who perform poorly in one particular match have a significantly higher likelihood of being temporarily demoted to a lower division in their subsequent assignment.[43] In contrast to the above-mentioned probit regression model, the response variable, *division in* t_{+x}, can have more than two outcomes and requires a different estimation technique. In that regard, the poisson regression model is a particularly well suited econometric tool as it appropriately models count data.

The characteristic feature of count data is that the dependent variable is discrete and takes only a finite number of non-negative integer values. Sometimes, count data are used to model rare or infrequent events such as the number of patents applied for by a company in a fiscal year. In the underlying study, the response variable can take the values 1, 2 and 3, depending on the assigned division. It should be noted that poisson regression, which is based on maximum likelihood estimators, usually requires a large sample size (Long and Freese 2006; Cameron and Trivedi 2009). This condition is satisfied by our data. Moreover, a statistically insignificant goodness-of-fit chi-squared test suggests that the poisson model fits our data reasonably well.

On the other hand – for the same reasons discussed for binary response variables – a linear model might not adequately reflect all values of the explanatory variables (see Gujarati and Porter 2009: 576-579; Wooldridge 2013: 580-585) and is therefore not considered in the analysis. However, since the dependent variable, division in t_{+x}, is ordinal, it makes sense to additionally estimate an ordinal logit or probit model (commonly referred to as ordered logit/probit). Here again, ordered logit and probit models are interchangeable and lead to almost identical results. In the following, we refer to the latter. The estimation technique of ordered probit is very similar to the bivariate probit model. The distinctive advantage of ordered probit is that it is the best estimator for response variables that have more than two outcomes and which are ordinal in nature, although the arithmetic gets rather complicated

[42] The results of these estimations are available from the authors upon request.
[43] Of course, the reverse case is also conceivable in the sense that particularly good performances are rewarded with an assignment in a higher division.

due to the non-trivial underlying distribution (see Gujarati and Porter 2009: 580). For both poisson and ordered probit we estimate a model of the following functional form:

$$\text{div} = \beta_0 + \beta_1 \text{grade} + \beta_2 \text{appointed division} + \beta_3 \text{FIFA referee} + \beta_4 \text{LogAttendance} + \beta_5 \text{experience} + \beta_6 \text{PosHET} + \beta_7 \text{season controls} + \varepsilon_i,$$

where div is the response variable, *division in* t_{+x}, indicating the division of a referee's subsequent assignment. All estimation results are reported in section 3.6. In the following section, we describe our second dataset that includes aggregate referee information at the season-level.

3.5.2 SEASON-LEVEL DATA

In order to investigate empirically whether (i) a referee's average performance as well as the performance consistency throughout the season have a statistically significant effect on the likelihood of (permanent) promotion to a higher division (i.e., $H_{promotion}$), and whether (ii) top division referees are in fact better than lower division referees in terms of average performance (i.e., $H_{ability}$), we compiled a second dataset with aggregate season-level information. To explicitly test the first hypothesis, we use a binary response variable, *promotion*, which indicates whether or not a referee is promoted to a higher division in the subsequent season.

In addition to that, we look at the reverse outcome(s), namely that a referee is (iii) permanently demoted to a lower division – or to a lower status in the case of FIFA referees – (i.e., *demotion*), exits the sample (iv) prematurely (i.e., *exit*) or (v) due to reaching the age threshold of 47 (i.e., *retirement*). Whereas *retirement* and *demotion* only occur rarely, we observe quite a number of premature exits every season (which might include demotions to the fourth division). Merging the observations of the two latter variables yields our second response variable of interest, *demotion or exit*, which is also a binary outcome variable. For the purpose of assessing individual career outcomes following the season 2011/12, we include information on a referee's status at the beginning of the season 2012/13. For detailed summary statistics see Table 3-4.

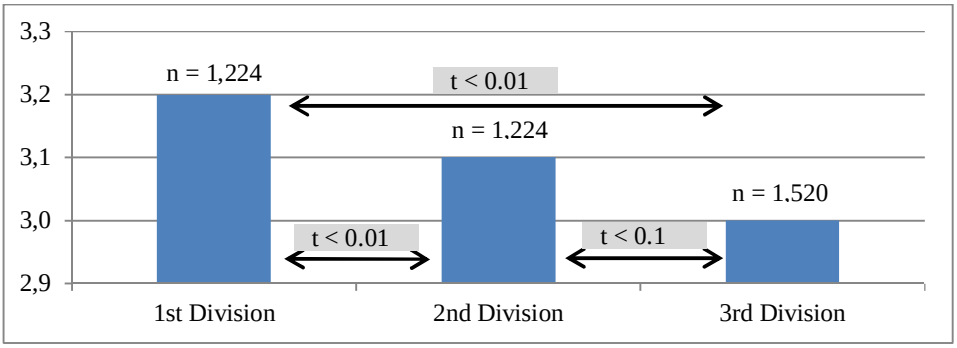
Table 3-4: Summary statistics of the season-level analysis on referee performance and career progress

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLES						
Promotion	Dummy: Equals 1 if referee is promoted to higher division or status in the subsequent season, 0 otherwise	251	0.10	-	0	1
Demotion or exit	Dummy: Equals 1 if referee is demoted to lower division or status in the subsequent season or retires early, 0 otherwise	251	0.12	-	0	1
Demotion	Dummy: Equals 1 if referee is demoted to lower division or status in the subsequent season, 0 otherwise	251	0.01	-	0	1
Exit	Dummy: Equals 1 if referee retires early, 0 otherwise	251	0.11	-	0	1
Retirement	Dummy: Equals 1 if referee retires due to the official age limit, 0 otherwise	251	0.01	-	0	1
INDEPENDENT VARIABLES						
AvgGrade	Average grade of all assignments of $referee_i$ in $season_j$	251	3.19	0.36	2.31	4.39
CV	Coefficient of variation of avgGrade (standard dev. divided by arithmetic mean), indicating performance consistency	251	0.33	0.07	0.12	0.53
Age	Referee's age	251	31.60	5.80	20	46
Appointed division	Categorical variable indicating a referee's qualification level (i.e., the division appointed prior to the season)	251	2.02	-	1	3
FIFA referee	Dummy: Equals 1 if referee is an official FIFA referee, 0 otherwise	251	0.15	-	0	1
Season assignments	Number of assignments of $referee_i$ in $season_j$	251	11.20	3.73	3	21
Temporary promotion	Dummy: Equals 1 if referee officiated in a higher division than was suggested by his appointed division in at least one match of the season, 0 otherwise	251	0.02	-	0	1
Season	Categorical variable indicating the season (1 corresponds to season 2008/09 etc.)	251	2.50	-	1	4

Among all independent variables, the individual average grade of all season assignments (*avgGrade*) is our main variable of interest. Using all 251 referee-season observations, we obtain an arithmetic mean of *avgGrade* of 3.19. Since outliers in performance in single matches are likely to be smoothed out throughout the season, *avgGrade* displays a rather

low standard deviation. On the other hand, the difference between the best seasonal average performance (2.31) and the worst outcome (4.39) is quite substantial, raising the question of whether individual performance or performance evaluations vary contingent upon the division in which the match takes place. Therefore, we separately computed the average grade for all matches played in the first, second and third division, respectively (see Figure 3-5).

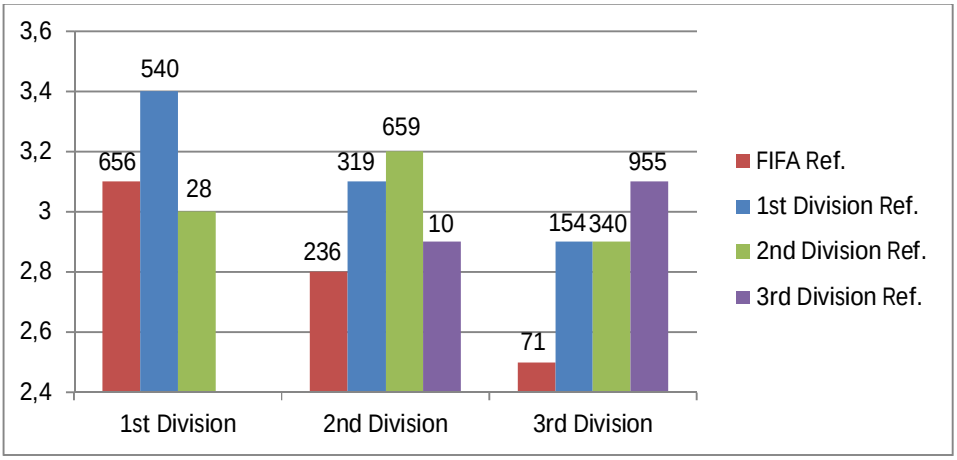
Figure 3-5: Average individual referee performance separated by division



Note: Average individual referee performance expressed by average referee grade (y-axis), separated by division (x-axis). Source: Own calculations based on data from www.kicker.de.

It appears from Figure 3-5 that referees who are assigned to a first division match receive on average poorer grades than referees who are active in a second division match which, in turn, tend to perform significantly worse than referees who monitor matches in the third division. But does this rather surprising descriptive result necessarily imply that designated first division referees perform worse than referees with a lower qualification level (or that their performances are simply evaluated more strictly)? Figure 3-6 aims to shed light on this somewhat ambiguous issue.

Figure 3-6: Average individual referee performance separated by division and qualification level



Source: Own calculations based on data from "Kicker Online".

The above diagram shows the average performance (i.e., the average grade) of referees (y-axis) conditional on the division of their individual assignment (x-axis) and their qualification level (colored bars). The number of observations for each allocation is indicated on top of each bar. With respect to first division matches it appears that FIFA referees receive on average significantly better grades than "ordinary" first division referees. Perhaps surprisingly, designated second division referees who are temporarily promoted to a first division match on average receive the best grades. However, only about two percent of all first division matches are monitored by second division referees. These referees are presumably the top performers among their peers and thus represent a positively selected sub-group. At least the underlying data reveal that four out of five designated second division referees who are temporarily promoted to one or more matches of the next higher division in a particular season are permanently promoted to the first division after the season.

A similar picture can be observed with regard to referee performance in second division matches. As expected, the higher the qualification level, the better the average performance. Again, in the rare case of a temporary promotion (n = 10 matches, monitored by 2 different referees who are both promoted to the second division after the season), designated third division referees seem to outperform their higher qualified colleagues – with the exception of FIFA referees. The sub-sample comprising all third division matches is probably most suitable to compare the performances of referees with varying levels of qualification as the number of observations for each qualification level is sufficiently high. Here,

FIFA referees receive on average by far the best grades, while "ordinary" first and second division referees seem to perform significantly better than third division referees.

These preliminary results are in line with the initial hypothesis, $H_{ability}$, stating that better qualified referees should, on average, perform better and thus receive better average grades than their less qualified colleagues. To take up the question that emerged from Figure 3-5 above, it appears from Figure 3-6 that referee performance in fact deteriorates towards higher leagues, irrespective of the individual qualification level. That is, referees who are active in a first division match tend to receive poorer grades than equally qualified colleagues who are appointed to a second or third division match. This holds true for FIFA and first division referees active in all divisions as well as second division referees active in the two lower divisions, while third division referees are almost exclusively assigned to matches in their appointed division.

What remains to be tested is whether the observed effect indeed reflects a drop in performance or whether poorer grades are simply the result of stricter performance assessments. Both are complementary rather than substitutable explanations. On the one hand, in line with most of the literature related to crowd effects (see section 3.2.1.4), the increasing number of spectators in higher divisions is likely to adversely affect referee performance. On the other hand, one of the side effects of the increasing media coverage of higher division football is that all individuals are subject to increased monitoring. As a consequence, the probability of detecting false referee decisions – be it a wrong offside call or a mistakenly awarded red card – is undoubtedly higher in a first division match where critical scenes are repeatedly shown from multiple camera angles than in a sparsely broadcasted lower division match.

Summarizing, the aim of the season-level analysis is to identify the determinants of long-term career success and failure. For that purpose, we estimate two different semi-parametric proportional hazard models (see Cox 1972) as well as two probit models which have the following general form:

$$\text{succ./fail.} = \beta_0 + \beta_1 \text{avgGrade} + \beta_2 \text{CV} + \beta_3 \text{appointed division} + \beta_4 \text{FIFA referee} + \beta_5 \text{season assignments} + \beta_6 \text{temporary promotion} + \beta_7 \text{age} + \beta_8 \text{season controls} + \varepsilon_i,$$

where succ./fail. stands for the respective response variable, *promotion* (*demotion or exit*), which equals 1 in case of a referee's promotion to a higher division/status (demotion to a lower division/status or early retirement) in the subsequent season, and 0 otherwise. Using the arithmetic mean and sample standard deviation of individual referee performance, we also compute the coefficient of variation (*CV*) which serves as a performance consistency measure. *Season assignments* is thought to be a good indicator of individual career success as the number of assignments per season should reflect both a referee's current reputation and the likelihood of success or failure in t_{+1}. *Temporary promotion* indicates whether a referee is assigned to one or more matches in a division higher than his actual qualification level at some point in the season.

It should also be mentioned that in the season-level analysis we use *age* instead of *experience* as an explanatory variable. The reason is that with respect to long-term career outcomes *age* supposedly has more explanatory power than is the case in short-term nomination decisions. For example, a referee aged 45 might have similar chances of being nominated for an important match compared to an equally endowed colleague who is a few years younger. When it comes to permanent promotion decisions, however, individuals approaching retirement age might potentially fare worse than, say, a 35-year-old referee with similar characteristics. This, of course, presupposes that the DFB aims at promoting its top referees at a rather young age, thus reducing turnover at the top level of refereeing and, at the same time, giving the best referees the opportunity to officiate in international matches for an extended period of time.

These assumptions – along with the hypotheses developed in section 3.3 – need to be tested econometrically. We also aim to shed additional light on the descriptive results emerging from the illustration and discussion of the datasets. The results of our empirical analyses are presented in the next section.

3.6 EMPIRICAL FINDINGS

In the following, we report the estimation results of both our match-level analyses, examining the influencing factors of referees' nomination outcomes in the short run (section 3.6.1), and our season-level analyses, investigating which factors determine career progress of referees in the long run (section 3.6.2). A short summary of the results is provided in section 3.6.3.

3.6.1 THE IMPACT OF REFEREE PERFORMANCE ON SHORT-TERM NOMINATION OUTCOMES

We begin with the estimation results of the pooled and separate OLS regressions (Table 3-5). Recall that here and in the following two result tables, all models are estimated using robust standard errors (adjusting for clustering of observations by *Referee ID*).

Table 3-5: OLS estimation results regarding potential short-term sanctions of referees

	Pooled OLS			Separate OLS		
Response Variable	Waiting Time in Games					
	All divisions			Division 1	Division 2	Division 3
Covariates						
Grade	0.019	0.024	0.021	0.019	0.059	-0.038
	(0.023)	(0.024)	(0.024)	(0.019)	(0.036)	(0.065)
Appointed division						
1	Reference category			Reference category		
2	**1.174*** **	**1.242*** **	**1.145*** **	0.050	**1.575*** **	**0.588*** **
	(0.105)	(0.102)	(0.116)	(0.260)	(0.118)	(0.107)
3	**1.708*** **	**1.759*** **	**1.629*** **	-	0.124	**1.382*** **
	(0.099)	(0.107)	(0.104)		(0.193)	(0.086)
FIFA referee	0.004	**-0.151*** **	-0.054	**-0.105***	-0.079	**0.329** **
	(0.052)	(0.047)	(0.041)	(0.059)	(0.088)	(0.135)
LogAttendance	-0.083**	-0.086**	-0.081**	-0.062	-0.058*	-0.076
	(0.037)	(0.037)	(0.037)	(0.050)	(0.032)	(0.057)
Age	0.025**	0.006	-	-	-	-
	(0.011)	(0.007)				
Experience	-0.002***	-	-0.001***	-0.001**	-0.000	-0.002**
	(0.001)		(0.000)	(0.000)	(0.001)	(0.001)
PosHET	0.020	0.018	0.018	0.040	**-0.442***	0.382
	(0.130)	(0.128)	(0.130)	(0.146)	(0.224)	(0.352)
Constant	1.985**	2.450***	2.716***	2.536***	2.406***	2.431***
	(0.421)	(0.383)	(0.368)	(0.519)	(0.351)	(0.506)
Season and division dummies included	Yes	Yes	Yes	Only season dummies	Only season dummies	Only season dummies
Observations	3,878	3,878	3,878	1,210	1,203	1,465
F	173.80***	168.95***	171.97***	3.17**	74.54***	92.94***
R^2	0.219	0.216	0.217	0.022	0.321	0.094

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Coefficients denoting statistical *and* economic significance are in bold.

In the first three columns, we report the pooled OLS regressions results which include all observations of our match-level dataset. All three model specifications are very similar and yield almost identical results. The only difference is the alternating inclusion of *age* and *experience*. Since both variables are highly correlated (at 0.86), the interpretation of the coefficients in the first model specification is likely to be biased due to multicollinearity.

A weakness of the second specification is that *age* alone appears to be a poor predictor of individual quality. *Experience*, on the other hand, seems to be a better predictor, as can be inferred from (i) its statistically significant (albeit economically irrelevant) coefficient, (ii) the "absorbance" of the *FIFA referee* effect observed in the second specification, as well as (iii) a marginally larger coefficient of determination (i.e., "R^2") in the third model specification as compared to the second one. The purpose of displaying the results of all three specifications (instead of presenting only the third column) is to demonstrate that *experience* is a better predictor of individual refereeing quality (and nomination outcomes). Columns 4-6 display the results of the models that we estimated separately for each division.

Perhaps the most striking result is that referee performance (i.e., *grade*) seems to have no statistically significant impact on the nomination outcome in the short-term. In other words, a referee's performance in one particular match seems to have no effect on the length of time until his subsequent employment. In what appears to be the main determinant of short-term nomination outcomes, a referee's qualification level (i.e., *appointed division* and, to a lesser extent, *FIFA referee*) exhibits both a statistically and economically significant correlation with individual waiting time: The higher a referee's qualification[44], the shorter the expected time period until the next assignment.[45] For example, a designated second division referee who is active in a second division match has to wait on average 1.6 games longer until his next assignment than a designated first division referee active in the same division, irrespective of the refereeing quality in that particular match.[46] In terms of earnings, longer waiting times between two assignments imply fewer nominations per season and thus considerable income losses. In the present example, second division referees have approximately five assignments less per season, resulting in foregone income of about 10,000 €. Regarding the other covariates of interest, reflecting stadium attendance, refereeing experience and match uncertainty, we are unable to find persistent and both statistically and economically significant effects.

To sum up, it appears that the waiting time between two assignments (and thus the total number of employments per season) is independent of individual referee performance in

[44] Recall that qualification increases with decreasing values for *appointed division*.
[45] The relatively low R^2 (of about 2 percent) in the separate estimation for the first division can be explained by the fact that out of 1,224 first division matches in our dataset only 28 were monitored by designated second division referees (see Figure 3-6), hence causing the major explanatory variable, *appointed division*, to offer almost no variation. Recall that the negative coefficient of *FIFA referee* in the separate estimation for the third division is due to the peculiarities at the start of the season that were discussed in section 3.5.1.2.
[46] Under normal circumstances (i.e., match days are scheduled according to a weekly rhythm), an additional waiting time of 1.6 games is equivalent to about 10 calendar days.

the previous match but rather seems to be predetermined by a referee's qualification level. Nevertheless, it is worth testing for potential immediate sanctioning effects. Table 3-6 lists the results of several probit estimations that seek to identify the determinants of a referee's nomination outcome in t_{+1}. For ease of interpretation, we report marginal effects at the means of the covariates.[47]

Table 3-6: Marginal effects after probit regression testing for potential immediate sanctions of referees

Marginal effects after probit regression Response Variable	Nomination in t_{+1}		
	Division 1	Division 2	Division 3
Covariates			
Grade	-0.003	-0.023*	0.002
	(0.011)	(0.014)	(0.014)
Appointed division			
1	Reference category		
2	-0.035	**-0.403*** **	**-0.143*** **
	(0.118)	(0.040)	(0.051)
3	-	**0.212***	**-0.250*** **
		(0.129)	(0.046)
FIFA referee	**0.071****	0.054	**-0.108*** **
	(0.029)	(0.036)	(0.032)
LogAttendance	**0.084*** **	0.006	0.025**
	(0.029)	(0.021)	(0.013)
Experience	0.000	0.000	0.001*
	(0.000)	(0.000)	(0.000)
PosHET	-0.082	**0.149***	**-0.156***
	(0.069)	(0.083)	(0.085)
Season dummies included	Yes	Yes	Yes
Observations	1,215	1,215	1,508
Wald Chi²	33.01***	206.79***	134.95***
Pseudo R²	0.011	0.198	0.062

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Marginal effects are computed at the means of all covariates, reporting a discrete change from the base level for categorical variables and an infinitesimal change for continuous variables.
Coefficients denoting statistical *and* economic significance are in bold.

The empirical evidence from the above table reveals a similar picture as in the previous estimations. Due to the low variation of *appointed division* in the first model specification, it is not surprising that the 'Pseudo R²' is quite low, too. Again, *grade* appears to have no effect on the probability of a nomination in t_{+1}. On the other hand, *FIFA referee* and

[47] Instead of the traditional 'mfx compute' Stata command we use the more recent 'margins' command, available in Stata 11 and newer versions, that enables us to accurately predict the values of independent categorical variables. That is, categorical variables (or factor variables), as for instance *appointed division*, are treated as such, while simultaneously computing the probabilities for all continuous variables as in the traditional 'mfx compute' command.

LogAttendance have the expected positive impact on the immediate nomination outcome: FIFA referees who are active in a first division match have a 7 percentage points higher chance of being nominated (for a match in any division) on the following match day than referees without FIFA status. Umpires who are assigned to matches/stadia attracting a huge crowd have a significantly higher likelihood of nomination in t_{+1} than referees who are nominated for less attended matches, all else being equal. When interpreting the coefficients, however, one has to be cautious with the direction of causality. It seems more likely that those referees who are selected for important matches are per se the ones who are favored by the DFB in terms of the number and quality of nominations.

The empirical evidence derived from matches played in the second and third division is perhaps more convincing because of the far more heterogeneous pool of referees. As a consequence, *appointed division* once more appears to be the main influencing factor of the nomination outcome: The better the qualification, the higher the likelihood of a successive nomination. For example, a designated first division referee active in a second division match has a 40 percentage points higher chance of being nominated in t_{+1} than a second division referee, all other conditions remaining the same. Interestingly, third division referees who are temporarily promoted to the second division fare significantly better than first and second division referees. Whether this result reflects favorable treatment of promotion candidates or simply represents a "statistical artefact" due to the small number of observations (n = 10) cannot be conclusively answered with the data used here. The counterintuitive result for FIFA referees active in the third division has been discussed before. *Grade* is statistically significant at the 10 percent level in the second model specification. The economic relevance, however, is rather low. A one unit change of *grade* – which is already substantial, given a mean of 3.11 and a standard deviation of 1.08 – leads to a two percentage points change in the probability of nomination. Match uncertainty (i.e., *PosH-ET*) has a statistically barely significant impact with the expected direction in the third division. That is, referees who are assigned to matches with two equally strong teams are more likely to be nominated in the subsequent match than similarly endowed referees monitoring two rather uneven opponents. Of course, this result is indicative of selection effects in the sense that "more able" referees are assigned to more balanced matches. Perhaps surprisingly, the reverse effect seems to be true for referees who are active in the second division. Therefore, the overall effect of match uncertainty should be interpreted cautiously.

The empirical analyses presented so far include an outcome variable that basically reflects a referee's waiting time between two assignments and the probability of a consecutive nomination, respectively. In both cases, we can draw conclusions as to what determines the mere number or frequency of assignments. In order to learn more about the determinants of the (income-relevant) quality (i.e., the division) of individual assignments and to empirically answer the question of whether poor referee performance is sanctioned in the form of a temporary demotion to a lower division, we have estimated separate poisson and ordered probit regressions for each division.

Table 3-7: Poisson and ordered probit estimation results on qualitative sanctions for referees

	Poisson Regression			Ordered Probit Regression		
Response Variable	Division in t_{+x}					
	Division 1	Division 2	Division 3	Division 1	Division 2	Division 3
Covariates						
Grade	0.014	-0.001	-0.005	0.038	-0.003	-0.102
	(0.010)	(0.007)	(0.004)	(0.026)	(0.028)	(0.063)
Appointed division:						
1	Reference category			Reference category		
2	-0.010	**0.426*** **	**0.346*** **	-0.023	**1.384*** **	**1.471*** **
	(0.039)	(0.050)	(0.040)	(0.105)	(0.192)	(0.192)
3	-	**0.556*** **	**0.677*** **	-	**2.221*** **	**4.570*** **
		(0.075)	(0.044)		(0.546)	(0.302)
FIFA referee	**-0.113*** **	-0.033	**-0.260*** **	**-0.308*** **	-0.079	**-1.000*** **
	(0.036)	(0.059)	(0.068)	(0.098)	(0.213)	(0.278)
Experience	-0.001***	-0.001***	0.000	-0.001***	-0.004***	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
LogAttendance	-0.019	0.025**	-0.005	-0.048	0.098**	-0.089
	(0.035)	(0.010)	(0.003)	(0.095)	(0.041)	(0.055)
PosHET	-0.021	-0.050	0.029	-0.051	-0.174	0.370
	(0.059)	(0.062)	(0.026)	(0.162)	(0.248)	(0.376)
Constant	0.765***	0.278***	0.462***	-	-	-
	(0.371)	(0.107)	(0.050)			
Season dummies included	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,210	1,203	1,465	1,210	1,203	1,465
Wald Chi²	93.38***	704.64***	2063.56***	88.02***	242.42***	409.55***
Pseudo R²	0.005	0.056	0.048	0.025	0.269	0.636

Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The computation and interpretation of marginal effects is not very useful in regression analyses where the dependent variable is a multiple outcome variable.[48] In particular with

[48] By estimating marginal effects we would undo the very advantage of poisson and ordered probit regression: The point of applying these kinds of regression techniques is that one can use the ordered nature of the

respect to ordered probit one has to be very cautious when interpreting the coefficients in this model (Greene 2000: 878). Therefore, we report 'plain' regression results which allow determining the direction and level of statistical significance, but not the exact magnitude of each effect. Despite this limitation, we obtain unambiguous results that are robust to different specifications. First, and in line with the above results, we are unable to find evidence for any performance-related short-term sanctions. In other words, no matter how poor (or excellent) a referee's performance in one particular match, it will not affect the appointed division (and, thus, the level of remuneration) in the subsequent assignment. Second, individual qualification/status appears to be the main determinant of the outcome variable: FIFA referees are less likely to be assigned to a lower division in t_{+x} than "ordinary" first division referees.[49] Moreover, the evidence from second and third division matches reveals that *appointed division* has the expected impact on the nomination outcome: The higher a referee's qualification level, the more likely is ceteris paribus a subsequent assignment to a higher division.[50] However, the latter effect has to be interpreted carefully due to a potential endogeneity problem: Since only few designated second (third) division referees are assigned to first (second) division matches, while third division referees are categorically excluded from first division matches (see Figure 3-6), a large part of the effect is likely to be caused by unobservable nomination restrictions on the part of the DFB.

All in all, the empirical evidence presented in this section suggests that short-term nomination decisions are, by and large, based on ex ante selection effects rather than the result of immediate performance evaluations.

response variable to estimate the overall effect for each explanatory variable. Marginal effects, on the other hand, show the effects for each category of the response variable, similar to a multinomial probit or logit regression (Liao 1994: 41-47; Borooah 2002: 12-15), which is clearly not the most suitable way to interpret our results.

[49] Note that this effect is statistically insignificant in the specifications which are estimated separately for second division matches.

[50] Recall that larger numbers are associated with a lower qualification level with respect to the x-variable and a lower division of assignment regarding the y-variable. With the first division being the reference category, positive coefficients indicate an expected lower division of assignment in the subsequent match.

3.6.2 THE IMPACT OF REFEREE PERFORMANCE ON LONG-TERM CAREER PROGRESS

In this section, we present and discuss the empirical evidence on the determinants of long-term career progress.

Table 3-8: Hazard ratios and marginal effects for career progress

	Cox Proportional-Hazard Model (1.1)	Probit Model (1.2)	Cox Proportional-Hazard Model (2.1)	Probit Model (2.2)
Response Variable	Promotion		Demotion or Exit	
	Hazard Ratio	dy/dx	Hazard Ratio	dy/dx
Covariates				
AvgGrade	**0.070*** **	**-0.159*** **	0.473	-0.043
	(0.056)	(0.045)	(0.299)	(0.054)
CV	0.080	-0.149	0.290	0.006
	(0.216)	(0.186)	(0.826)	(0.230)
Appointed division				
1		Reference category		
2	0.236	-0.033	0.767	-0.033
	(0.279)	(0.044)	(0.570)	(0.041)
3	0.446	-0.038	**6.763** **	**0.197** **
	(0.429)	(0.047)	(5.411)	(0.092)
FIFA referee	-	-	0.708	0.050
			(0.673)	(0.077)
Season assignments	0.911	-0.004	**0.747*** **	**-0.030*** **
	(0.104)	(0.006)	(0.062)	(0.008)
Temporary promotion	**12.17*** **	-	-	-
	(6.571)			
Age	**0.891** *	**-0.009** **	**1.162*** **	**0.010*** **
	(0.061)	(0.004)	(0.060)	(0.004)
Season dummies included	-	Yes	-	Yes
Observations	251	251	251	251
No. of Failures	24	-	29	-
Time at Risk	627	-	627	-
LR Chi² / Wald Chi²	35.06***	38.43***	28.93***	55.51**
Pseudo R^2	-	0.151	-	0.213

In parentheses, we report robust standard errors for models 1.1 and 2.1, while we report clustered robust standard errors for models 1.2 and 2.2.

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

Marginal effects are computed at the means of all covariates, reporting a discrete change from the base level for categorical variables and an infinitesimal change for continuous variables. In some estimations, *temporary promotion* predicts success and failure perfectly and is therefore omitted. *FIFA referee* is omitted in the first two model specifications due to the impossibility of further promotion. Season dummies are of course not included in the Cox estimations, while the inclusion of a time trend left the results unchanged. Probit regressions are estimated with clustered and robust standard errors, using *referee ID* as cluster variable. Results are robust to models estimated separately for each division, which are available from the authors upon request.

In Table 3-8 we report the results of our estimations. We present both hazard ratios[51] and marginal effects that have to be interpreted differently. The hazard ratios in the first model specification indicate an individual's probability of being promoted to a higher division/status in t_{+1}, conditional on having been active until the previous season. Hazard ratios smaller than 1 imply a reduction of the promotion probability. The closer the hazard ratio is to zero, the less likely is the event to be explained. In the underlying example, *age* displays a hazard ratio of 0.89, implying that a one year older referee has a ceteris paribus 11 percent lower likelihood of promotion than his younger counterpart. That is, every additional year of age appears to impede career progress quite substantially. Although this result is barely statistically significant at the 0.1 level, it remains robust in the probit specification (see results in the second column), albeit the magnitude of the effect is smaller there. Moreover, in accordance with our assumption in section 3.5.2, referees who are temporarily promoted to one or more matches of the next higher division in a particular season are almost certainly permanently promoted to the higher division after the season. Since *temporary promotion* almost perfectly predicts *promotion* (as well as *demotion or exit*), it is not included in the remaining estimations.

A particularly striking result is that although referee performance appears to have no impact on short-term nomination outcomes (as shown in section 3.6.1), *avgGrade* has a statistically significant and economically relevant effect on long-term career advancement (i.e., the probability of promotion): The better the average performance throughout the season, the higher the likelihood of promotion. This result is robust to a wide range of different specifications (of which only the above Cox and probit estimations are presented).

Perhaps surprisingly, neither performance consistency, *CV*, nor the number of nominations per season, *season assignments*, nor a referee's qualification level, *appointed division*, have an effect on long-term career progress. In the face of these results, it appears that the DFB's tournament-like ranking system is effective – at least in the long term – in identifying and rewarding the on average most productive agents.

[51] A methodological note with regard to the Cox proportional hazard models is warranted: Unlike in "traditional" survival analysis, where subjects exit the sample after the event of interest (e.g. death or career end), the individuals in our sample are usually still active after a promotion (or demotion) and hence need to be continuously observed. Otherwise, one would waste possibly relevant information. Moreover, some referees are promoted to a higher division/status more than once during the observation period. One way of analyzing multiple failure time data without losing relevant information is to estimate the models without specifying a multiple record ID variable (here: *referee ID*) in Stata.

The empirical evidence on the determinants of career setbacks is somewhat different. First, neither *avgGrade* nor *CV* seem to have a systematic impact on the probability of a referee's demotion to a lower division or a premature end of the refereeing career. Instead, career failure appears to be strongly affected by a referee's age as well as his current qualification level (i.e., *appointed division*). More specifically, designated third division referees are at a far greater risk of being demoted or of terminating their career[52] than referees with similar characteristics, but who are "higher up the hierarchy". Second, and in line with the above finding that every additional year of age significantly impedes career advancement, *age* also has the expected impact on career setbacks: The older a referee, the more likely is ceteris paribus some form of career-related failure. Third, it appears that the number of individual assignments per season affects the outcome in the expected direction. That is, the larger the number of assignments per season, the lower is the likelihood of individual failure, and vice versa. This result, however, should be interpreted with caution, as the direction of causality is unclear. Do referees react to personal underemployment (and thus foregone income) by terminating their career and instead falling back on potentially more lucrative outside options? Or is a smaller number of assignments the result of a long-planned career decision on behalf of the DFB? Besides that, one should be aware of a potential endogeneity issue: Referees who terminate their career in mid-season almost necessarily have fewer assignments than their colleagues. The data do not allow determining whether a referee in fact terminates his career in the middle of the season, or whether he is simply not considered for nomination in the second part of the season. This information is contained in the error term which in turn might be correlated with the independent variable (i.e., *season assignments*).

3.6.3 SUMMARY AND DISCUSSION OF THE RESULTS

Irrespective of the already mentioned limitations, the intuition arising from the accumulated evidence on the determinants of referees' (short- and long-term) career outcomes is straightforward.

With regard to the match-level analyses it appears that short-term nomination decisions are unaffected by individual performance (in the previous match). That is, even very poor refereeing does not entail any short-term sanctions whatsoever; nor are particularly good per-

[52] Due to data limitations – refereeing activity is only observed in the three highest divisions – we cannot distinguish between demotion and exit in the case of third division referees. Yet, anecdotal evidence suggests that only few referees are in fact demoted to a lower division while the majority terminates their career prematurely.

formances rewarded with the prospect of a follow-up assignment and/or a nomination to a higher division (which would be associated with considerably higher earnings). In what seems to be the main determinant of short-term nomination decisions, the individual qualification level (i.e., *appointed division*) significantly influences *when*, *where* and *how often* individual assignments occur. These findings lead one to suspect that the DFB – contrary to official statements – determines referee assignments well in advance (thus ruling out potential short-term performance effects).

The season-level analyses, in turn, suggest that in the long term, individual performance indeed matters and has a statistically significant impact on career progress: The better a referee's average performance throughout a particular season (as measured by individual "Kicker" grades), the more likely he is to get promoted to a higher division or status in the subsequent season. Those failing to "move up the ladder" are increasingly threatened by demotion or even premature termination of their career with increasing age.

3.7 Conclusions, Limitations and Implications for Future Research

Applying a wide range of different regression models to two distinct and hitherto unavailable datasets that include referee- and match-specific information spanning four seasons of the top 3 German football divisions, we analyze the determinants of referees' career outcomes. Although (short-term) nomination decisions at the match-level appear to be independent of a referee's performance in the previous assignment (but are rather subject to ex ante selection effects), long-term career progress seems to heavily depend on individual performance. Thus, in line with the recommendations emerging from tournament theory, the DFB apparently succeeds in identifying and rewarding the most productive agents in the long term. This, in turn, suggests that, in the long run, labor market inefficiencies due to favoritism, random selection, and the like are presumably rather small (if existent at all). The evidence resulting from the match-level analyses, however, gives rise to question the effectiveness of the football association's recurring hiring decisions. If neither poor nor particularly good performances in single matches have an effect upon the subsequent nomination outcome, individuals are perhaps less inclined to constantly exert maximum effort. Short-term nomination decisions with a stronger focus on the most recent individual performance could help to align the interests of the football association and the referees.

Moreover, it appears that the DFB prefers an 'up-or-out' principle to career progression. That is, referees are either promoted to the next higher division on a regular basis or somehow forced to terminate their career. Whether the (few) observed cases of premature career terminations are indeed actively enforced by the DFB or not, cannot explicitly be tested with the available data. It is also conceivable that referees who are "stuck" in a lower division and who are attracted by potentially more lucrative outside options more or less voluntarily end their career. Qualitative data (e.g. interviews with referees who quit their career prematurely) could shed further light on this matter.

A limitation of this study is that all information concerning referee performance is based on performance evaluations conducted by independent "Kicker experts", whereas official evaluations carried out by the DFB were not available for this period. Yet anecdotal evidence suggests that there is a rather high correlation between both grading systems. After all, "Kicker" as well as official DFB grades are disclosed (internally, for the latter) a few days after each match. Given the extensive broadcasting of critical referee decisions in the meantime, systematic differences in the performance assessments of referees appear very unlikely.

Finally, the recent professionalization of referees in German association football represents an institutional change worth investigating. Future research in this field could examine whether the (still ongoing) modification of the pay regime has an effect on e.g. individual career length. One would expect that the introduction of an additional annual fixed salary reduces the expected value of the outside option and thus induces referees to "stay in the business" for as long as possible.

3.8 APPENDIX B

The following example intends to illustrate how betting odds can be used to quantify the degree of heterogeneity (i.e., match uncertainty) of opposing teams. On match day 13 of the 2011/12 season, Bayern Munich played Borussia Dortmund at home, and was clearly favored by the bookmakers. The odds for a home win, draw and away win were 1.60, 3.89 and 5.50, respectively, implying that bettors received an amount of 5.50 € for every euro they had placed on the away team (as Borussia Dortmund won this match). By summing up the inverse of these quotes, we obtain the so-called payout ratio:

$$\text{payout ratio} = \frac{1}{\frac{1}{\text{Payoff home win}} + \frac{1}{\text{Payoff draw}} + \frac{1}{\text{Payoff away win}}}.$$

Using the above odds yields the following equation:

$$\text{payout ratio} = \frac{1}{\frac{1}{1.60} + \frac{1}{3.89} + \frac{1}{5.50}},$$

which results in a payout ratio of 0.94, indicating that 94 percent of the money staked by bettors is paid out to them again, while the remaining share of 6 percent represents the bookmakers' average margin for this particular game (which indeed varies considerably between different bookmakers). In the next step, the implicit winning probabilities can be calculated by dividing the payout ratio by the payoffs associated with the different match outcomes. This yields an implicit probability of a home win of 0.59, while the implicit probabilities of a draw and an away win are 0.24 and 0.17, respectively. In order to quantify the degree of heterogeneity between the teams, *HET* is introduced which is defined as the difference between the implicit probabilities of a home win (P_H) and an away win (P_A):

$$HET = P_H - P_A.$$

HET can take values between -1 and +1, where negative values are indicative of the away team being favored by the bookmakers while positive values point at a home team favorite. Values close to zero suggest that the opposing teams are homogeneous in their abilities (and prospects of winning) and hence imply a high degree of match uncertainty. Values close to -1 and +1, on the other hand, indicate a presumably less balanced match with rather heterogeneous opponents. In the above example *HET* amounts to 0.42, identifying Bayern Munich as clear home favorite (the average home advantage – as expressed by a *HET* value of 0.17 in Table 3-3 – is considerably smaller).

For ease of interpretation we refrain from a differentiation between home and away favorites and introduce PosHET, which takes only positive values varying between zero and one. The advantage of PosHET is that larger regression coefficients indicate a higher degree of heterogeneity whereas with HET one would have to disentangle the effects of home and away favorites (which is not necessary for the purpose of this study).

4 GENDER DIFFERENCES IN COMPETITIVENESS: EMPIRICAL EVIDENCE FROM LONG-DISTANCE RACES

4.1 INTRODUCTION

The prevalence of a substantial gender gap in competitiveness[53] has been documented and discussed by a large and diverse body of literature. Yet, the reasons for its emergence (and persistence) remain highly contested. One school of thought states that genetic and/ or psychological predispositions cause the observed sex differences. More recent studies show that particularly in the sports environment the gender gap – although still existent – has narrowed over the years (see Frick 2011a, 2011b). This supports the 'culture-and-incentives-hypothesis' which posits that (i) changing socio-cultural conditions, which foster a similar socialization of boys and girls in a growing number of countries over the world, enable more women to engage in competitive sports, and (ii) adjusted incentive systems (such as identical prize money levels and distributions) encourage women to train as hard as equally talented men.

Since long-distance races signal enduring competitiveness and both men and women usually compete at the same time and under equal conditions, they are an ideal setting to analyze gender differences in competitiveness. Using repeated cross-sectional data from the annual long-distance triathlon World Championship (also known as "Ironman Hawaii"), the "Swiss Alpine Marathon", a mountain ultra-marathon, as well as the "Vasaloppet", a long-distance cross-country ski race in Sweden, it appears that women have indeed become more competitive over time (at least in the first two events, while the results are somewhat different for the "Vasaloppet"). This, however, seems to be caused by an influx of (intrinsically) motivated female athletes who cannot reasonably target the top positions (and thus the "money ranks"), suggesting that societal changes rather than monetary incentives explain the observed reduction of gender differences. Hence, this paper adds to the existing (sports economics) literature by providing evidence for a decrease of gender differences in competitiveness among less talented and rather intrinsically motivated contestants.

The remainder of the paper is structured as follows: Section 4.2 provides a selective review of the recent literature – both experimental and non-experimental. Section 4.3 describes the

[53] Competitiveness is defined here as the ability and willingness to deliberately perform a task in a competitive environment.

data and methodology and offers some descriptive evidence. Section 4.4 presents the econometric evidence while the main findings are discussed in section 4.5.

4.2 PREVIOUS EMPIRICAL RESEARCH

Despite recent gender policy developments in industrial countries, women are still underrepresented in leading positions in management and academia. For instance, the boards of directors and top executive positions of the FTSE 100 are predominantly reserved for men: In 2012, only 14.8 percent of the seats were held by women. If only top executive positions are considered, the share of women is considerably smaller, equaling roughly 6 percent (Sealy and Vinnicombe 2013). Although gender inequalities seem to be most pronounced in the labor market – and in particular on the far right tail of the income distribution – there is ample evidence for significant gender differences in various other competitive settings.[54]

Regardless of the already large and still growing literature on gender differences in competitive environments, a consensus concerning the reasons for the emergence (and persistence) of the observed gender gap in competitiveness has yet to be reached. A contested issue, for instance, is whether gender differences are due to genetic differences or behavioral characteristics. While a number of empirical studies suggest that a gender gap in competitiveness is most likely evoked by particular circumstances – such as a promotion tournament, where the pressure to succeed is high – an equally large number of laboratory studies using data from "real life experiments" are unable to find statistically significant gender differences. Finally, a wide-ranging and still increasing number of publications use professional sports data to analyze and explain gender differences of highly (self-)selected

[54] The observable inequalities on the playing field are often – at least to some extent – explained by deeply embedded gender differences caused by inherent traits and societal norms which are only gradually changing (see, inter alia, Gneezy et al. 2009 and Booth and Nolen 2012). A large part, however, has to be attributed to discriminatory practices. Schneider and Bauhoff (2013), for example, find a surprisingly large share of job adverts discriminating with respect to gender *despite* the General Act on Equal Treatment that has been enacted in Germany in 2006. Using the so-called correspondence method to study discrimination in the German labor market of apprentices, Kolle (2014) emphasizes that gender discrimination in hiring seems to be most pronounced in male-dominated jobs. In an attempt to tackle discriminatory practices in the recruitment process – one of the "opposable" causes for gender inequalities – anonymous application procedures (AAP), which are already common practice in the US, were introduced in pilot projects in several European economies. As a result, women had an increased probability of being offered a job under AAP (Åslund and Nordström Skans 2012). Looking at "blind" auditions of professional musicians (each candidate's identity was concealed by a screen) Goldin and Rouse (2000) find similar evidence in the sense that the proportion of women in symphony orchestras increased under the "anonymous" (or "blind") audition.

athletes with similar aspirations and "competitive motivations".[55] Notwithstanding the above-mentioned merits, all three strands of literature also suffer from (major) shortcomings which, to some extent, might explain the difficulties to obtain unanimous results when trying to answer one of the presumably most pertinent questions in the field of gender economics (i.e., do men and women indeed differ in their (competitive) behavior, risk preferences, productivity, and the like, and, if so, how much of the apparent gender inequalities in the field can be explained by these differences?).

4.2.1 REAL-LIFE EVIDENCE

Using real-world data drawn from "naturalistic experiments", a number of studies find considerable gender differences in risk-taking and in performance under competitive pressure. Analyzing quantitative data obtained from test scores of Czech secondary-school graduates applying to tuition-free selective universities, Jurajda and Münich (2008, 2011) find that women perform significantly worse than equally able men in competitive situations. Since they show that women do not shy away from selecting themselves into highly competitive application programs the authors reject differences in risk aversion as a possible explanation for the observed gender gap. Along the same lines, Ors et al. (2008, 2013) provide similar evidence by showing that the performance of female applicants during admission tests at a French elite university is significantly worse than the performance of male applicants. Again, differences in risk aversion and ability cannot explain the observed results, because women perform significantly better in high school exams as well as during their first year after admission to the university, suggesting that women tend to "choke" in competitive environments. Attali et al. (2011) compare the performance on the *Graduate Record Examinations* (GRE), a standardized test that is used by many graduate schools in the US as a selection tool, with the performance of an experimentally designed second GRE taking place immediately after the "real" test. The authors document a greater gender gap in the competitive and largely incentivized "real" GRE than in the experimental GRE which they attribute to men performing better under competitive pressure, i.e. when the stakes are particularly high. Price (2008) examines how students at various high-ranked academic institutions in the US respond to the "Mellon Foundation's Graduate Education

[55] Using data from 100 meter races (ranging from world-class competitions such as the finals of the Olympic Games and the IAAF World Championships to national and regional races in Germany), Frick and Scheel (2013) find that women exhibit a lower degree of competitiveness than men in the finals of federal state and national races while they cannot find any significant gender differences at the international level; i.e. the "closeness" or "competitive balance" of international top races is the same for female and male competitions. Note that large parts of the following literature review are borrowed from Frick and Scheel (2013).

Initiative", a competitive scholarship program that was initially designed to encourage students to make quick progress toward their degree. As a result, the time to candidacy was significantly reduced for male, but not for female students. Moreover, the reduction in time to candidacy was greater for both males and females when a larger fraction of the competing cohort was female.

Hogarth et al. (2012) find gender differences in a competitive environment which they attribute to differences in risk preferences. Exploiting the natural experiment of a TV game show, where candidates answer general knowledge questions in multiple rounds, the authors reveal that women earn 40% less than male contestants and exit the game voluntarily and earlier than men particularly when being in a minority.[56] Corroborating the latter result, Neelakantan (2010) finds that approximately 10% of the gender gap in accumulated wealth of elderly Americans ($194,000 for men and $95,000 for women) can be attributed to differences in risk-taking as women are found to be more conservative investors with a significantly lower probability to invest in stocks.

The following studies are unable to find any persistent behavioral gender differences and thus stand in stark contrast to the findings presented so far. Using data from the British "Workplace Employees Relations Survey", Manning and Saidi (2010) analyze earnings and work effort under performance pay and find very modest (if any) evidence for gender specific differences with respect to sorting into the performance pay scheme. Moreover, they find only a small effect of performance pay on earnings which does not significantly differ by gender. Delfgaauw et al. (2009) conduct a field experiment in a Dutch retail chain consisting of 128 stores. Following the introduction of short-term sales competitions among randomly chosen subsamples of these stores they find large positive effects on sales growth, but only in stores where the manager and a large fraction of the workforce are of the same gender. That is, shops with mostly female staff and a female manager improve their performance under competitive pressure to the same extent as male-dominated shops with a male manager. Interestingly enough, these results hold true even in the absence of

[56] In a similar type of study, using data from the TV game show "The Weakest Link", Antonovics et al. (2009) find that women's likelihood of correctly answering a question is unaffected by the opponent's gender whereas men are more likely to give a correct answer when competing against a woman. Säve-Söderbergh and Lindquist (2011) analyze the behavior of candidates in the game show "Jeopardy" and find that women play more conservatively when facing male competitors only, which is associated with significantly lower earnings for women.

any monetary rewards, suggesting a high symbolic value of winning a tournament.[57] In a similar academic setting as in the aforementioned studies by Jurajda and Münich (2008, 2011) and Ors et al. (2008, 2013), which document the existence of a considerable gender gap in highly competitive and stressful situations, Leuven et al. (2011) find gender differences neither in sorting nor in performance among male and female students of an introductory microeconomics course. In a study using performance data of Israeli teachers, Lavy (2008, 2012) finds that the performance of mathematics and language teachers is neither affected by the introduction of a competitive remuneration scheme, nor by the gender composition of the group of teachers at a particular school. The evidence suggests that after the introduction of a performance-related bonus system neither the average rank, nor the probability of winning a prize (or the size of the prize) differs by gender and that the gender composition of the groups of teachers competing with each other did not affect the performance of female teachers either. Exploring gender differences in risk aversion and negotiating practices, Feidakis and Tsaoussi (2009) compare the behavior of Greek attorneys, Greek business students and a control group consisting of young employees in public and private organizations and are unable to find any statistically significant gender differences among attorneys. They argue that gender differences are unlikely to occur in groups of employees with a distinct professional culture.

4.2.2 LABORATORY EXPERIMENTS

While the majority of the studies quoted above come to the general conclusion that women perform worse under pressure than men, a number of laboratory experiments provide different evidence. Shurchkov (2008) for example asked students from Harvard Business School, MIT and Boston University to perform a verbal task (generally considered to favor women) under varying conditions. The participants were divided in groups of four and required to solve word-in-word puzzles in a limited amount of time.[58] To examine the influence of competitive pressure on participants, the experiment was conducted in two stages. In the first round, a piece-rate reward scheme was used, i.e. participants solved their tasks in a non-competitive environment as their compensation was determined by their absolute rather than their relative performance. In the second round, a winner-takes-all tournament was introduced which only rewarded the winner of a respective group while

[57] This assumption is further supported by the fact that despite the considerable variation in team size the authors are unable to find any evidence for free-riding.
[58] In a word-in-word puzzle the objective is to form as many words as possible from the letters of another long word.

the other group members did not receive any compensation for their efforts. The behavior of male and female participants proved to be similar in the non-competitive as well as in the tournament setting. In fact, both men and women were found to not increase their performance as a result of competitive pressure. Hence, Shurchkov (2008) finds no evidence that women underperform relative to men in a competitive environment.[59]

Much in line with these latter results, Dreber et al. (2011) find no difference in performance under competitive pressure of seven to ten year old boys and girls in Sweden. Following the initial research conducted with nine to ten year old school children in Israel by Gneezy and Rustichini (2004)[60], Dreber et al. (2011) replicated the field experiment with children from eleven primary schools in the Stockholm area. In addition to competing in short distance races (60m), the children engaged in rope skipping and modern dancing, two tasks that were deliberately chosen as they are perceived as rather advantageous for girls. In the first round of the experiment the children had to perform the respective task separately with no direct competition. In the second round the children were matched in pairs according to their performance in the first round. As expected, the boys on average ran faster while the girls performed better in rope skipping and dancing. However, no differences in the performance under competitive pressure were found between boys and girls. A possible explanation for the contradictory results of both studies is that children in rather male-dominated societies (e.g. Israel) experience a different socialization and nurture than children in a more gender-equal society (e.g. Sweden).[61] In a similar vein, Cárdenas et al. (2012) examine gender differences in competitiveness and risk taking among children aged 9-12 in Colombia and Sweden, two countries ranked 55th and 4th in terms of gender equality according to various macro-economic indices. Surprisingly, boys and girls proved to be equally competitive in all tasks in Colombia whereas the results for Sweden were mixed.

[59] Corroborating this result, Cotton et al. (2010) find that even in a typically male-dominated task (i.e. a multiple-stage math tournament) the initial underperformance of women disappears in the later stages of the competition. In a more recent study Shurchkov (2012) examines two task stereotypes (math exercises vs. verbal tasks) under different competitive regimes and time constraints and finds that men outperform women in a high-pressure math-based tournament while women prevail in a low-pressure verbal task.

[60] The children had to run a track of 40 meters twice, first alone and then matched with a child of equal ability. While boys increased their effort (i.e., they ran faster) in the second (competitive) stage, the girls' performance was unaffected by the setting.

[61] Following a meta-analysis of 109 studies using performance data of the so-called "20 meters shuttle run test" (performed in 37 countries between 1981 and 2003, including more than 400,000 children), Cazorla et al. (2006) come to a different result. They find that the differences between the sexes are consistent across a large number of countries with different social, political and economic systems, hence supporting the "biology-and-predispositions-hypothesis". However, unlike in the two afore-mentioned studies the children were not matched in pairs but all ran alone, i.e. the observed results do not allow any conclusions with respect to gender differences in competitive environments.

However, girls are more risk averse in both countries, with a smaller gender gap in Sweden. Finally, Ivanova-Stenzel and Kübler (2011) examine whether gender differences in competitiveness depend on the gender composition of teams by introducing a real-effort task with wages either based on the teams' absolute performance or on the teams' relative performance in a competitive environment. Their results suggest that relative to a single-sex composition gender diversity decreases gender differences in a competitive setting while the gap increases in a non-competitive environment.

Another stylized fact emerging from a number of experimental studies is that women shy away from competition. Niederle and Vesterlund (2007) find that when having the choice between performing a real-effort task in a competitive (tournament) versus a non-competitive setting (piece rate scheme) twice as many men as women self-select into the tournament (with similar findings see e.g. Kamas and Preston 2009, Vandegrift and Yavas 2009, Gupta et al. 2011). Since the performances of men and women in the two different settings are similar, men are apparently overconfident and enter tournaments too often while equally able women frequently shy away from competition.[62]

In an attempt to explore whether the observed gender differences are innate or caused by society ("nature versus nurture"), a number of experimental studies use data from natural settings. Examining the competitive choices of girls from single-sex and coeducational schools, Booth and Nolen (2012) find that girls from single-sex schools behave more like boys when being randomly assigned to gender-diverse experimental groups. Drawing on the distinctly different circumstances in matrilineal and patriarchal societies, a number of studies have looked at gender differences in competitiveness within these idiosyncratic social environments. Gneezy et al. (2009) examine the competitiveness of men and women in two profoundly different societies, the Massai, a patriarchal society in Tanzania, and the Khasi, a matrilineal society in Northeast India. They find that in the patriarchal Massai

[62] See also Balafoutas and Sutter (2010), Cason et al. (2010), Sutter and Rützler (2010), Wozniak et al. (2010), Healy and Pate (2011), Niederle and Vesterlund (2011), Price (2012), Niederle et al. (2013) and, using data from professional tennis players, Wozniak (2012). A noteworthy exception is Price (2010) who uses a similar experimental design but is unable to find gender differences in competition aversion. This is mainly due to the fact that the participants in this study did not display any gender differences in confidence which is considered more important in explaining gender differences in competitiveness than differences in risk preferences (as examined in numerous experimental studies; see e.g. Charness and Gneezy 2012, Garcia-Gallego et al. 2012, Halko et al. 2012, and, for a summary, Croson and Gneezy 2009). A good example for men's overconfidence is presented by Reuben et al. (2012). In a two-stage real effort task, groups select a leader to compete against other group leaders. It is found that men are selected significantly more often as leaders than is suggested by their individual performance in the first stage and that this effect is mainly driven by men's overconfidence.

society the gender gap is similar to that in Western societies. However, the gender gap is reversed in the matrilineal society, i.e. here women are found to be more competitive than men.[63]

4.2.3 SPORTS DATA

Sports data have been found to be particularly well suited to explore gender differences in competitive behavior (Frick 2011a, 2011b). First, professional athletes represent a highly self-selected sample of persons with a competitive motivation, enabling researchers to analyze sex differences among homogeneous individuals with a distinct professional attitude. Second, since the contestants have very specific ex ante information about prize structures and their opponents' abilities, the self-selection process as well as the incentive effects of tournaments can be examined in detail. Moreover, the high degree of transparency of the athletes' performance and capabilities is likely to reduce gender differences in overconfidence, which, in turn, might lead to a reduction in the gender gap in competitiveness.

Garratt et al. (2011) look at the self-selection process of male and female runners in the "State Street Mile", a running event that offers its participants the choice between entering a less competitive (non-incentivized) race and a more competitive one including prizes. The authors find that qualified women and older runners are less likely to self-select into the competitive race than qualified young men. Only the fastest younger women, being aware of their abilities, do not shy away from competition as they always enter the more competitive race. These results are in line with Nekby et al. (2008) who find that in an elite 10,000-meter race in Sweden women are at least as likely as men to self-select into starting groups with average running times that are beyond their current physical abilities, suggesting that within the latter groups, overconfidence is equally likely for men and women. Frick (1998) examines how male and female professional marathon runners respond to changes in prize money levels and structures. It appears that women respond more to an increase of the total purse as well as to changes in its distribution while their performance (i.e. their finishing time) is unaffected by bonus payments. These behavioral differences can be explained by the fact that the female marathon elite at that time (in the 1980s and early 1990s) was more heterogeneous (i.e., less balanced) than the male elite. Due to that heterogeneity and an equal number of lucrative races for men and women, the top female

[63] Andersen et al. (2013) compare the competitiveness of children in patriarchal and matrilineal societies and show that the gender differences in competitiveness start to evolve around puberty, with the more pronounced changes occurring in the patriarchal society. For additional studies using data from matrilineal and patriarchal societies see e.g. Andersen et al. (2008), Hoffman et al. (2011) and Gong and Yang (2012).

athletes were initially able to avoid competing against each other by strategically entering certain events only. Thus, it was possible for a woman, but not for a man (who, due to the greater homogeneity in the field, always faced competitors of similar strength), to win a marathon with a "suboptimal" performance (i.e., a finishing time that was well above the world record). More recently, the gender gap in competitiveness has significantly narrowed especially in long-distance and ultra-marathon running (see e.g. Frick 2011a, 2011b). Further evidence from road races is presented by Lynch and Zax (2000), Maloney and McCormick (2000) and Frick and Prinz (2007), with the first study questioning the incentive effects of prizes in tournaments (and instead attributing faster finishing times to sorting effects).

In a similar approach, several studies explore the response of professional tennis players to "competitive pressure". Paserman (2007) analyzes aggregate set-level data from Grand Slam tournaments and finds that for both men and women the quality of the game deteriorates with increasing stakes. This drop in performance is greater for women in the decisive set, albeit not statistically significant. Yet, when examining point-by-point data, Paserman (2010) finds that women are more likely to produce "unforced errors" at crucial points of the match, while this does not hold true for men. Hence, there is evidence for a statistically significant gender gap in competitiveness under pressure even among highly self-selected professional athletes. Sunde (2009) uses data from the final two rounds of all ATP Master and Grand Slam tournaments in the years between 1990 and 2002 to analyze the behavior of male tennis professionals. He finds that in uneven contests both the favorite and the underdog perform worse than in "balanced" contests, which is in line with the incentive hypothesis. Lallemand et al. (2008) find exactly the opposite result for female tennis professionals, because here a greater heterogeneity of the two contestants results in a larger number of wins of the favorite and a larger number of losses of the underdog, supporting the capability hypothesis.

Employing a large panel dataset including 1.4 million chess games recorded over a period of 11 years, Gerdes and Gränsmark (2010) analyze the behavior of male and female chess experts. Their results suggest that women are more risk averse than men in the sense that they choose more conservative opening strategies, whereas men, especially when facing female opponents, choose more aggressive strategies even though such strategies reduce their winning probability. This, in turn, is indicative of men being overconfident. Using the same dataset, Gränsmark (2012) finds that women perform worse under time pressure.

Unlike male chess experts, who are found to be too impatient (i.e., men play shorter games but at a higher price), women tend to overstrain reflection time at the early stage of the game, leading to time pressure and a performance deterioration later.

Ehrenberg and Bognanno (1990a, 1990b) find that the overall prize money as well as its distribution in professional golf tournaments have a significant impact on player performance. The higher the total purse and the larger the prize differential, the lower are the scores of the individual golfer (note that lower scores are indicative of a better individual performance). In addition, it appears that during the last round a golfer's performance is positively correlated with the marginal returns to effort, i.e. the higher the rewards for improving one's rank, the lower the number of strokes required to finish that particular round. However, replicating that study with comparable data from the Ladies Professional Golf Association Tour (LPGA) in the year 2000, Matthews et al. (2007) are unable to confirm the results for female golf professionals. They show that an increase of the total purse leads to higher scores and thus a weaker performance. A possible explanation for the observed gender differences in competitiveness is that women are likely to "choke" (i.e., they succumb to the pressure that comes with large prizes at stake) whereas men respond positively to an increase of prizes by delivering a better performance.

Summarizing the literature review, it can be stated that the majority of studies find persistent gender differences in risk-taking, overconfidence, willingness to compete and the like. Most of the sports economics literature thereby focuses on highly selected top athletes vying for monetary incentives. This study contributes to this strand of literature by empirically investigating the gender responsiveness to competition among less talented and rather intrinsically motivated individuals. In the following, the datasets and methodology are described and some (preliminary) descriptive results presented.

4.3 DATA, METHODOLOGY AND DESCRIPTIVE EVIDENCE

The data come from three prestigious long-distance races, namely the "Swiss Alpine Marathon", a mountain ultra-marathon in Davos, Switzerland, the "Ironman Hawaii", the official long-distance triathlon World Championship, and the "Vasaloppet", a long-distance cross-country ski race in Sweden.[64] These competitions take place annually and include a

[64] Race results and background information regarding the competitions can be obtained from the websites www.swissalpine.ch, www.ironman.com, www.ironmanworldchampionship.com, www.vasaloppet.se and www.worldloppet.com.

large number of participants, offering several advantages with regard to the empirical analyses. First, in all races, male and female athletes compete at the same time and under identical conditions. This ensures that the performances of men and women can be compared in a natural and mostly undistorted competitive environment.[65] Second, the finishing times of all participants are available for an uninterrupted period of almost one decade (2002-2010), allowing to observe changes over time. Third, today the prize money levels and distributions are identical for men and women.[66] Thus, the "returns to winning" are the same for male and female contestants, ruling out potential earnings-related selection effects. Unlike in e.g. professional tennis, where it is not possible to accurately measure differences in competitiveness between male and female players (simply because they do not compete against each other in official individual matches), the data used in this study allow observing and quantifying gender differences in competitiveness over time.

The Races

The "Swiss Alpine Marathon" is a mountain ultra-marathon which covers a distance of 78 kilometers and a total altitude change of more than 2,200 meters. Starting and ending in Davos, Switzerland, at an altitude of 1,538 meters, a part of 21 kilometers of the run leads through high alpine terrain with the highest point at 2,632 meters above sea level. The first run was organized in 1986 and has taken place annually since then. Prize money levels are available following the year 1995 and can be accessed on www.arrs.net, a website that is operated by a group of (former) competitive long-distance runners who are at the same time devoted statisticians. Although the (inflation-adjusted) total purse increased by more than 50 percent from 2002 to 2010, only the first 3 male and female finishers receive comparatively small (and equal) monetary rewards of currently CHF 4,000, 2,000 and 1,000 respectively. With more than 1,000 runners participating every year, monetary incentives should thus have no effect for the majority of the contestants whose expected *monetary* utility is fairly close to zero.

[65] One objection that might be raised against this argument is that in mixed contests, although men and women compete independently of each other in their respective class, it cannot be ruled out that women's motivations and performances are, at least to some extent, affected by the performance of male contestants, and vice versa. On the other hand, potential externalities such as weather conditions, which can strongly influence the performance of athletes in long-distance races, are equal for all contestants. In fact, over the years, the finishing times vary significantly (and in a similar pattern for men and women) in both races which can, for the most part, be attributed to changing conditions in the weather, surface, etc.

[66] Recently, a number of sports have either reduced or even abolished the "gender pay gap" in the sense that prize money levels of men and women have been (completely) adjusted. The presumably most prominent example is professional tennis, where male and female players nowadays compete for equal prizes in all four "Grand Slam" tournaments.

The "Ironman Hawaii" is the oldest and most famous long-distance triathlon in the world and represents the official annual Ironman World Championship. The first Ironman took place on Oahu, Hawaii, in 1978. In 1981, the race was moved to Big Island, Hawaii, with the start and finish in Kailua-Kona. The "Ironman Hawaii" is considered one of the toughest long-distance races in the world. Athletes have to swim a distance of 3.86 kilometers, cycle 180.2 kilometers and run a full marathon of 42.195 kilometers. Besides the extraordinarily long distance, contestants face temperatures of sometimes more than 100 degrees (40 degrees Celsius) and crosswinds of up to 70 km/h, which are an enormous challenge in particular during the cycling part where slipstream riding behind or next to a participant is strictly forbidden (as in every Ironman event). Moreover, every triathlete has to qualify for the "Ironman Hawaii" by meeting the qualification norms in one of the other Ironman events around the world. Thus, contrary to the "Swiss Alpine Marathon", which can be entered by recreational athletes, too, the starting field at the "Ironman Hawaii" consists of highly (self-)selected and rather homogeneous athletes with respect to their abilities and (competitive) motivations. Women participate since 1979, although the number of female starters was very small in the early years[67] and only increased after the equalization of prize money levels for men and women in 1986. The "Ironman Hawaii" offers a comparatively large (and growing) purse of more than $ 500,000 which is equally distributed among the top ten male and female contestants, with both the male and female champion receiving more than $ 100,000 each.

The "Vasaloppet" (literally: Vasa race) is an annual 90 km cross-country ski race that takes place in Sweden. The first race was held in 1922, and today, every year more than 10,000 skiers participate in the oldest, longest and most reputable cross-country ski race in the world. Notably, women were not allowed to enter the race between 1924 and 1980 and did not receive any awards from 1981 to 1997. The ban was introduced because the stresses and strains associated with the race were considered bad for women's health. Later, it was rather the concern that female participants would damage the event's reputation as a tough challenge that upheld the ban. Since 1998, the "Vasaloppet" includes both a men's and a women's classification, with the mixed field starting at the same time. Unlike the "Swiss Alpine Marathon" and the "Ironman Hawaii", no prize money is awarded in the "Vasaloppet".

[67] In fact, Lyn Lemaire was the first *and only* woman in the 1979 "Ironman Hawaii", thus becoming the first female champion practically "by default".

In all three events, top athletes can expect (additional) income through lucrative endorsement contracts. In particular long-distance triathletes are wooed by sport equipment suppliers as they can credibly commit to the use of high quality equipment in different disciplines. However, economic incentives in the form of on-site prize money and/ or endorsement deals should only matter for the top athletes. Moreover, the prize money levels and distributions for men and women in the "Swiss Alpine Marathon" and the "Ironman Hawaii" had been equalized some time before the start of the observation period, hence offering no variation in the covariate (monetary incentives). For an illustration of the structure of the datasets see Table 4-1:

Table 4-1: Structure and composition of the datasets

Competition	Coverage	Participants per event/year		Number of observations	
		Male	Female	Total	Comparable
Mountain Ultra-Marathon "Swiss Alpine Marathon"	2002 - 2010	800 - 1,200	100 - 250	9,000	900
Long-Distance Triathlon "Ironman Hawaii"	2002 - 2010	1,100 - 1,300	250 - 470	16,000	2,250
90km Cross-Country Skiing "Vasaloppet"	2002 - 2010	10,000 - 13,000	800 - 1,600	120,000	7,200

Since the number of male participants is considerably higher than the number of female participants in all three events, the "total" number of observations has to be reduced to a "comparable" number of observations. That is, the minimum number of female finishers in a given event over the whole period of observation marks the maximum number of male and female participants whose finish times can be compared. This leaves us with 100 comparable observations (i.e., the finish times of 100 male and female athletes each) in the "Swiss Alpine Marathon" per year, 250 annual observations for the "Ironman Hawaii" and 800 annual observations for the "Vasaloppet".

In order to compare the performances of men and women, the individual finish time of each athlete is used. In the next step, the difference in finish times between male and female athletes who achieved the same rank is calculated. Following Frick (2011a, 2011b), I calculate not the absolute but the percentage time difference to ensure comparability across the different race types and distances.

$$PD_{ij} = ((FFT_{ij} - MFT_{ij}) / MFT_{ij}) * 100 \quad (1)$$

where PD_{ij}: percentage difference in finish time between male and female contestant on rank i in race j.

MFT_{ij}: male finish time on rank *i* in race *j*.

FFT_{ij}: female finish time on rank *i* in race *j*.

Table 4-2 below reveals that the percentage difference in finish time between men and women finishing on the same rank increases with rank. This, in turn, suggests that men are more competitive than women. The estimations in section 4.4 support the general assumption that men show a higher degree of competitiveness than women (i.e., men's races are more "balanced" than women's races). This is in line with Frick (2011a, 2011b), Frick and Scheel (2013) and most of the literature discussed above.

Table 4-2: An illustration of the structure of the samples

Example: Long-Distance Triathlon, "Ironman Hawaii".

Source: http://ironman.com, http://ironmanworldchampionship.com and own calculations.

Year	Rank	Male Athlete	Time	Time Difference (in %)	Female Athlete	Time	No. of finishers	Proportion of Women (in %)
2002	1	T. Deboom	8:29:56	7.45	N. Badman	9:07:54	1455	23.44
	2	P. Reid	8:33:06	8.05	N. Kraft	9:14:24	1455	23.44
	3	C. Brown	8:35:34	9.09	L. Bowden	9:22:27	1455	23.44
	…							
	248	J. Olsen	10:13:37	29.58	N. Rotovnik	13:15:07	1455	23.44
	249	M. Weijers	10:13:42	29.62	G. Osterlund	13:15:30	1455	23.44
	250	P. Mahon	10:13:51	29.60	J. Guetz	13:15:33	1455	23.44
…								
2010	1	C. McCormack	8:10:37	9.78	M. Carfrae	8:58:36	1770	26.50
	2	A. Raelert	8:12:17	10.91	C. Steffen	9:06:00	1770	26.50
	3	M. Vanhoenack.	8:13:14	11.52	J. Dibens	9:10:04	1770	26.50
	…							
	248	A. Gordon	9:44:05	19.54	T. Nishikawa	11:38:13	1770	26.50
	249	K. Glah	9:44:15	19.53	R. Wilson	11:38:22	1770	26.50
	250	E. Barbet	9:44:16	19.54	J. Smalec	11:38:25	1770	26.50

In order to ensure comparability across years and to control for the robustness of the results, I introduce two additional variables: the proportion of women among all finishers

(%FEM)[68] and the total number of finishers in a given year and event (ΣFIN). While the importance of the control for women's participation is obvious, the following thought experiment should help clarifying as to why the inclusion of the total number of finishers is necessary.

Suppose the relative share of women in two consecutive years remains stable at 20 percent but the overall number of athletes increases from 1,400 to 1,700. Then the absolute number of female athletes would increase from 250 to 340, implying that the less talented women finishing in the lowest quintile of the estimation (i.e., between the ranks 201 and 250) clearly differ in their characteristics and (relative) sporting performance: whereas the slowest observed women in the estimation in fact represent the lowest quantiles of the overall population of female athletes in the first year, this does not hold true for the second year where a considerable share of the least talented female athletes is not included in the estimation.

A very straightforward approach to tackle this problem would be to compare the performances of men and women finishing, for instance, in the same deciles. However, there is a certain time limit in every race that predetermines the finish time of the slowest – but still fast enough to escape the "broom wagon" – participants. Due to this ex ante fixed lower boundary, the percentage time differences would decrease towards the lower quantiles and even disappear in the "slowest" decile, representing an obvious bias. For the reasons just mentioned it seems that the above control variables most appropriately capture variations in the composition of the starting field. Summary statistics for the datasets of the "Swiss Alpine Marathon", the "Ironman Hawaii" as well as the "Vasaloppet" are displayed separately in the following three tables.

[68] Expressed as (F/N*100), where "F" is the number of female finishers and "N" is the number of total finishers.

Table 4-3: Summary statistics "Swiss Alpine Marathon"

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLE						
PD	Percentage difference in finish time between female and male contestant on rank i in race j	900	27.90	6.11	6.88	46.69
INDEPENDENT VARIABLES						
Rank	Position of runner i in race j	900	50.5	-	1	100
Rank²	Square value of rank	900	3,383	-	1	10,000
Rank³	Cubic value of rank	900	255,025	-	1	1,000,000
TT	Linear time trend (2002 = 1, … 2010 = 9)	900	5	-	1	9
%FEM	Proportion of female finishers among all finishers in race j	900	13.06	1.53	11.48	16.88
ΣFIN	Total number of finishers in race j	900	1,024	199.5	839	1,487

On average, about 1,000 individuals (of whom roughly 130 are female) finish the "Swiss Alpine Marathon" every year. It should be noted that both the relative share of female athletes (%FEM) and the total number of finishers in a given race (ΣFIN) increase over time in an approximately linear way.

Table 4-4: Summary statistics "Ironman Hawaii"

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLE						
PD	Percentage difference in finish time between female and male contestant on rank i in race j	2,250	17.29	3.29	6.73	29.66
INDEPENDENT VARIABLES						
Rank	Position of runner i in race j	2,250	125.5	-	1	250
Rank²	Square value of rank	2,250	20,959	-	1	62,500
Rank³	Cubic value of rank	2,250	3,937,563	-	1	1.56e+07
TT	Linear time trend (2002 = 1, … 2010 = 9)	2,250	5	-	1	9
%FEM	Proportion of female finishers among all finishers in race j	2,250	25.91	1.53	23.44	27.95
ΣFIN	Total number of finishers in race j	2,250	1,626	83.80	1,455	1,770

Despite the restrictive qualification norms, on average, more than 1,600 triathletes participate (and successfully complete) the "Ironman Hawaii" every year. The share of female athletes is significantly larger than in the other two competitions and ranges around 26 per-

cent. Quite remarkably, performance differences between men and women who finish on the same rank seem to be comparatively small, even among less talented individuals. The empirical analyses – and in particular the discussion in section 4.4.2 – will shed further light on this issue.

Table 4-5: Summary statistics "Vasaloppet"

Variable	Operationalization	# of Obs.	Mean	SD	Min	Max
DEPENDENT VARIABLE						
PD	Percentage difference in finish time between female and male contestant on rank i in race j	7,200	67.07	18.99	1.7	109.5
INDEPENDENT VARIABLES						
Rank	Position of runner i in race j	7,200	400.5	-	1	800
Rank²	Square value of rank	7,200	213,733.5	-	1	640,000
Rank³	Cubic value of rank	7,200	1.28e+08	-	1	5.12e+08
TT	Linear time trend (2002 = 1, … 2010 = 9)	7,200	5	-	1	9
%FEM	Proportion of female finishers among all finishers in race j	7,200	8.80	1.04	7.7	11.06
ΣFIN	Total number of finishers in race j	7,200	13,146	1,328	11,028	14,947

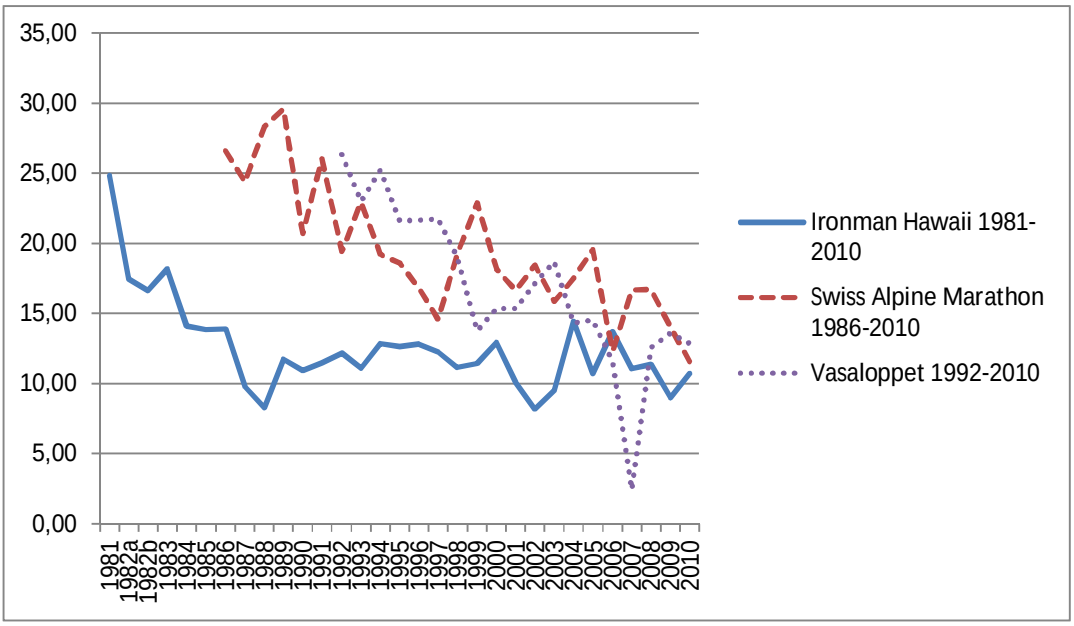
The "Vasaloppet" differs considerably from the other two races because, firstly, it is a mass participation event attracting up to 15,000 – predominantly recreational – athletes per year. Secondly, long-distance cross-country skiing appears to be a rather male-dominated domain, with an average proportion of female finishers of less than 10 percent. Along the same lines, gender differences in competitiveness seem to be more pronounced in the "Vasaloppet", as expressed by larger mean and maximum values for *PD*.[69] And, thirdly, athletes are less diverse with respect to their country of origin: About 75 percent of the athletes are from Sweden, while the majority of the remaining 25 percent come from other Scandinavian countries (see Figure 4-3 in Appendix C).

Apart from the aforementioned (and anticipated) result that men are, on average, more competitive than women – as expressed by the increasing percentage time difference towards the lower ranks – a question of particular interest is whether the observed gender differences change over time. Figure 4-1 below includes data on the top 3 male and female

[69] These statistics might however be "driven" by the large number of observations in the Vasaloppet in comparison to the other races.

finishers in the "Swiss Alpine Marathon", "Ironman Hawaii" and "Vasaloppet", covering finish times over a period of 25, 30 and 19 years, respectively.[70]

Figure 4-1: Performance gap development between top 3 male and female finishers over time



Note: The y-axis shows the average percentage time difference between the top 3 male and female athletes in a given race.

It appears from Figure 4-1 that the gender gap in competitiveness has decreased over time (i.e., the male elite and the female elite, consisting of the top 3 finishers in every race in every year, have converged with respect to their performances). This convergence is most pronounced at the beginning of the observation period and seems to be asymptotically bounded from below by a value of about 10 percent. These (preliminary and descriptive) findings are in line with Thibault et al. (2010) who find that after 1983 the gender gap in world records has stabilized at a mean difference of 10.0 percent (± 2.94) for all Olympic swimming, track-and-field, cycling and speed skating disciplines (with the latter exhibiting the lowest gender gap of 7.0 percent).

4.4 EMPIRICAL RESULTS

The descriptive results derived from Figure 4-1 above are based on a very limited number of observations, covering only the top 3 men and women in every race. The empirical

[70] Comparable data for the earlier years were only available for the medal ranks. The following estimations in section 4.4 include data on all finishers for the years 2002 to 2010.

analyses in this section include data on a considerably larger number of male and female finishers and aim to explore whether the "overall" gender gap in competitiveness (i.e., the performance differences not only of the fastest men and women but of a representative share of the respective starting fields) exhibits statistically significant changes over time. To uncover these changes, the following regression model is estimated[71]:

$$PD_{ij} = \alpha_0 + \alpha_1 RANK_{ij} + \alpha_2 RANK^2_{ij} + \alpha_3 RANK^3_{ij} + \alpha_4 TT + \alpha_5 \%FEM_j + \alpha_6 \Sigma FIN_j + \varepsilon \quad (2)$$

where PD_{ij}: percentage difference in finish time between female and male contestant on rank *i* in race *j*.

RANK: position of runner *i* in race *j*.

TT: linear time trend (2002 = 1, … 2010 = 9).

$\%FEM_j$: proportion of female finishers among all finishers in race *j*.

ΣFIN_j: total number of finishers in race *j*.

ε: unexplained random error.

4.4.1 RESULTS "SWISS ALPINE MARATHON"

In the following, the estimation results of the "Swiss Alpine Marathon" are presented. The first three model specifications include the top 100 male and female finishers of each race and vary with regard to the control variables included in the estimation. In model specifications 4 and 5, the top 10 and top 25 finishers, respectively, are excluded from the estimation, for reasons that will be explained later in this section.

Table 4-6 supports the descriptive evidence presented in the previous section, showing that the pool of male ultra-marathon runners is far more homogeneous than the pool of female ultra-marathoners. The statistically significant and negative (positive) coefficients of the squared (cubic) term of RANK suggest that the percentage time difference between male and female runners on the same rank is increasing with rank, first, at a declining and, in the higher ranks (i.e., among the less talented athletes), at an accelerating rate. This is in line with the results reported by Frick (2011b) who identifies a similar pattern for male and female ultra-marathon runners. What is salient in this context is that the coefficient of the linear time trend is negative and statistically highly significant even if the control variables

[71] Since the distribution of the dependent variable PD_{ij} is significantly different from the standard normal distribution, while the exact theoretical distribution is unknown, all regressions are bootstrapped (Efron 1979) with 200 replications to ensure robustness of standard errors.

"proportion of female finishers" and "total number of finishers" are included (model 3). Thus, there is robust evidence for a decrease of the gender gap in competitiveness over time (i.e., the performance differences of male and female ultra-marathoners competing in the "Swiss Alpine Marathon" have decreased in the years 2002 to 2010). As a further robustness check, the model is estimated with year dummies instead of a linear time trend. However, I am unable to identify any statistical breaks or outliers. In fact, the inclusion of year dummies corroborates the assumption that the decrease of the gender gap follows a linear trend.[72]

Table 4-6: Estimation results: "Swiss Alpine Marathon", mountain ultra-marathon

(OLS, bootstrapped standard errors with 200 replications)

	Model 1	Model 2	Model 3	Model 4	Model 5
	Top 100 finishers included			Top 10 finishers excluded	Top 25 finishers excluded
Covariates	Dependent Variable: PD_{ij}				
RANK	0.471***	0.471***	0.471***	0.412***	0.537***
	(12.71)	(13.19)	(15.19)	(5.95)	(4.19)
$RANK^2$ #	-0.065***	-0.065***	-0.065***	-0.054***	-0.075***
	(-7.79)	(-8.18)	(-9.55)	(-4.24)	(-3.46)
$RANK^3$ ##	0.039***	0.039***	0.039***	0.033***	0.044***
	(7.10)	(7.61)	(8.81)	(4.53)	(3.78)
TT	-0.827***	-0.406***	-0.609***	-0.741***	-1.005***
	(-20.12)	(-7.34)	(-9.77)	(-11.33)	(-17.39)
%FEM		-0.891***	-0.060	0.185	0.654***
		(-10.30)	(-0.40)	(1.43)	(6.39)
∑FIN			-0.005***	-0.007***	-0.010***
			(-7.64)	(-11.11)	(-19.78)
CONST	20.36***	29.89***	25.57***	25.38***	21.35***
	(38.65)	(26.73)	(19.37)	(16.22)	(8.59)
Observations	900	900	900	810	675
Wald Chi^2	2,510***	2,873***	2,750***	2,394***	2,363***
Adj. R^2	0.789	0.808	0.816	0.790	0.812

multiplied by 10 for ease of presentation

multiplied by 1,000 for ease of presentation

z statistics in parentheses, *** p<0.01.

[72] Results of these estimations are available from the author upon request.

Given the development of the percentage time differences between the top 3 male and female runners over the last 25 years (as illustrated earlier in Figure 4-1), which seem to have decreased asymptotically (leveling at a performance gap of about 10 percent), the changes are likely to have occurred in the lower ranks. Therefore, it appears worthwhile to exclude the top finishers from the analysis. This has been done in models 4 and 5 which exclude the top 10 and top 25 male and female finishers of each race. By focusing on less talented individuals (who have little prospect of reaching the "money ranks"), it is possible to rule out – or at least reduce – potential effort or selection effects that are induced by monetary incentives.

The estimation results of models 4 and 5 suggest that the gender gap has decreased over time, irrespective of monetary incentives. Closer inspection of the "time trend" coefficient reveals that the reduction of the gender gap is even more pronounced among lower ranked individuals. More precisely, the gap between male and female ultra-marathoners finishing between ranks 26 and 100 has decreased, on average, by one percentage point per year. That is, the 50^{th} ranked man who needed approximately 7.5 hours to finish the race in 2002, and who was more than 2.5 hours faster than the 50^{th} ranked woman in that year, was ceteris paribus less than 2 hours faster than the 50^{th} ranked woman in 2010. Hence, over a period of only 9 years, the gender gap in this particular event appears to have narrowed by about 9 percentage points, leading to a dramatic decrease of absolute performance differences between men and women.

This, in turn, implies that not only the quantity but also the quality of female runners has increased and that women's races have become more "balanced". For example, whereas the relative performance gap between the fastest and the 50^{th} ranked female athlete was 47.5 percent in 2002, the gap was reduced to 38.9 percent in 2010. The "performance dispersion" in men's races, on the other hand, remained largely unchanged, with a relative performance gap between rank 1 and rank 50 of 30.6 (28.8) percent in 2002 (2010). Possible explanations for the observed quality improvements among female runners are that (i) more "able" women select into the races and that (ii) women increased their training workload and/ or exert more effort during the race. Disentangling these effects is rather intricate and requires additional athlete-specific information. Qualitative interviews could shed additional light on individuals' (competitive) motivation. The above explanatory approaches will be further elaborated in section 4.4.3. In the following, the results of the "Ironman Hawaii" are presented.

4.4.2 RESULTS "IRONMAN HAWAII"

An inspection of the estimation results of the "Ironman Hawaii" reveals that the coefficients have the same direction as the ones presented in Table 4-6, suggesting that (i) male long-distance triathletes, too, are more competitive than female long-distance triathletes, albeit (ii) these differences are diminishing over time.

Table 4-7: Estimation results: "Ironman Hawaii", long-distance triathlon

(OLS, bootstrapped standard errors with 200 replications)

	Model 1	Model 2	Model 3	Model 4	Model 5
	Top 100 finishers included			Top 10 finishers excluded	Top 25 finishers excluded
Covariates	Dependent Variable: PD_{ij}				
RANK	0.471***	0.471***	0.471***	0.412***	0.537***
	(12.71)	(13.19)	(15.19)	(5.95)	(4.19)
$RANK^2$ #	-0.065***	-0.065***	-0.065***	-0.054***	-0.075***
	(-7.79)	(-8.18)	(-9.55)	(-4.24)	(-3.46)
$RANK^3$ ##	0.039***	0.039***	0.039***	0.033***	0.044***
	(7.10)	(7.61)	(8.81)	(4.53)	(3.78)
TT	-0.827***	-0.406***	-0.609***	-0.741***	-1.005***
	(-20.12)	(-7.34)	(-9.77)	(-11.33)	(-17.39)
%FEM		-0.891***	-0.060	0.185	0.654***
		(-10.30)	(-0.40)	(1.43)	(6.39)
∑FIN			-0.005***	-0.007***	-0.010***
			(-7.64)	(-11.11)	(-19.78)
CONST	20.36***	29.89***	25.57***	25.38***	21.35***
	(38.65)	(26.73)	(19.37)	(16.22)	(8.59)
Observations	900	900	900	810	675
Wald Chi^2	2,510***	2,873***	2,750***	2,394***	2,363***
Adj. R^2	0.789	0.808	0.816	0.790	0.812

multiplied by 10 for ease of presentation

multiplied by 1,000 for ease of presentation

z statistics in parentheses, *** p<0.01.

Not surprisingly, the RANK coefficient has a considerably smaller magnitude than the RANK coefficient displayed in the "Swiss Alpine Marathon" results table.[73] This implies

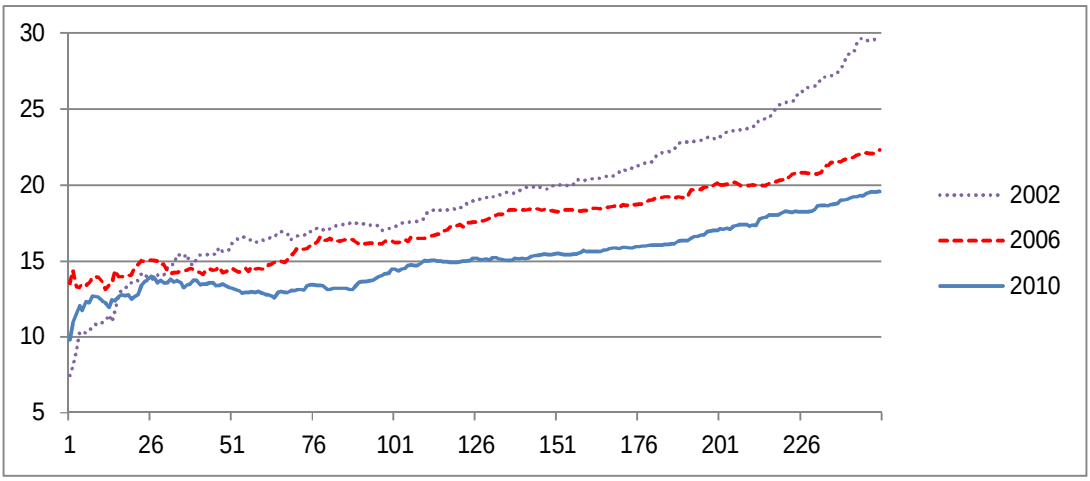
[73] Notably, these differences are robust to a wide range of model specifications and additional robustness checks, including estimations based on a reduced – and comparable – sample of 100 annual observations. The respective estimation results are available upon request.

that men's races are still more "balanced" than women's races (i.e., men are more homogeneous in their performance than women) but to a lesser extent in the "Ironman Hawaii" than in the "Swiss Alpine Marathon". I am tempted to attribute this finding to the highly selective entry criteria (i.e, the adherence to strict qualification norms) in the "Ironman Hawaii" that ensure participation of male and female athletes who are less diverse with regard to their abilities, aspirations and competitive motivations.

Again, the results are robust to different model specifications, including the estimations based on the reduced samples (see models 4 and 5). This, in turn, implies that monetary incentives alone cannot explain the narrowing of the gender gap, as the exclusion of those athletes vying for the rewards leaves the results largely unchanged.

Figure 4-2 below captures the main results emerging from the analysis: First, the development of the percentage time difference among the top male and female athletes does evidently not follow any particular (linear) time trend. In line with the descriptive evidence provided in Figure 4-1, the fastest men are about 10 percent faster than the fastest women, with some variation across years. Second, the gender gap in performance increases towards the lower ranks (i.e., among the less talented individuals), suggesting that the performance dispersion is significantly smaller in men's races than in women's races. In other words, men's races (still) seem to be more competitive than women's races. It appears, however, that the gender gap has decreased considerably throughout the last decade. For example, in 2002, the 250th ranked woman was about 30 percent slower than the man on the same rank. In 2010, the time difference on the same rank was less than 20 percent. Given the finish time at this particular point of the distribution of about 10 hours for male long-distance triathletes, it can be noted that the absolute time difference between a man and a woman on rank 250 was reduced by approximately one hour.

Figure 4-2: Development of the percentage differences in finishing times between male and female athletes by tank ("Ironman Hawaii", 2002-2010)



Note: The y-axis shows the percentage time difference between a male and a female triathlete on the same rank, while the ranks (1-250) are indicated on the x-axis. Observations of interjacent years are omitted for a better illustration of the results and are available from the author upon request. Standard errors are not included here; however, the existence of statistically significant differences between years is corroborated by the empirical analysis. Source: www.ironman.com, www.ironmanworldchampionship.com and own calculations.

Admittedly, Figure 4-2 leads us to suspect a cubic relationship between *RANK* and the percentage time difference for the year 2002, whereas an almost linear relationship can be observed for the later years. Therefore, a number of robustness checks were performed. However, neither the inclusion of year dummies instead of a linear time trend, nor the exclusion of the 2002 observations from the initial model did affect the results.[74] This supports the general assumption that women have become *continuously* more competitive over time.

4.4.3 RESULTS "VASALOPPET"

In the following, I report the estimation results of the "Vasaloppet". Model specifications 1-5 are based on a comparatively large sample including the top 800 male and female finishers of each race (although the top finishers are excluded in models 4 and 5). For a better comparison with the results in the previous sections, the model is also estimated on the basis of a reduced sample (n = 250, model specifications 6-8).

[74] Results tables of the robustness checks are available from the author upon request.

Table 4-8: Estimation results: "Vasaloppet", long-distance cross-country skiing

(OLS, bootstrapped standard errors with 200 replications)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
	Top 800 finishers included			Top 10 finishers excluded	Top 25 finishers excluded	Top 250 finishers included	Top 10 finishers excluded	Top 25 finishers excluded
Covariates	Dependent Variable: PD_{ij}							
RANK	0.195***	0.195***	0.195***	0.177***	0.167***	0.426***	0.316***	0.265***
	(53.94)	(57.32)	(54.11)	(52.70)	(42.45)	(29.29)	(21.56)	(15.34)
RANK² #	-0.003***	-0.003***	-0.003***	-0.002***	-0.002***	-0.022***	-0.013***	-0.010***
	(-23.50)	(-28.75)	(-27.44)	(-23.34)	(-18.87)	(-17.83)	(-11.29)	(-7.22)
RANK³ ##	0.0002***	0.0002***	0.0002***	0.0001***	0.0001***	0.005***	0.003***	0.002***
	(15.90)	(21.10)	(20.36)	(16.01)	(13.00)	(14.98)	(9.32)	(5.77)
TT	-0.812***	0.509***	0.823***	0.848***	0.875***	0.103***	0.156***	0.200***
	(-29.81)	(18.39)	(22.63)	(26.53)	(25.14)	(3.02)	(4.86)	(6.31)
%FEM		-5.836***	-6.846***	-6.953***	-7.111***	-1.856***	-2.000***	-2.215***
		(-69.74)	(-59.99)	(-70.53)	(-68.25)	(-18.68)	(-23.31)	(-26.87)
∑FIN			0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
			(11.52)	(12.22)	(11.45)	(11.48)	(14.79)	(13.94)
CONST	29.50***	74.27***	71.12***	73.90***	76.50***	24.47***	29.38***	33.83***
	(91.78)	(96.62)	(82.94)	(83.44)	(84.99)	(21.40)	(29.54)	(30.84)
Observations	7,200	7,200	7,200	7,110	6,975	2,250	2,160	2,025
Wald Chi²	62,569***	61,947***	60,549***	59,830***	63,045***	13,431***	13,370***	14,405***
Adj. R^2	0.852	0.921	0.922	0.921	0.918	0.907	0.899	0.889

multiplied by 10 for ease of presentation

multiplied by 1,000 for ease of presentation

z statistics in parentheses, *** p<0.01.

The empirical evidence emerging from the "Vasaloppet" is somewhat different from the results presented so far. Although the performance dispersion among long-distance cross-country skiers appears to be significantly smaller for men – thus corroborating the aforementioned results – the gender gap seems to have increased over time. It should be mentioned that this result only holds true when controlling for the proportion of female finishers (as well as the overall number of finishers), while it is robust to the exclusion of the top 10 and top 25 finishers (models 4 and 5). Moreover, quantile regression estimates show that changes in the performance gender gap are most pronounced in the lowest tier of the distribution (see Table 4-8 in Appendix C). As a further robustness check – and for reasons of comparison – models 6-8 are estimated with a reduced number of observations. Although changes in the gender gap over time appear to be smaller among more talented ath-

letes (i.e., among the top 250 male and female finishers), men seem to have become increasingly competitive as compared to women. Possible explanations for this rather surprising result will be discussed in the following section.

4.4.4 DISCUSSION OF THE RESULTS

A more or less undisputed finding resulting from the above analyses is that men are, on average, more competitive than women. In other words, due to a supposedly more homogeneous talent pool of male athletes, men's races are more balanced than women's races. Yet, whereas gender differences among top athletes seem to have stabilized over the years (i.e., the fastest men are consistently about 10 percent faster than the fastest women), the observed gender gap in competitiveness appears to have narrowed among the less talented athletes, at least in competitions with culturally heterogeneous contestants.[75]

That is, the majority of (intrinsically motivated) women (who cannot reasonably target the top positions) have become more competitive in recent years. This finding suggests that sex differences in performance depth cannot only be explained by evolved predispositions and physical differences between men and women, as contended by a number of medical experts and sports scientists (see e.g. Cheuvront et al. 2005, or Deaner 2006a, 2006b, 2012), but seem to be affected by time constraints and differences in opportunities available in society, too, as argued by Becker (1993).[76]

One still unresolved question is whether the observed relative performance improvements of females are due to selection effects or increased individual effort. Given that the changes predominantly occur in the lower ranks (i.e., among the less talented individuals), effort is not likely stemming from pecuniary incentives. Instead, it is possible that highly selected groups of ambitious ultra-marathoners and long-distance triathletes display an exceptional-

[75] Whereas both the "Swiss Alpine Marathon" and "Ironman Hawaii" attract athletes from all over the world, individuals who are active in the "Vasaloppet" appear to be culturally less diverse (see Figure 4-3 in Appendix C). This might, to some extent, explain the perhaps unexpected increase of the performance gender gap among long-distance cross-county skiers: It is conceivable that saturation and substitution effects impact the results in the sense that women from mostly gender-equal societies reduce investments in traditional and rather male-dominated activities at the expense of more "trendy" pastimes. This, of course, remains to be tested in future research.

[76] What has not been considered in this context is whether there are any age-related changes in the level of competitiveness. This has been done recently for the "Swiss Alpine Marathon" in a study by Eichenberger et al. (2012) who find that the participation of elderly male and female mountain ultra-marathon runners increased across the years while the performance in terms of running times remained unchanged for women and even deteriorated among elderly men. Using data covering 25 years of "Ironman Hawaii" results, Knechtle et al. (2012) find that both the relative participation and performance of males older than 44 years and females older than 40 years increased while the gender difference in total time performance significantly decreased.

ly high level of intrinsic motivation which, in turn, induces them to exert maximum effort during *any* athletic contest, *irrespective* of potential rewards. Coffey and Maloney (2010) describe this as the thrill of victory (TOV) effect. In an unprecedented "naturalistic experiment" the authors examine horse and dog racing to separate TOV from incentive effects while simultaneously accounting for selection effects.

The underlying data of this study do not allow accounting for selection effects. Yet, assuming that (i) monetary incentives have no effect on the performance of amateur athletes finishing well beyond the "money ranks", and that (ii) intrinsically motivated individuals do not systematically vary their effort from year to year, a possible alternative explanation is that due to societal changes a growing number of women have gained access to leisure time. That is, while discrimination of females in accessing leisure time was prevalent in many societies in the past (and continues to be in various parts of the world), nowadays more gender-equal societies enable women to engage in time-consuming leisure activities – including, among others, challenging sport activities – at declining (social) costs. This suggests that both selection effects (more women are able (and willing) to enter competitive races) and TOV effects (women train harder and are thus able to exert more effort during the race) work in the same direction, causing gender differences to diminish. A similar development was observed for female professional mid- and long-distance runners by Frick (2011a): Until the mid-1980s, the female mid- and long-distance running scene was dominated by runners from the US as well as from Eastern and Western Europe. Following a sudden influx of world-class runners from Africa (and here particularly from Ethiopia and Kenya) the "cultural heterogeneity" increased dramatically, associated with a considerable intensification of the competition. A notable difference of Frick's (2011a) study is that pecuniary incentives are likely to have caused much of the observed relative performance improvements of women, whereas TOV (and selection) effects appear to dominate in the present study.

4.5 SUMMARY, LIMITATIONS AND IMPLICATIONS

The empirical evidence presented in this study suggests that gender differences between highly competitive and culturally heterogeneous long-distance athletes have diminished over time.[77] Since the observed changes occur in the lower parts of the rankings and thus

[77] Gender differences between culturally less diverse athletes in a rather male-dominated sport, on the other hand, seem to have slightly increased and might be explained by changing leisure activity preferences among women.

irrespective of pecuniary incentives (which are equal for men and women), it appears that particularly the less talented and rather intrinsically motivated females have made up ground on equally endowed male contestants. That is, while the performance dispersion in men's races remained largely unchanged throughout the last decade, women's races have become increasingly competitive over time, leading to a narrowing of the gender gap among amateur athletes. Unlike the influx of female mid- and long-distance runners from Africa in the mid-1980s, which was rather economically motivated and resulted in a narrowing of the gender gap among the top athletes, the quantitative as well as qualitative increase of (intrinsically motivated) female talent in mass participation events in recent years cannot be explained by financial incentives. Rather, it seems that changing socio-cultural conditions enable more women to invest in time-consuming sports activities at declining (social) costs.

One objection that might be raised against the results is that men and women do not necessarily compete *against* one another. But even if we assume two independent competitions *separated* by gender, men and women still compete at the same time and under equal conditions, allowing to draw inferences about relative performance differentials among men and women. Unfortunately, it is not possible to compare absolute performances across years, because due to changing weather conditions finish times vary considerably from year to year. Even the respective winning times in the "Ironman Hawaii" and "Swiss Alpine Marathon" vary by about thirty minutes for both men and women. For comparison, the variation in winning times in the "New York City Marathon" over the same time period (i.e., 2002-2010) is two and a half minutes for men and about six minutes for women. Therefore, the question of how the best women today would fare against past performances by men cannot be answered with the underlying data.

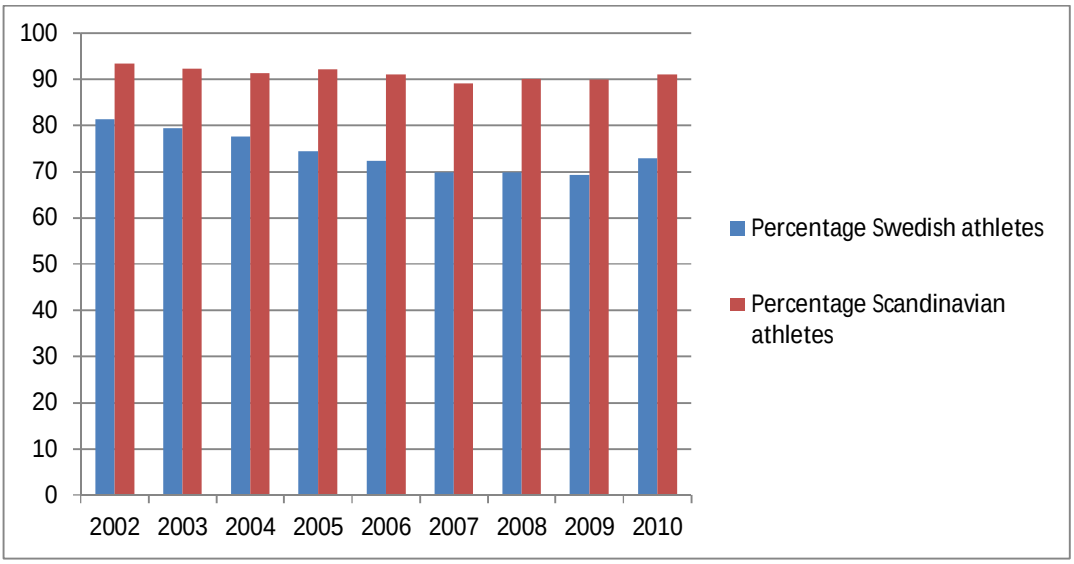
A further limitation is that the study focuses on a (seemingly) non-representative sample of highly selected individuals with very similar aspirations and competitive motivations. Even though these are exactly the same preconditions that hold true for men and women in top positions in management and academia, conclusions about gender responsiveness in the labor market must be drawn very cautiously. As a consequence of decreasing opportunity (or social) costs in increasingly gender-equal societies, for example, it is possible that the "pool" of women who prefer to invest in a professional career (rather than in tasks that are associated with traditional role models) increases in a quantitative as well as qualitative sense, thus reducing gender differences in competitiveness in the general labor market. Of

course, such a development would presuppose transparent and non-discriminatory market mechanisms. Only under these premises is it possible (and economically reasonable) to increase the share of women in top positions – even in the absence of government-imposed measures such as gender quotas and the like.

4.6 APPENDIX C

Figure 4-3 illustrates that the starting field in the Swedish "Vasaloppet" is composed of individuals who are largely homogeneous in terms of their cultural background. About 90 percent of the athletes are from Scandinavian countries which are known to be among the most gender-equal societies in the world. The remaining 10 percent of athletes are predominantly from other European countries with a tradition in winter sports, such as Germany and Austria. Only few athletes (i.e., less than 1 percent) are of non-European origin. Therefore, changes in the level of competitiveness (among men and women) are most likely independent of changes in the "cultural composition" of the starting field.

Figure 4-3: Cultural homogeneity of athletes in the Swedish "Vasaloppet", 2002-2010



Source: www.worldloppet.com and own calculations.

Table 4-9: Estimation results: "Vasaloppet", long-distance cross-country skiing

(Quantile regressions)

Covariates	.10 Quantile	.25 Quantile	.50 Quantile	.75 Quantile	.90 Quantile
	Dependent Variable: PD_{ij}				
RANK	0.232***	0.204***	0.169***	0.172***	0.145***
	(185.13)	(139.00)	(97.99)	(144.34)	(211.21)
$RANK^2$ #	-0.004***	-0.003***	-0.002***	-0.002***	-0.002***
	(-105.14)	(-72.38)	(-39.67)	(-60.73)	(-76.58)
$RANK^3$ ##	0.0002***	0.0002***	0.0001***	0.0001***	0.0001***
	(79.78)	(54.30)	(26.67)	(42.58)	(49.39)
TT	1.581***	0.093***	0.035	1.420***	2.291***
	(159.49)	(5.26)	(1.19)	(58.92)	(146.53)
%FEM	-9.053***	-4.977***	-4.612***	-7.517***	-8.872***
	(-308.24)	(-100.96)	(-54.02)	(-97.16)	(-172.95)
∑FIN	0.003***	0.0002***	-0.001***	0.001***	0.002***
	(143.29)	(7.04)	(-12.50)	(11.29)	(45.58)
CONST	51.09***	62.59***	77.16***	82.41***	82.20***
	(190.49)	(182.36)	(169.12)	(228.62)	(391.15)
Observations	7,200	7,200	7,200	7,200	7,200
Pseudo R^2	0.763	0.763	0.763	0.763	0.763

multiplied by 10 for ease of presentation

multiplied by 1,000 for ease of presentation

t statistics in parentheses, *** p<0.01.

The quantile regression results reveal that the increase of the performance gap between male and female long-distance cross-country skiers varies across quantiles and seems to be most pronounced in the lowest tier of the distribution. That is, relative performance differences between men and women appear to have increased the most among the less talented athletes. Perhaps surprisingly, no statistically significant changes are observed in the mid-tier of the distribution, suggesting that the overall increase of the performance gender gap is caused by high and low performers.

5 COMPETITIVE BALANCE IN DOMESTIC SPORTS LEAGUES – A GENDER COMPARISON

5.1 INTRODUCTION

If we were to believe the words of Henry Ford who once said "*competition is the keen cutting edge of business, always shaving away at costs*" (Henry Ford II, 1949), we would surely question the many examples of successful businesses that did exceedingly well by dominating an entire market. Former chairman and CEO of General Electric, Jack Welch, found a perhaps more contemporary formulation for some of today's business strategies: "*Number one, cash is king […] Number two, communicate […] Number three, buy or bury the competition*" (Blodget 2009). It appears that at least in some industries competition seems to be "negligible" at the "expense" of a monopoly.

In sporting events and especially in professional team sports the situation is different. Unlike businesses, that can sell their *own* products independently from those of their competitors, sport teams to some extent rely on the performance of their opponents. In a seminal paper in the sports economics literature, Rottenberg (1956: 242) argues that in professional sports "*competitors must be of approximately equal 'size' if any are to be successful; this seems to be a unique attribute of professional competitive sports*". In a completely unbalanced contest the exact outcome can be predicted with certainty, eventually causing the fans to lose interest in the competition. Conventional wisdom holds that some degree of outcome uncertainty is needed to attract fan interest in the form of stadium attendance and TV viewers. Neale (1964) describes this as the "League Standing Effect": In a very unbalanced league fan interest in the weaker teams will decline, which in the long run also causes fan interest in the stronger teams to fall. Neale (1964: 2) conceptualizes what every team must hope for: "*Oh Lord, make us good, but not that good*". Yet, in the more recent literature there is debate about whether a decline in competitive balance will in fact result in a reduction of fan interest (with mixed evidence emerging from empirical studies). Given the attendance figures in both Premier League and German Bundesliga, where stadium utilization rates are constantly well above 90 percent, it is unlikely that a temporary decline in competitive balance will noticeably affect fan demand. Moreover, due to stadium capacity restrictions it is often impossible to quantify the exact ticket demand for a match. In partic-

ular in the event of sold-out stadia information concerning the additional number of tickets that could have been sold is simply not available.

Therefore, the purpose of this paper is *not* to analyze the relationship between competitive balance and attendance, as this has been done in numerous papers before (see section 5.2). Instead we contribute to the existing sports economics literature by empirically analyzing the long-term development of competitive balance in selected domestic sports leagues, with a particular focus on gender differences in competition. Our data cover a period of 21 consecutive seasons (1990/91 – 2010/11) and stem from six different football leagues (men's and women's English Premier League, men's and women's English second division and men's and women's German Bundesliga) as well as two handball leagues (men's and women's German Handball Bundesliga).[78]

The remainder of the paper is structured as follows: Section 5.2 provides a selective review of the earlier and more recent sports economics literature with a focus on competitive balance in professional team sports. Section 5.3 describes the data and competitive balance measures, while the (predominantly descriptive) evidence is presented in section 5.4. Finally, section 5.5 concludes with a brief summary of the main results and some limitations as well as suggestions for future research.

5.2 REVIEW OF THE EXISTING SPORTS ECONOMICS LITERATURE

Apart from the already mentioned fundamental works by Rottenberg (1956) and Neale (1964) which opened up the sports economics literature, influential early studies include Sloane (1971) and El-Hodiri and Quirk (1971, 1974). Notable follow-up works, including Jennett (1984), Cairns et al. (1986), and Quirk and Fort (1992), to name but a few, are thoroughly summarized and reviewed by Fort and Quirk (1995). In the following years, a number of studies (Kesenne 1996, 1999, 2000; Hoehn and Szymanski 1999; Dobson and Goddard 2001; and, for a review of this line of literature, Marburger 2002) specifically focus on competitive balance in European football – as opposed to the North American Major Leagues.

[78] Data can be accessed on www.rsssf.com and www.statto.com (men's football Premier League and second Division), Rothmans Football Yearbooks and www.women.soccerway.com (women's football Premier League and second division), www.kicker.de (men's football Bundesliga), www.dfb.de (women's football Bundesliga), and www.bundesligainfo.de (men's and women's Handball Bundesliga).

In response to a symposium in the *Journal of Sports Economics* in 2002, where several articles from different disciplines (e.g. Humphreys 2002; Noll 2002; Sanderson 2002; Zimbalist 2002; Hall et al. 2002) present partially contradictory evidence, Fort and Maxcy (2003) review the relevant literature up to that time.[79] Thereby, they classify the empirical literature on competitive balance in two different strands.

The first strand of literature is the analysis of competitive balance (ACB) literature which analyzes in what way institutional changes in the business practices in professional sports leagues affect competitive balance over the course of time. In this respect, institutional changes may occur in the form of variations in the respective clubs' market size (e.g. Demmert 1973; Eckard 2001a; Fort 2001), the switch from a "reserve clause" system to "free agency"[80] (e.g. Balfour and Porter 1991; Eckard 2001b; Maxcy 2002), a change of the ownership structure of professional (football) teams (e.g. Szymanski and Kuypers 1999), or the expansion of a (single) league (e.g. Schmidt 2001) as well as the reorganization of a (multi-tier) league (e.g. Dobson et al. 2001). More recent studies in this "niche" of the competitive balance literature are presented by Hadley et al. (2005) who find that the 1994 players' strike in Major League Baseball (MLB) led to a significant deterioration of competitive balance in the post-1994 seasons, and Feddersen (2006) who estimates the impact of increasing UEFA Champions League payouts on the financial and sporting imbalance in the German football Bundesliga. Along the same lines, Haugen (2008) examines the effects caused by the change of the point award system (i.e., the transition from a "2-1-0" to a "3-1-0" system), while Pawlowski et al. (2010) treat the modification of the Champions League payout system at the turn of the millennium as a natural "quasi-experiment" to analyze competitive balance levels in the Top-5 European football leagues before and after these institutional changes.

The second strand of literature on competitive balance is much broader. It analyzes the effect of competitive balance on attendance and thus tests Rottenberg's (1956) uncertainty of outcome hypothesis (UOH). In contrast to the ACB literature which examines changes in the level of competitive balance across seasons (Butler 1995) the UOH literature usually applies "within-season" competitive balance measures to capture changes in attendance

[79] In particular, Zimbalist's (2002) article is criticized for several statements questioning the fan's perception of competitive balance. For a response to Fort and Maxcy's (2003) comment see Zimbalist (2003).

[80] In North American professional team sports leagues, the "reserve clause" has entitled the teams to own the rights of a player even after the expiration of the player's contract. Today, the reserve clause system has almost entirely been replaced by "free agency", implying that a player who is currently not under contract (i.e., a "free agent") is eligible to sign a contract with any club or franchise.

figures contingent on the degree of competitive balance in a given season. For a comprehensive review of the relevant UOH literature up to that time one can refer to Dobson and Goddard (2001) and Garcia and Rodriguez (2002).

In the following, the more recent UOH literature will be reviewed in detail. Focusing on MLB, Schmidt and Berri (2001) aim to empirically test two hypotheses, namely (i) that competitive balance in MLB decreased in the 1990s as a result of the increasing gap between rich and poor teams and, (ii) that there is a relationship between competitive balance and aggregate league attendance.[81] Their results suggest that the financial imbalance between the teams did not affect competitive balance in the predicted negative way, as the 1990s turned out to be the most competitive decade in MLB's history. The time-series analysis supports the positive impact of competitive balance on attendance on the aggregate level, although attendance per game appears to have increased only slightly. On the other hand, the panel data analysis revealed that improvements in single-year competitive balance had a statistically significant and negative impact on attendance, whereas improvements in long-term competitive balance over three and five years positively affected gate attendance.

Garcia and Rodriguez (2002) analyze the determinants of football match attendance in the Spanish Primera División and find no statistically significant correlation between attendance and the uncertainty of the outcome, as measured by the squared difference of league positions.[82] A positive effect on attendance was only found for home teams that were in contention for the league championship. Buzzacchi et al. (2003) develop a dynamic competitive balance measure to compare the "open" football leagues in Europe, characterized by promotion and relegation systems, with the North American "closed" leagues and observe that the "open" leagues are less balanced than the "closed" leagues under consideration in terms of the outcome of play while at the same time providing more equal opportunities regarding potential new entrants to a league. Feddersen and Maennig (2005) empirically analyze competitive balance developments in four European football leagues (i.e., English Premier League, Spanish Primera Division, German Bundesliga and Italian Serie A) and four US Major Leagues (i.e., MLB, NBA, NFL and NHL) over a time period of

[81] Previous research had focused on the relationship between competitive balance and within-season attendance. Schmidt and Berri (2001) conduct a time-series analysis spanning the entire history of MLB and additionally use a panel dataset for the years for which price data exist in order to test whether this relationship holds at the aggregate level.

[82] Since no betting odds were available for the Spanish league in the period of observation, the authors introduced a measure based on the relative difference in league positions.

more than 30 years. Thereby, they identify 48 trends of which 19 (12) are found significantly negative (positive), pointing at a growing (im)balance, with the remaining 17 trends insignificantly different from zero. For an extensive review of the UOH literature in European football see Groot (2008).

Recently, a number of studies focused on the North American Major Leagues. Drawing on attendance data in MLB covering a period of 50 years, Krautmann et al. (2010) use a within-season competitive balance measure to analyze the impact of what they term "playoff-uncertainty", i.e., a team's probability of making the playoffs, on attendance. They find that towards the end of the season tighter pennant races evoke an increase in league-wide attendance. Fort and Quirk (2011) seek to identify optimal (i.e., welfare-maximizing) competitive balance levels in the National Football League (NFL), representing a "closed" sports league where the major part of ticket sales comes from season tickets – as opposed to single-ticket leagues like MLB where a relatively large part of tickets is sold on the match day. Their research suggests that careful empirical analyses – incorporating potential influencing factors such as talent choice, regulatory governing structures and antitrust remedies – are required in order to determine whether an increase in competitive balance enhances welfare in a season-ticket league. A first attempt to close this gap in the empirical literature is made by Mills and Fort (2011, 2014) who look at annual league-level per-game attendance (LAPG), i.e., a league's average per game attendance in a given year. Using attendance data from the NBA (seasons 1955/56-2009/10), NFL (1934/35-2009/10) and NHL (1960/61-2009/10) they find little evidence supporting Rottenberg's UOH hypothesis. On the other hand, they find that the time series of LAPG in all three Major Leagues are non-stationary but stationary with break points which correspond to historical occurrences such as World Wars I and II, league mergers, league expansions, player strikes and the like.

Summarizing, there is an already large and still growing literature on the determinants (and consequences) of competitiveness in professional team sports. While the majority of studies focus on the US Major Leagues, a growing number of sports economists have drawn their attention to European team sports. In the following section, we discuss several competitive balance measures that serve as a basis for our analysis.

5.3 MEASURING COMPETITIVENESS IN EUROPEAN FOOTBALL

The most commonly used measure of competitive balance is the dispersion of winning percentages within sports leagues. Scully (1989), Vrooman (1995) and Quirk and Fort (1997) use the standard deviation of winning percentages to assess the performance dispersion of teams in North American sports leagues. Although European football is quite different in its organizational form to North American sports[83] – and in particular with respect to the point award system that allows tied games and which experienced a transition from a 2-1-0 to a 3-1-0 system in the mid-nineties – Szymanski (2001) argues that winning percentage is still a reliable measure of competitive balance in English (and hence European) football.[84] In order to take potential draws into account, the winning percentage (*WPCT)* is defined as the ratio of wins and weighted draws to total games played. In a league with *N* teams over a period of *T* seasons $WPCT_{i,t}$ is defined as the winning percentage of team *i* in season *t*. Following Humphreys (2002), the standard deviation of winning percentages for this league is:

$$\omega = \sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(WPCT_{i,t} - 0.5)^2}{NT}}, \qquad (1)$$

where the idealized *WPCT*[85] is subtracted from each team's individual *WPCT* in season *t* while the squared term is divided by the number of teams and seasons. Since ω largely depends on the league size in any given season *t* (ω decreases the greater the value for *N*) it is only possible to compare leagues of equal size. To take changes in league size into account, Leeds and von Allmen (2008: 156) introduce an additional measure, *Ratio R*. In order to obtain *Ratio R* the "ideal" standard deviation of winning percentages – representing the "ideal" competitive balance in a league with *N* teams – is computed as follows:

[83] One distinctive difference is that European football is characterized by a multi-tier "open" league system where the best teams get promoted to a higher division (and the worst teams are relegated), whereas the North American Major Leagues are based on a more rigid "closed-shop" system. Moreover, the labor market of European football players is less regulated and comes close to a competitive market (Szymanski and Smith 1997), in particular since the so-called "Bosman-ruling" in 1995 (see e.g. Simmons 1997, Antonioni and Cubbin 2000, Feess and Muehlheusser 2002, 2003a, 2003b, Feess et al. 2004, Frick et al. 2007 and Frick 2009), while collective bargaining agreements, salary caps and rookie draft systems regulate the labor markets of the North American Major Leagues. See also Fort (2000) for a comparison of the European and North American sports markets.

[84] Buzzacchi et al. (2001, 2003) introduce a dynamic measure of competitive balance which is particularly suitable for open leagues with relegation and promotion systems as can be found in most European sports leagues.

[85] The dispersion of winning percentages in a completely balanced league where all teams are of equal strength and claim the same number of wins and draws is 0.5.

$$\omega_p = \frac{0.5}{\sqrt{G}}, \tag{2}$$

where G is the number of match days per season in a given league. Here again, ω_p is decreasing in the number of games played. *Ratio R* is defined as the ratio of the "simple" standard deviation to the "ideal" standard deviation of winning percentages:

$$R = \frac{\omega}{\omega_p}. \tag{3}$$

If *R* takes the value of 1 the league is perfectly balanced; i.e., all teams perform equally well. Generally speaking, the closer *R* is to 1 the greater is the competitive balance in the league. A value of 2 would indicate that the league is rather unbalanced, while a value of 2.5 or above indicates that the league is completely unbalanced (see also Perline and Stoldt 2007a). With this measure it is possible to compare the level of competitive balance of diverse sports leagues of various sizes and across different sports; yet it also has some shortcomings. Although *R* is a convenient measure when applied for within-season competitive balance comparisons, it is incapable of capturing across-seasons changes in relative league standings within a particular league. This point is emphasized by the subsequent example following Humphreys (2002): Consider two hypothetical three-team leagues over a period of 3 seasons (Table 5-1). In one league each team wins the championship once but also finishes runner-up as well as at the bottom of the league table in one season whereas the win-loss record (and thus the rank) of each team in the other league is identical in each season. Consequently, *Ratio R* shows the same value for each league and for every season. However, competitive balance across-seasons is undoubtedly greater in *League 1* which "produces" three different champions in three years, whereas *League 2* is consistently dominated by the same team.

Table 5-1: Win-loss records in two hypothetical three-team leagues

League 1				*League 2*			
Team	Season 1	Season 2	Season 3	Team	Season 1	Season 2	Season 3
A	2-0	0-2	1-1	A	2-0	2-0	2-0
B	1-1	2-0	0-2	B	1-1	1-1	1-1
C	0-2	1-1	2-0	C	0-2	0-2	0-2

Source: Own illustration following Humphreys (2002).

This example illustrates that additional measures of competitive balance are needed in order to make valid assertions concerning the level of competitiveness in professional team sports. A method often used by industrial economists is to measure the "relative" competitiveness of an industry.[86] One measure that captures the concentration or distribution of firms within an industry is the *Herfindahl-Hirschman Index* (*HHI*). The *HHI* is defined as the squared sum of all companies' market shares in an industry:

$$HHI = \sum_{i}^{N} (MS_i)^2 \quad , \tag{4}$$

where MS_i is the market share of the i^{th} firm on a scale from 0 to 1. *HHI* can take values between 1/*N* (perfect parity among all firms in the industry) and 1 (pure monopoly). As this measure requires the market shares of *N* - 1 firms, potential data limitations might hamper research endeavors applying the *HHI* method. Using MLB as an example, Depken (1999) states that in professional sports data requirements do not constitute a problem as a team's "market share" can be measured as a team's percentage of total wins in the league. Since in European football not only wins and losses but also draws are possible, we use a team's percentage of total points in the league as a proxy for its "market share".[87] Taking into account changes in the league size over time, Depken (1999) introduces an additional measure, *dHHI* = *HHI* - 1/*N*, which adjusts the *HHI* in a linear way. We use the following – to some extent modified – measure:

$$HHI_{adj} = \frac{\sum_{i=1}^{N} s_i^{\,2}}{20/N} x100 \quad , \tag{5}$$

where *HHI* is divided by the quotient of 20 and the number of teams per league (*N*), representing the adjusted *HHI* for a league consisting of 20 teams.[88] In order to illustrate to what extent *HHI* is sensitive to changes in league size and how the new measure HHI_{adj} adjusts the results, consider a perfectly balanced league with *N* = 2, […], 20 teams (see Table 5-2).

[86] See, inter alia, Gilbert (1984) who analyzes the relative competitiveness of the banking industry, Sullivan (1985) for evidence in the cigarette industry, MacDonald (1987) for an analysis of the railroad industry, and Borenstein (1989) who investigates the airline industry.

[87] For the conversion of wins and draws into points we use the three-point system which came into effect in all major European football leagues in the mid-1990s. For inter-seasonal comparability, the accumulated points in the seasons prior to the transition from a 2-1-0 to a 3-1-0 system are also converted on the basis of the three-point system.

[88] For presentation purposes, this term is multiplied by 100.

Table 5-2: *HHI* and *HHIadj* values in a hypothetical perfectly balanced league of varying size

Number of teams	*HHI* x 100	HHI_{adj}
2	50	5
3	33.3	5
4	25	5
:	:	:
10	10	5
:	:	:
20	5	5

Source: Own illustration and own calculations.

While *HHI* is strongly influenced by the number of teams in a league (the more teams the lower *HHI*), we can control for the variation in league size using HHI_{adj}. Similar to the aforementioned *Ratio R* indicator, this measure can be used to compare competitive balance levels of different leagues of various sizes. However, HHI_{adj} suffers from the shortcomings that (i) (similar to *Ratio R*) changes in relative league standings over the years cannot be captured, thus making the results of across-seasons competitive balance comparisons less meaningful, and that (ii) due to the nonlinear distribution of points in the 3-1-0 point award system HHI_{adj} values for leagues of seemingly identical competitive balance levels can vary contingent on the number of draws. As illustrated in Table 5-3 below, two hypothetical leagues can be equally balanced as measured by wins, draws and losses. Yet, since draws lead to fewer points accumulated by both teams (i.e., two points for a draw instead of three points in case of a win) the distribution of points within each league can vary considerably. This, in turn, is reflected in differing HHI_{adj} values.

Table 5-3: The effect of draws on *HHIadj*

	League 1				*League 2*			
Team	Wins	Draws	Losses	Points	Wins	Draws	Losses	Points
A	6	0	0	18	6	0	0	18
B	2	2	2	8	3	0	3	9
C	2	2	2	8	3	0	3	9
D	0	0	6	0	0	0	6	0
		HHI_{adj}: 7.82				HHI_{adj}: 7.50		

Source: Own illustration and own calculations.

Admittedly, such extreme cases seem highly unlikely as the number of draws is relatively evenly distributed across leagues and over time.[89] Several studies (e.g. Fernandez-Cantelli and Meeden 2003; Brocas and Carrillo 2004; and Haugen 2008) have shown that the number of draws remained constant even after the transition from the former 2-1-0 point system to the 3-1-0 system. Therefore, the variation of HHI_{adj} is – if at all – only marginally influenced by a variation of the number of draws across leagues (and seasons) and can therefore be disregarded.

Another within-season competitive balance measure is the *Concentration Ratio* (CR_i), which is defined as the cumulated share of points (s_i) of the top $N = 1, 2, \ldots, i$ clubs in relation to the accumulated share of points won by the remaining teams in the league. In European football, usually the top 5 clubs of each season are compared to the rest of the league. In analogy to the previous two measures, CR_5 has to be standardized in order to account for changes in the size of the league(s). This is done as follows:

$$C5 - Index = \frac{\sum_{i=1}^{N} s_i^{\,2}}{5/N} x100 \quad , \tag{6}$$

where the *C5-Index* is defined as the quotient of the observable CR_5 to the CR_5 of a perfectly balanced league. By implication, the *C5-Index* in a perfectly balanced league would take a value of 100, while greater values imply a reduction in competitive balance. Similar to the already introduced competitive balance measures, *Ratio R* and HHI_{adj}, the *C5-Index* is convenient for the comparison of competitive balance levels of different leagues of different size. Moreover, Michie and Oughton (2004, 2005) point out that the *C5-Index* offers an advantage inasmuch as the exact percentage changes in the level of competitive balance can be assessed. If, for example, the *C5-Index* of a perfectly balanced league takes a value of 130 in the subsequent season, the league experienced a reduction in competitive balance of 30 percent. But here again, the pitfall is that changes in relative league standings over time cannot be captured.

[89] This can be corroborated by looking at the cumulated points in any given league *l* in season *t* and by comparing this value to the cumulated points in other seasons and other leagues. For example, the total points accumulated annually in the English Premier League can – given a league size of $N = 20$ teams – theoretically take values between 760 (only ties) and 1140 (all games decided). In fact, our own calculations reveal that the mean cumulated points per season in the Premier League in the previous 16 seasons (prior to that the league consisted of 22 teams for several consecutive years) is 1039.5, with a standard deviation (sd) of 10.4 (min.: 1021, max.: 1063). For comparison, the possible range of points in the German Bundesliga (given that $N = 18$) is [760, ..., 1140], while the mean cumulated points per season in the same observation period is 838.5 (sd: 12.0, min.: 810, max.: 855).

An alternative to the hitherto discussed within-season competitive balance measures is the dispersion of championship titles over the years. Similar to *HHI* (or HHI_{adj}) which measures the dispersion of wins (or points) within a given season, $HHI_{acrossseasons}$ quantifies to which extent a league is dominated by one or a few teams. It is defined as the sum of squared championship titles won by *i* teams, divided by the number of seasons:

$$HHI_{acrossseasons} = \frac{\sum_{i=1}^{N} CT_i^2}{N} \tag{7}$$

In the women's football Bundesliga, for example, seven different teams were able to win the championship in the 21 seasons between 1990/91 and 2010/11. Most titles were won by 1. FFC Frankfurt (7), followed by Potsdam (5), Siegen (4), FSV Frankfurt (2), while Duisburg, Brauweiler and Niederkirchen each won the championship once. The corresponding across-seasons competitive balance value can be computed as follows:

$$HHI_{acrossseasons} = \frac{(7^2 + 5^2 + 4^2 + 2^2 + 1^2 + 1^2 + 1^2)}{21} = \frac{97}{21} = 4.62 \tag{8}$$

The men's football Bundesliga, on the other hand, seems to be less balanced in terms of championship titles. Over the same period, six different teams won at least one championship, with Bayern Munich (10 titles) clearly dominating the league. The remaining titles were divided between Dortmund (4), Stuttgart, Kaiserslautern and Bremen (2 each), while Wolfsburg won their first and hitherto only title in the season 2008/09. The corresponding $HHI_{acrossseasons}$ takes a value of 6.14 and corroborates the finding that championship titles in the past two decades were divided more equally in women's football as compared to men's football.

Although this measure allows quantifying the degree of long-term competitiveness in sports leagues, it also suffers from major shortcomings. First, by measuring the dispersion of championship titles across seasons one cannot determine the "closeness" of the championships. For example, a league with 10 different champions in 10 different seasons might nevertheless be somewhat unbalanced if the respective championships were decided by a margin of 20 points or more. A league that is seemingly dominated by one single team, on the other hand, might in fact be fiercely contested, with the same team repeatedly finishing ahead of its closest competitors by a narrow margin. Second, since $HHI_{acrossseasons}$ is a highly aggregated measure that provides only one observation for the longitudinal data of each

league, it is only possible to descriptively compare the results between different leagues. Third, in light of the relegation system in European team sports leagues it makes no sense to measure the dispersion of championship titles in lower divisions as this would simply "produce" spurious evidence due to sample selection bias.[90]

Given the benefits and drawbacks of the competitive balance measures discussed earlier in this section (and taking the vast literature related to this topic into account), it appears that *the* ideal measure of competitive balance has yet to be developed. Meanwhile we remain with a whole set of different (complementary rather than substitutable) competitive balance indices of which only a few are discussed in this paper. In the next section we present the predominantly descriptive results of our analysis which are then supported by some basic test statistics.

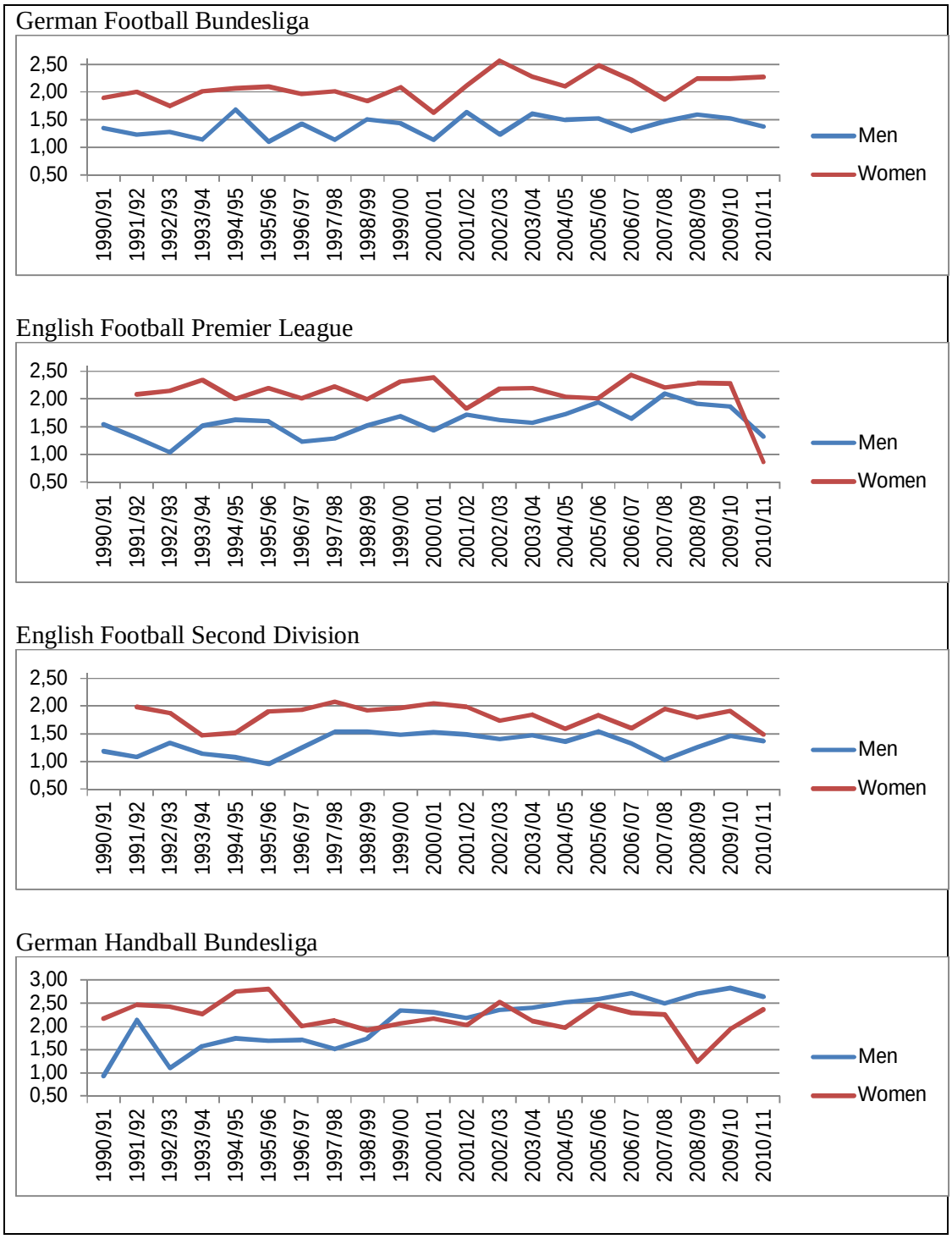
5.4 RESULTS

We begin with the presentation of the results by illustrating the long-term development of competitive balance levels in the earlier mentioned team sports leagues, with a particular focus on gender differences. For this purpose we compare annual *Ratio R* indices for each league, spanning a period of 21 seasons (and 20 seasons in the case of the English women's football leagues).[91]

[90] Those teams winning the championship e.g. in the second division are promoted to the first division in the subsequent season and are necessarily excluded from the pool of potential second division champions for at least one season. See Heckman (1979) for a thorough discussion of potential bias resulting from non-randomly selected samples.

[91] The women's football league was reintroduced in England in the season 1991/92 after a lengthy ban of women's football on the part of the English Football Association. Indeed, women's football has a much longer history than commonly assumed. Already in the 1920s, some women's football matches in England were played in front of more than 50,000 spectators (see Matheson and Congdon-Hohman 2013).

Table 5-4: Development of within-season competitive balance levels in all leagues (measured by *Ratio R*)

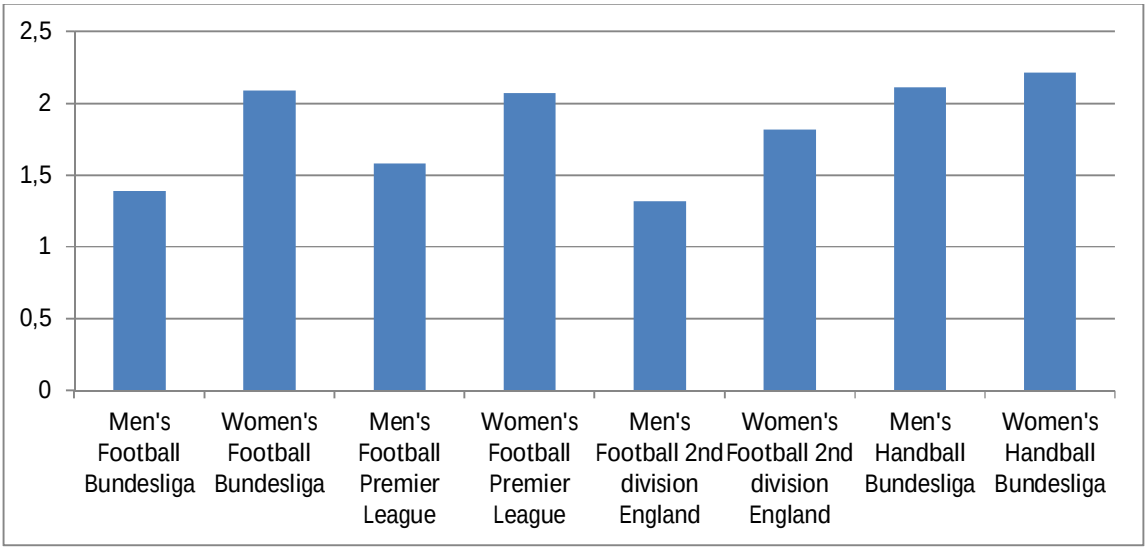


Note: *Ratio R* values are on the y-axis, while seasons are indicated on the x-axis. Own illustration based on own calculations.

It appears that in European football the men's leagues are more balanced than the women's leagues and that these differences are persistent over time. With the exception of a few seasons in the English Premier League where men's and women's leagues display similar levels of competitiveness, the observed gender differences seem to prevail in different geographical environments as well as across different divisions. On the other hand, the evidence from the German Handball Bundesliga is somewhat different. Although the men's league appears to be far more balanced than the women's league in the early seasons, there are almost no differences in the "middle years" while the gender gap in competitiveness seems to be reversed in the last third of the observation period. The results are very similar when measuring the level of competitive balance on the basis of the HHI_{adj} or the *C5-Index* (see Appendix D).

It should be stated that the results so far are of a rather descriptive nature as we cannot test for statistical significance with respect to single (i.e., annual) observations. Yet, it can be assumed that the (highly aggregate) information that is comprised in every single competitive balance indicator provides conclusive evidence regarding the level of (within-season) competitive balance in the respective leagues. Nevertheless, it is worth testing for the existence of statistically significant gender differences over time. This can be done by a simple one-way analysis of variance (ANOVA), where the response variable is the respective competitive balance measure while the distinguishing factor variable is gender. Since we conduct a two-sample ANOVA the differences between group means would be identical to the results of a t-test. Figure 5-1 below shows the results of the mean comparison tests.

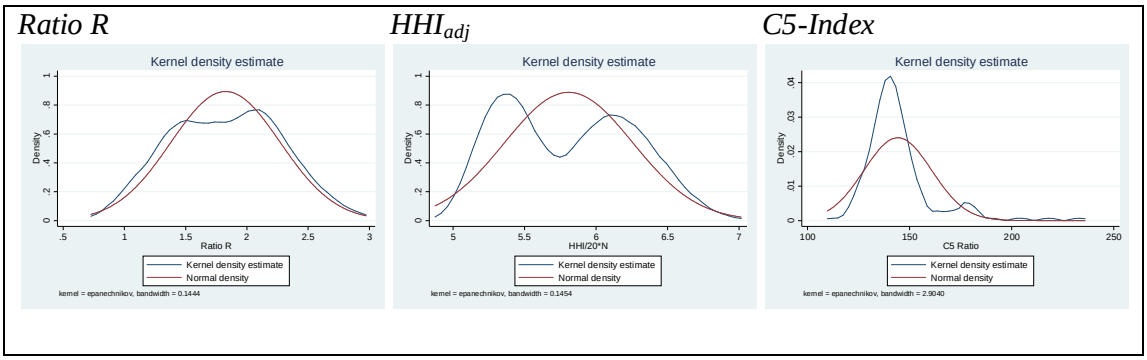
Figure 5-1: Comparison of average competitive balance levels between leagues (measured by *Ratio R*)



Note: *Ratio R* values are on the y-axis, while leagues are displayed on the x-axis. Own illustration based on own calculations. By conducting F-tests in Stata 11 for each "pair" of leagues we can reject the null hypothesis that the means of annual competitive balance levels are equal in the respective men's and women's football leagues (Prob> F < 0.0001). The marginal differences in the average competitiveness in men's and women's Handball Bundesliga, on the other hand, are found to be statistically insignificant. This result changes, however, when comparing the means of the HHI_{adj} indices (see Appendix D). In this case, we find statistically significant (albeit still comparatively small) differences even in German handball. Gender differences in competitiveness as expressed by the *C5-Index* are less pronounced and even found to be statistically insignificant in the English Premier League as well as in the Handball Bundesliga (see Appendix D).

The descriptive as well as econometric results indicate that there are considerable and persistent differences in competitiveness between the respective men's and women's football leagues. At the same time, the (overall) gender gap in German handball appears negligible, following an increase of competitiveness in the women's league and a simultaneous decrease of competitiveness in the men's league. In addition to the mean comparison tests we conducted Bartlett's (1937) test for equal variances for each "pair" of leagues. The results, however, have to be interpreted cautiously, since Bartlett's test is very sensitive to departures from normality (see e.g. Markowski and Markowski 1990). The assumption that the data are drawn from a normal distribution is obviously violated for all three indices (see Figure 5-2).

Figure 5-2: Epanechnikov kernel density estimates of alternative competitive balance measures



Note: Brown and Forsythe (1974) propose an alternative test that is robust under non-normality. This test can be conducted using the "sdtest"-command in Stata 11 and yields very similar results supporting the descriptive evidence above. All test results are available from the authors upon request.

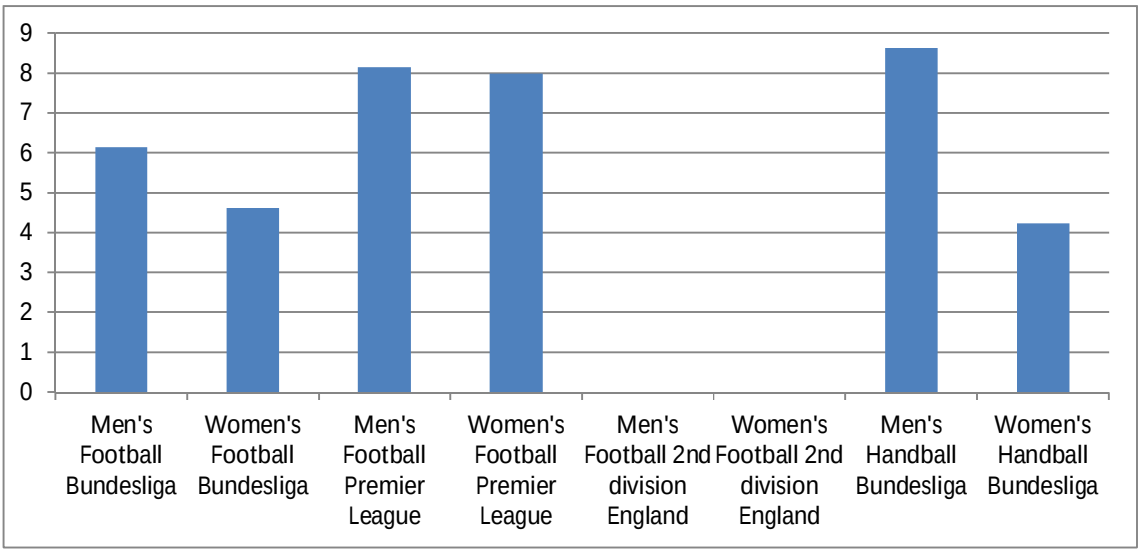
An interesting subsidiary aspect is that both the English men's and women's second division football leagues appear to be more "balanced" in terms of competitiveness than any of the first division leagues in England or Germany. A possible explanation for this perhaps surprising result is that the turnover of teams in the lower divisions is significantly larger than in any top-tier division. That is, while in the highest division only the weakest teams are being relegated (while the top teams can extend their lead over time), the regulation of the sporting equality in lower divisions works at both tails of the distributions: The strongest teams are promoted to a higher division while the weakest teams are demoted to a lower division.

As mentioned earlier, the women's football Bundesliga appears more balanced in terms of the dispersion of championship titles over time than the men's league. This finding is even more pronounced in German handball, where the men's league is clearly dominated by one team, THW Kiel, who won 13 out of 21 possible titles, whereas the titles in the women's league are more equally split among a number of teams. In English football, both leagues are equally unbalanced in terms of championships, with Manchester United and Arsenal London (12 titles each) dominating the men's and the women's Premier League, respectively (see Figure 5-3 for an overview of $HHI_{acrossseasons}$ indices).

These results are somewhat counterintuitive as one would normally expect a larger dispersion of championship titles in a "more balanced" league than in a seemingly unbalanced league. On the other hand, this example emphasizes that the analysis of competitive balance is far from trivial and can sometimes lead to markedly different results, depending on

the index applied. Considering the benefits and drawbacks of each method (see previous section), we devote greater weight to the within-season competitive balance measures.

Figure 5-3: Comparison of long-term competitive balance levels between leagues (measured by $HHI_{acrossseasons}$)



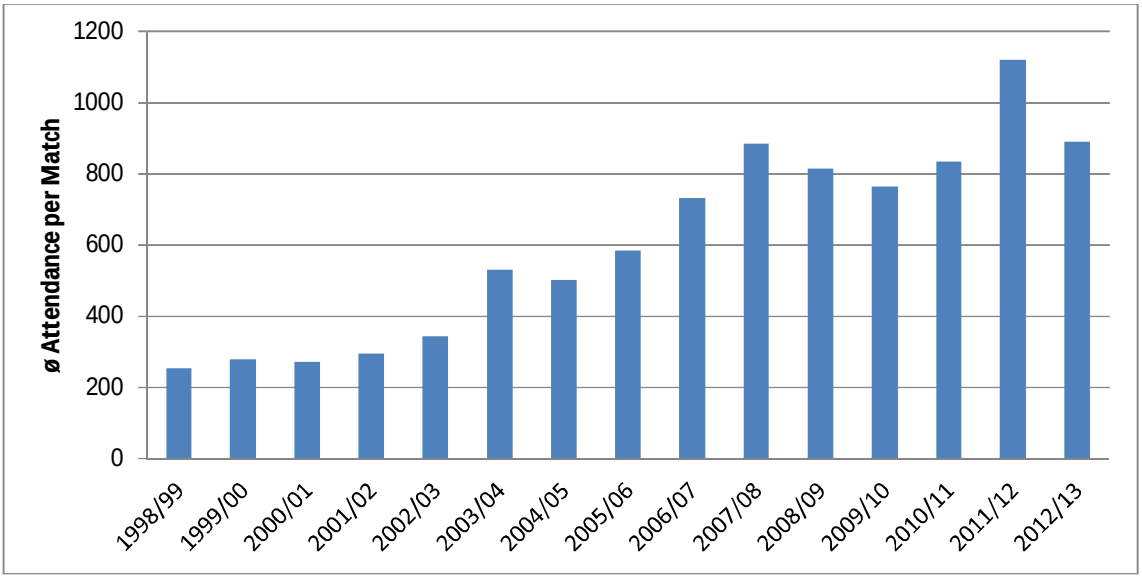
Note: $HHI_{acrossseasons}$ values are shown on the y-axis, while leagues are displayed on the x-axis. For reasons discussed in the previous section the dispersion of championship titles has not been computed for men's and women's second division football.

-A Brief Look at Attendance in Women's Football and Handball in Germany-

As mentioned in the introductory part of this paper, it is not our goal (nor would it be sensible) to analyze the relationship between the "closeness" of the competition and stadium attendance. This is due, in particular, to measurement problems arising from stadium capacity restrictions in the case of men's football: For the econometrician it is, for example, not possible to quantify the exact ticket demand for sold-out games. Nevertheless, it appears worthwhile to provide some descriptive statistics on fan demand in women's leagues.

Attendance in women's football, for example, is much lower than in the men's leagues but seems to be affected by largely televised events boosting the reputation of women's football in general. At least in Germany, the highest women's league seems to have heavily benefited from the women's Football World Cup taking place every four years (see Figure 5-4).

Figure 5-4: Average stadium attendance in the women's football Bundesliga over time



Note: Average per game attendance is displayed on the y-axis, while seasons are indicated on the x-axis. For reasons discussed in the previous section the dispersion of championship titles has not been computed for men's and women's second division football.

Obviously still in its infancy, the women's football Bundesliga received only little fan interest in terms of stadium attendance until the season 2002/03. Throughout the following seasons, attendance rates increased, although not on a constant level. A striking aspect in this context is that there are dramatic surges in attendance in the seasons following the women's Football World Cup (e.g. a 55% increase in the season 2003/04, 21% in 2007/08 and 42% in 2011/12). It appears that in particular the notable achievements of the German national team (which won the World Cup in 2003 and 2007) and the World Cup hosted in Germany in 2011 proved very beneficial for the promotion of women's football in Germany – at least in terms of ticket sales for women's Bundesliga games. Yet, the surges in attendance in the season following a World Cup were often followed by a decline in fan interest in the subsequent season(s).

An admittedly ad hoc (but nevertheless conceivable) explanation for the 25 percent increase in attendance in the season 2006/07 is the men's FIFA World Cup that was also hosted in Germany in 2006 and that might have generated positive spillover effects for women's football. However, in the absence of counterfactual evidence this remains highly speculative. That is, without information on how average capacity utilization rates in the women's football Bundesliga would have developed without the men's FIFA World Cup

on home soil, it is not possible to quantify the impact of the World Cup on fan demand. Besides, the same holds true for the impact of the respective women's World Cup competitions on fan demand, although the repeating pattern suggests that a causal connection exists.

Attendance data for the German handball leagues are only available for a few seasons. An inspection of the average attendance of teams active in the highest two divisions in the season 2010/11 (see Table 5-7 in Appendix D) suggests that the gender gap in fan demand is considerably smaller than in football. Indeed, some of the teams in the women's Bundesliga draw the same number of fans as men's second division teams, while one team, HC Leipzig, even comes close to the lower bound of attendance in the men's Bundesliga. Again, the data do not allow drawing any conclusions regarding the impact of the league's attractiveness (or competitive balance) on attendance. It might as well be the case that handball – being a sport that had originally been designed for women – is perceived more gender-neutral than football and is thus more appealing to fans. This, in turn, could (i) attract more fans to women's games and (ii) encourage more women to select into a competitive environment. The latter would explain why talent in women's handball is apparently more equally distributed (relative to the men's league) than is the case in women's football.

5.5 CONCLUSIONS

We analyze the long-term development of competitive balance in selected European team sports leagues. Thereby, we contribute to the existing literature by focusing on gender differences in competitive balance.[92] Examining the within-season competitive balance over a period of two decades, we find that in football (i.e., the German Bundesliga, English Premier League and second division) competition is far more balanced among men's teams than among women's teams. A possible explanation for the persistent gender differences is that the pool of highly skilled individuals is significantly smaller among women and that the comparatively heterogeneous female talent is unevenly distributed. The evidence emerging from a supposedly less male-dominated competitive environment (i.e., the German Handball Bundesliga) is somewhat different. Here, we are unable to find persistent and statistically significant gender differences in competitive balance. Yet, the descriptive evidence

[92] Similar research exploiting data from the North American Basketball leagues has been conducted by Perline and Stoldt (2007b) and Matheson et al. (2013).

suggests that the women's league has become increasingly competitive over time, while the men's league has become less balanced. Given the limited number of annual observations, it is of course too early to infer on systematic changes in the talent distributions among men and women. On the aggregate, however, it seems that gender differences in competiveness are considerably smaller in German top division handball than in some European football leagues.

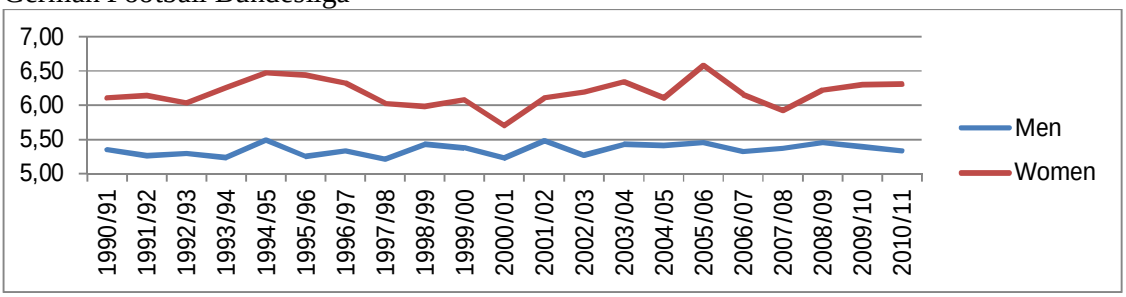
Although we deliberately refrain from *empirically* analyzing the effects of competitive balance on attendance, a *descriptive* analysis of attendance in some of the women's leagues provides two interesting insights. First, attendance in the women's football Bundesliga seems to be strongly influenced by international tournaments such as the women's Football World Cup. These largely televised events appear to have boosted the popularity of the sport in Germany which, in turn, led to a dramatic increase of spectators in the women's Bundesliga in every season following a World Cup event. Second, the most recent attendance figures from the German handball leagues reveal that the gender gap in attendance is considerably smaller than in the football leagues examined in this paper. Whether this is due to the increasing (relative) attractiveness (or "closeness") of the women's Handball Bundesliga remains to be tested in future research. Unfortunately, the underlying data do not allow conducting more sophisticated empirical analyses. A sample including a number of different sports in different countries and across different divisions would serve as a basis for a multivariate analysis.

Nevertheless, it remains difficult to determine causal relationships between competitive balance and attendance. A contested issue among sports economists is, for example, whether competitive balance indeed matters for fans. At least in European football it is commonly acknowledged that "David-against-Goliath encounters" attract significantly more fans than matches between two rather weak opponents with ex ante similar prospects of winning. Buraimo and Simmons (2008) explain this with the home fans' desire to be in the stadium in the rare event that David beats Goliath. It is also conceivable that superstar effects play a role, in the sense that the favorite team's star players draw "marginal" fans (i.e., people who under "normal" conditions would have preferred to stay at home) to the stadium. Disentangling these effects poses a huge challenge for the econometrician but at the same time provides avenues for future research.

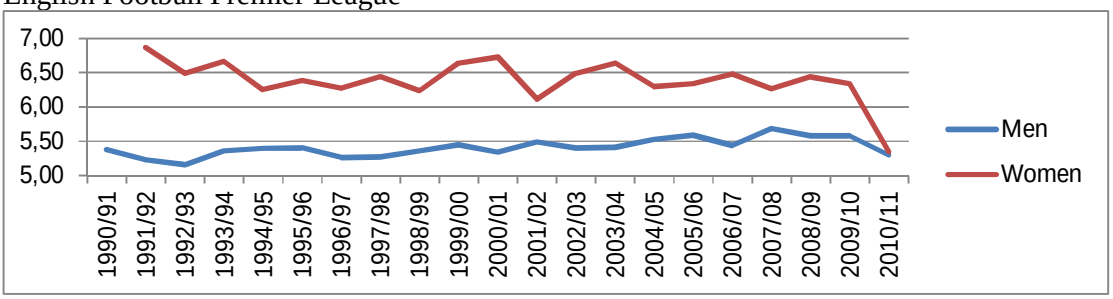
5.6 APPENDIX D

Table 5-5: Development of within-season competitive balance levels in all leagues (measured by HHI_{adj})

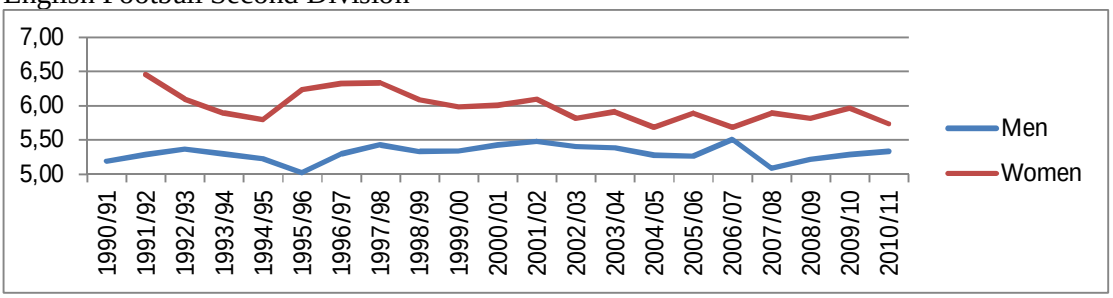
German Football Bundesliga



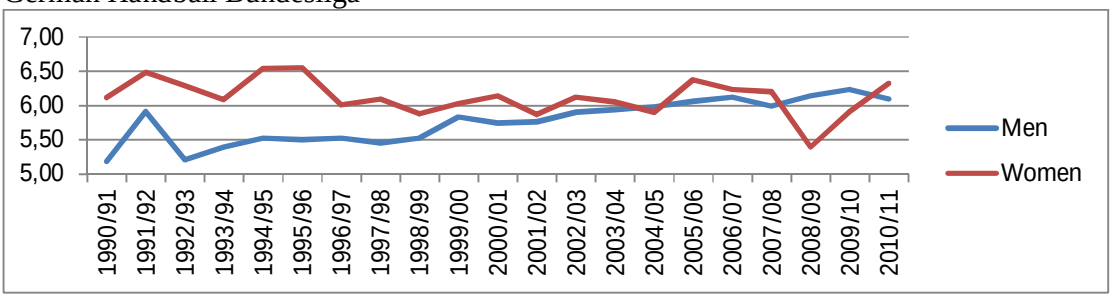
English Football Premier League



English Football Second Division

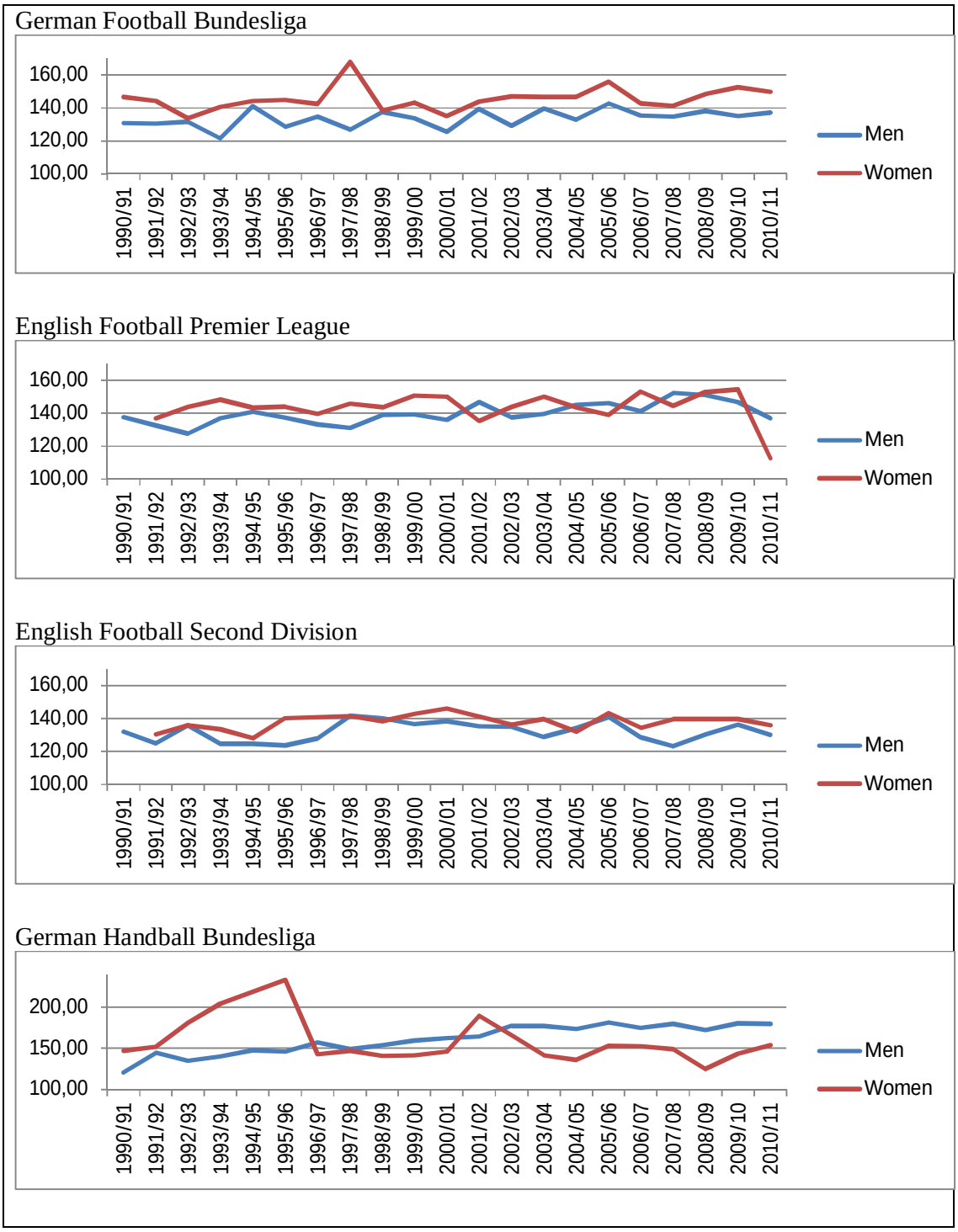


German Handball Bundesliga



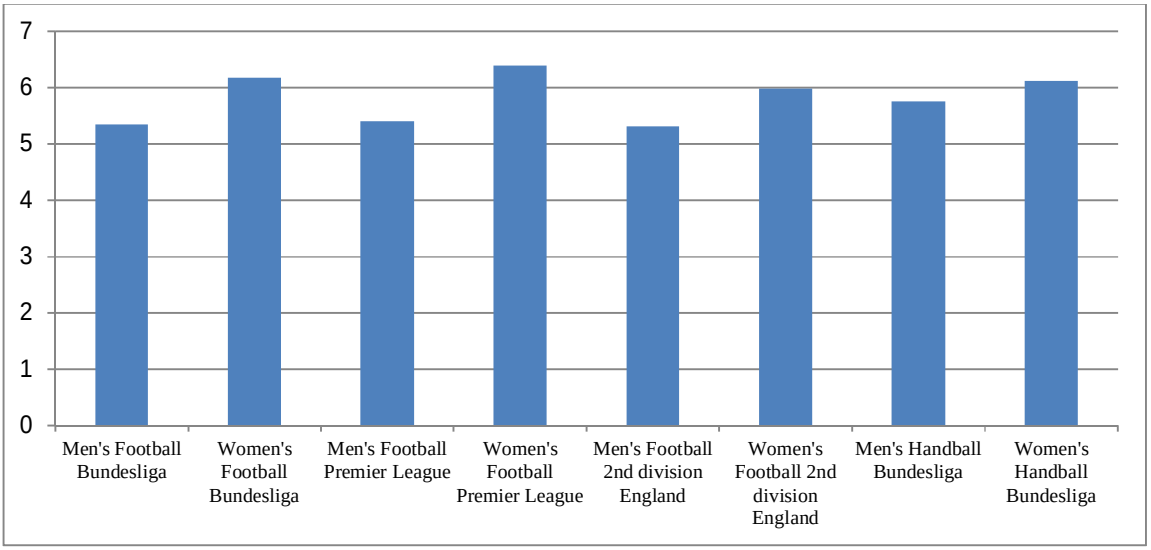
Note: HHI_{adj} values are on the y-axis, while seasons are indicated on the x-axis. Own illustration based on own calculations.

Table 5-6: Development of within-season competitive balance levels in all leagues (measured by *C5-Index*)



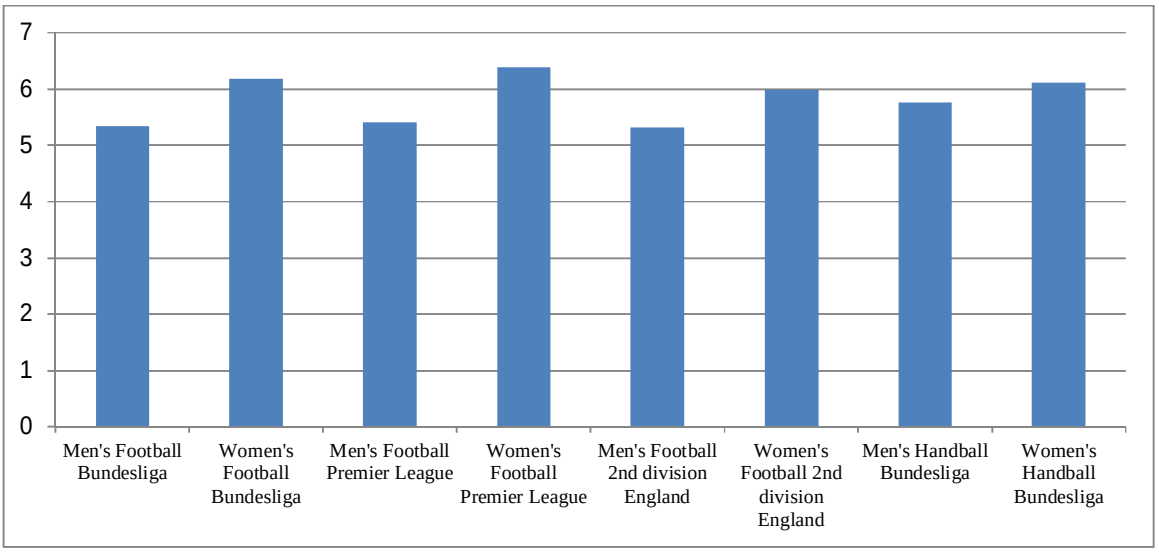
Note: *C5-Index* values are on the y-axis, while seasons are indicated on the x-axis. Own illustration based on own calculations.

Figure 5-5: Comparison of average competitive balance levels between leagues (measured by HHI_{adj})



Note: HHI_{adj} values are on the y-axis, while leagues are displayed on the x-axis. Own illustration based on own calculations.

Figure 5-6: Comparison of average competitive balance levels between leagues (measured by *C5-Index*)



Note: *C5-Index* values are on the y-axis, while leagues are displayed on the x-axis. Own illustration based on own calculations.

Table 5-7: Average attendance in German handball in the season 2010/11

Rank	Team	Overall Attendance	Average Attendance	Home Games
1	HSV Hamburg	181,722	10,689	17
2	THW Kiel	174,250	10,250	17
3	Füchse Berlin	138,946	8,173	17
4	Rhein-Neckar Löwen	136,559	8,032	17
5	SG Flensburg-H.	100,580	5,916	17
6	SC Magdeburg	88,594	5,211	17
7	Frisch Auf Göppingen	81,000	4,764	17
8	TBV Lemgo	74,788	4,399	17
9	HSG Wetzlar	60,324	3,548	17
10	TSV Hannover-Burgdorf	54,796	3,223	17
11	TV Großwallstadt	54,762	3,221	17
12	ASV Hamm-Westfalen	50,949	2,997	17
13	VfL Gummersbach	45,576	2,680	17
14	TuS N-Lübbecke	42,057	2,473	17
15	HBW Balingen-Weilstetten	41,240	2,425	17
16	MT Melsungen	38,361	2,256	17
17	GWD Minden	37,760	2,221	17
18	TV Bittenfeld	37,567	2,209	17
19	DHC Rheinland	36,967	2,174	17
20	TSG Friesenheim	35,400	2,082	17
21	HC Leipzig	27,020	2,078	13
22	VfL Bad Schwartau	32,232	2,014	16
23	SV Post Schwerin	30,657	1,916	16
24	TuSEM Essen	29,917	1,869	16
25	Eintracht Hildesheim	28,403	1,775	16
26	ThSV Eisenach	27,554	1,620	17
27	Bergischer HC	26,795	1,576	17
28	HG Saarlouis	25,099	1,476	17
29	Buxtehuder SV	19,283	1,377	14
30	Wilhelmshavener HV	21,915	1,369	16
31	SG BBM Bietigheim	22,309	1,312	17
31	HSG Nordhorn-Lingen	20,552	1,284	16
32	TV Emsdetten	20,465	1,279	16
33	TV Hüttenberg	22,488	1,249	18
34	HC Erlangen	20,800	1,223	17
35	TV 1893 Neuhausen	20,800	1,223	17
36	EHV Aue	20,700	1,217	17
37	HC Empor Rostock	19,316	1,207	16
38	Dessau-Roßlauer HV 06	18,430	1,151	16
39	VfL Oldenburg	14,360	1,104	13
40	HSG Düsseldorf	18,663	1,097	17
41	HSC Coburg	18,550	1,091	17
42	Thüringer HC	14,544	1,038	14
43	DJK/MJC Trier	11,110	1,010	11
44	FHC Frankfurt/Oder	11,667	972	12
45	SG H2Ku Herrenberg	16,509	971	17
46	Bayer Leverkusen	10,370	864	12
47	TSG Wismar	9,380	852	11
48	Frisch Auf Göppingen	9,200	836	11
49	SVG Celle	10,670	820	13
50	TuSpo Obernburg	13,900	817	17
	Average	41,683	2,522	16

Men's first division teams

Men's second division teams

Women's first division teams

Women's second division teams

Source: www.handball-world.com and own calculations.

6 GENDER DIFFERENCES IN DECISION-MAKING UNDER RISK: EVIDENCE FROM TV GAME SHOW DATA

6.1 INTRODUCTION

A wide range of theories of decision making under risk have been developed and have contributed to the understanding of differences in individual behavior, including von Neumann and Morgenstern's (1944) normative expected utility theory as well as Kahneman and Tversky's (1979) prospect theory. Nevertheless, the empirical testing of individual decision making, especially with regard to risk and overconfidence, has proven to be difficult. Laboratory experiments, for example, are characterized by a high degree of internal validity but are often criticized for their lack of external validity. One point of criticism is that laboratory experiments are limited in terms of their budget, which often forces experimenters to incentivize participants with small or even hypothetical stakes. This, in turn, could result in subjects not being motivated to act realistically and, thus, bears the risk of subjects not revealing their actual preferences and beliefs (Post et al. 2008). In this context, Holt and Laury (2002) find evidence that subjects tend to be more risk averse when they face large monetary stakes in contrast to hypothetical or small amounts of money.

One possibility to circumvent these problems is to use field data[93], and in particular data from TV game shows. In fact, television shows provide a good natural context in which individuals face well-defined decision-making tasks that are usually linked to large monetary amounts at stake. Game shows that have been in the focus of economic research include, among others, *Card Sharks* (Gertner 1993), *Jeopardy!* (Metrick 1995; Säve-Söderbergh and Lindquist 2011), *Lingo* (Beetsma and Schotman 2001), *Hoosier Millionaire* (Fullenkamp et al. 2003), *Who Wants to be a Millionaire?* (Daghofer 2007; Johnson and Gleason 2009; Hartley et al. 2013), *The Weakest Link* (Levitt 2003; Antonovics et al. 2005, 2009) and the recently widely exploited TV show *Deal or No Deal* (Post et al. 2008; Deck et al. 2008; Brooks et al. 2009; Mulino et al. 2009; De Roos and Sarafidis 2010; Bombardini and Trebbi 2012).

The game show that is analyzed in this paper, *The Million Pound Drop* (as well as the German and Swiss equivalents of this show, *Rette die Million* and *Die Millionenfalle*) has,

[93] Rather than being a substitute for lab-based experiments, Harrison and List (2004) conclude in an extensive review of the relevant literature that field and laboratory experiments complement one another.

to the best of our knowledge, not been empirically researched before, but provides an ideal setting to analyze gender differences in decision making under risk. The first and main advantage is that the stakes are exceptionally high from the beginning of the game: Already the first decision involves an amount at stake of one million pounds. This stands in stark contrast to other game shows, such as *Who Wants to be a Millionaire,* where participants get to the high-stakes questions only after having "survived" a number of rounds by correctly answering questions. We are thus able to rule out – or at least reduce – any knowledge-specific selection effects. This is particularly true since the rules of the game are simple and well-known to the participants. Another distinctive feature is that contestants compete together in pairs. Hence, besides the possibility to observe and compare the performance of single-gender and mixed teams, it is possible to examine individual behavior of male and female contestants depending on the team composition.

Using data covering 45 episodes, 94 different teams and 567 individual decisions from the above-mentioned game show, we test, first, if male teams are less risk averse than female teams, and second, if men are overconfident in particular when playing with a female team partner. We contribute to the already large and still growing literature by analyzing gender differences in overconfidence and risk behavior in a high-stakes real-life environment. At the same time, the results have important implications for the labor market inasmuch as they can partly explain the observed gender differences on the playing field, e.g. the underrepresentation of women in leading positions in management, politics, academia, and the like.

The paper proceeds as follows: Section 6.2 provides a selective review of the literature on (gender) differences in decision making under risk as well as some theoretical considerations. Section 6.3 describes the basic rules of the game show. In section 6.4, we present the data, develop our hypotheses and provide some descriptive evidence. The empirical analyses are presented and discussed in section 6.5, while section 6.6 offers concluding remarks, limitations and suggestions for further research.

6.2 LITERATURE REVIEW AND THEORETICAL CONSIDERATIONS

There are two strands of literature related to our research. The first strand of literature analyzes (gender) differences in decision making under risk, while the second addresses one of the potential causes of differences in individual risk preferences, namely the phenomenon of overconfidence in economic decision making. Overconfidence refers to the stylized fact

emerging from the social psychology literature stating that boundedly rational (as opposed to rational and utility-maximizing) individuals tend to overestimate their own capability relative to others (see e.g. Larwood and Whittaker 1977; Svensson 1981; Alicke 1985; Benoît and Dubra 2011). The analysis of individual risk preferences and overconfident behavior followed in the wake of alternative theories of non-rational behavior. Simon (1955) was among the first to question the standard economic model in which utility maximization and perfect rationality are inherent traits of the "homo economicus". In what has become popular as prospect theory, Kahneman and Tversky (1979) (and advancements to this theory by Tversky and Kahneman 1991, 1992) argue that individuals are sensitive to losses and gains based on (subjective) reference points rather than (objective) probabilities. This, in turn, often leads to suboptimal outcomes from a general welfare (or neo-classical) perspective. In this context, other behavioristic traits, such as myopia, anticipated regret or perceptions of fairness and reciprocity that could influence individual decision making (see, for example, Bruni and Sugden 2007) are not considered in the present study.

In the following, we review some of the "classical" and more recent works related to either of the two above-mentioned strands, beginning with the literature referring to gender differences in decision making under risk. A large part of the literature on gender differences in individual risk preferences uses data from laboratory experiments. The general tenor of these studies is that women are, on average, less confident and more risk averse than men. The following Table 6-1 summarizes some of the available literature with a particular focus on gender differences in risk behavior. This list is of course far from exhaustive.

Table 6-1: Overview of selected experimental studies on (gender) differences in risk behavior

Summary Table of Gender Differences in Risk Behavior				
Author(s)	**Setting**	**Experimental Design**	**Results**	**Significantly More Risk Averse Sex**
Brinig (1995)	Abstract	Subjects were given the choice of one of three urns, having a 90% chance of winning a "very small" prize, a 20% chance of winning a "slightly larger" prize and a 5% chance of winning a "very large" prize.	Male subjects have a greater preference for risk from the onset of adolescence to around the mid-forties.	Female
Eckel, Grossman (2002)	Abstract	Subjects had to choose one out of five lotteries which were associated with different risk choices and expected returns.	On average, women are found to be consistently more risk averse than men.	Female
Harbaugh, Krause, Vester-lund (2002)	Abstract	Subjects from age 5 to 64 were faced with choices between a certain outcome and a lottery in order to examine how risk attitudes change with age.	*Gain domain:* Subjects are risk seeking when they face high-probability prospects over gains and are risk averse when they are confronted with relatively small-probability prospects.	Neither
			Loss domain: On the other hand, subjects are risk averse when they face high-probability losses and are risk seeking when confronted with small-probability losses.	Neither
Holt, Laury (2002)	Abstract	Subjects faced a lottery-choice experiment with the intention of measuring the degree of risk aversion over a wide range of low and high payoffs.	*Low payoff:* Women are slightly more risk averse than men.	Female
			High payoff: Men and women become more risk averse.	Neither
Johnson, Powell (1994)	Contextual	Men and women made betting decisions on horse and dog races.	Women prefer gambles with a high probability of low returns and, therefore, take less risk, whereas men prefer gambles with a lower chance of some higher return.	Female
Powell, Ansic (1997)	Contextual	*Insurance setting:* Subjects faced insurance cover decisions.	*Insurance setting:* Female subjects more often buy (extensive) insurance than male subjects and are more risk averse.	Female
		Currency market experiment: Subjects faced currency market decisions and had to trade one currency for another in a risky market with the intention of making gains.	*Currency market experiment:* Female subjects are more risk averse than men when they have to decide whether to avoid risk while holding the wealth level or trading one currency for another in a risky market environment and possibly losing money.	Female
Gysler, Kruse, Schubert (2002)	Contextual	Subjects were confronted with twelve lotteries which included different risk levels in a financial decision context. The subjects were asked to bet on whether or not the lotteries would post a daily market price increase.	Controlling for factors such as competence, knowledge of financial markets and confidence in their own judgements, Gysler et al. show that women are significantly more risk averse than men.	Female
Agnew, Anderson, Gerlach, Szykman (2008)	Contextual	Subjects at retiring age had the choice between purchasing a fixed immediate lifetime annuity or investing their savings on their own.	Women are more likely than men to choose the annuity and are less willing to invest the savings on their own.	Female
Eckel, Grossman (2008)	Contextual	Subjects faced specific choices and had to make decisions with regard to the investment in a share of stock of one of five different companies	Male participants are significantly more risk prone than female subjects.	Female

Source: Own illustration.

The finding that women are, on average, less inclined to take risk than men is corroborated by a comprehensive review of the experimental economics literature by Croson and Gneezy (2009). This, in turn, has important implications for the participation of men and women in the labor market. If women systematically make other choices and behave differently than men, this can explain part of the gender gap observed in payment and key management positions. On the one hand, risk averse women are less likely to select into competitive career paths than equally-endowed men (see e.g. Nekby et al. 2008 for evidence from already highly selected road runners). On the other hand, if employers believe that women lack the necessary confidence or risk-loving attitude required for certain positions, they are likely to prefer male candidates over female candidates in recruitment or promotion decisions. Generally, most management decisions are made in uncertain and risky environments and are associated with uncertain opportunities, threats or costs.

However, results from laboratory studies of the above kind are often criticized for their lack of external validity and should, therefore, be interpreted with caution. A contested issue is, for example, whether the behavior of men and women in a "non-managerial" population typically consisting of undergraduate students allows drawing inferences on the behavior of males and females in top management positions. Along these lines, Johnson and Powell (1994) explore gender differences in the nature of decisions taken by a "managerial" population of potential and actual managers and compare these results to a "non-managerial" population of undergraduate students. The results suggest that the commonly observed stereotypes and gender differences may not apply to highly selected subpopulations. Indeed, males and females with a managerial background are found to display similar risk preferences and make decisions of equal quality.[94]

Despite these exceptions, the intuition is that among the general (i.e., "non-managerial") population women are significantly more risk averse than men. In the following, we discuss potential explanations for the observed gender differences. First, as posited by some social psychologists (e.g. Fujita et al. 1991; Brody 1993; Larkin and Pines 2003), women show greater emotional reactions to situations involving risk and uncertainty. As a result, women are more likely to avoid negative social outcomes especially when being in the

[94] Similar evidence (and criticism) is provided by Schubert et al. (1999). The intuition that some subpopulations tend to be distinctively different than the general population in terms of their attitudes toward risk is supported by Heß et al. (2013). Drawing on survey data from the German Bundestag as well as the German Socio-Economic Panel (GSOEP), the authors show that career politicians in Germany have significantly stronger risk preferences than the general population of GSOEP respondents.

center of public attention. Moreover, Croson and Gneezy (2009) conjecture that women experience emotions more intensely than men and that this trait can eventually adversely affect their decision making under risk. Another explanatory approach is that individuals differ in their perception of risk (see, inter alia, Arch 1993; Sitkin and Weingart 1995; Croson and Gneezy 2009). Whereas men tend to appraise risky situations as a challenge, women are more likely to perceive a risky environment as a threat and thus shy away from such environments. Finally, and to some extent related to the previous point, individuals tend to be overconfident (Camerer and Lovallo 1999; van den Steen 2004; Hoelzl and Rustichini 2005)[95], with men being more overconfident than women (Niederle and Vesterlund 2007, Nekby et al. 2008; Gerdes and Gränsmark 2010; Reuben et al. 2012).

Although the first two explanations for gender differences in risk behavior are valid in everyday contexts, they are supposedly less conclusive for potential differences as observed in the quiz show *The Million Pound Drop*. Due to the contestants' deliberate decision to participate in a risky (and of course public) environment, we observe a sub-population of males and females who are likely to be less risk averse than the general population. This has two important implications for our analysis. First, if we are able to identify systematic differences in the risk behavior of male and female game show participants, these effects are likely to be even more pronounced among the general population (i.e., we under- rather than overestimate the effects). Second, due to the aforementioned selection effects we can rule out some of the common explanations for differences in risk preferences and are thus able to examine overconfident behavior in an isolated manner.

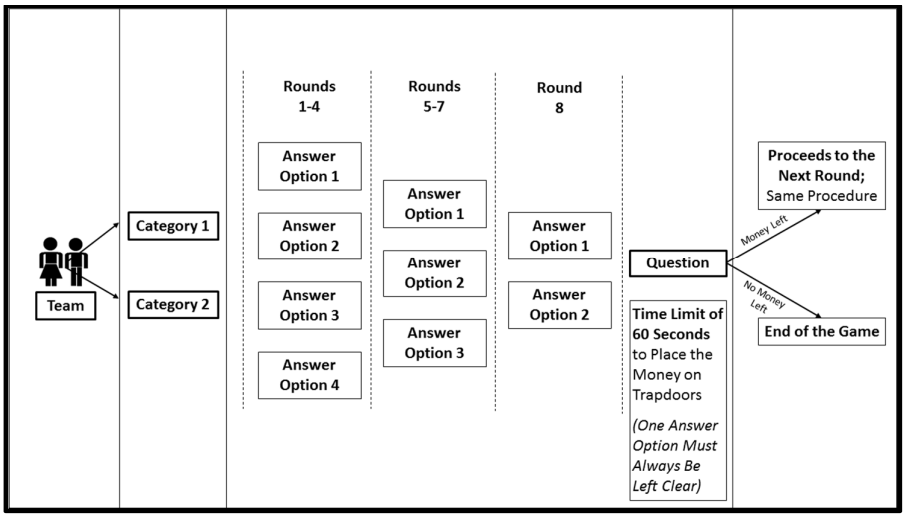
6.3 DESCRIPTION OF THE GAME SHOW

The TV game show *The Million Pound Drop* was developed by the Dutch production company *Endemol* and first aired in the United Kingdom in May 2010. The game show soon became popular and was exported to dozens of other countries, including Germany and Switzerland. The following description applies to the episodes of *The Million Pound Drop* aired in the United Kingdom. Except for the slightly different and (due to currency fluctuations) varying monetary amounts, the main structure of this quiz show is identical to the German and Swiss versions.

[95] Further evidence regarding the causes and consequences of overconfidence has recently been provided by Nöth and Weber (2003), Koszegi (2006), Menkhoff et al. (2006) and Garcia et al. (2007). Overconfidence is not only observed among the general population but also among CEOs who are found to overstate future returns of their companies (Malmendier and Tate 2005, 2008) and among financial analysts who tend to become overconfident following a series of accurate predictions (Hilary and Menzley 2006).

The contestants, who apply for and play in the show together in pairs, are given one million pounds in £50 notes which are divided into 40 bundles of £25,000 each. The teams must then answer eight questions with the aim of retaining as much of their initial prize-money as possible. At the beginning of each round, teams have to pick one out of two categories and then have to answer a multiple-choice question related to this category. The first four questions have four possible answer options, while the next three questions have three and the final question only has two answer options. In each round, the candidates can either bet all of their (remaining) money on one answer option or wager their bets, depending on how confident they are about the answer. However, there are two important restrictions. First, one answer field must always be kept clear with no money on it. That implies that the splitting of stakes is no longer possible in the eighth and last (all-or-nothing) round. Second, after the question and the corresponding answer options have been presented by the anchorwoman, the team has 60 seconds to place the money onto trapdoors corresponding to the answer(s) chosen. If the candidates are confident about their answer, they are able to stop the clock before the time limit is up. The trapdoors that are related to incorrect answers then open and any money placed on these trapdoors is lost. Moreover, those bundles that are not placed on either answer option by the end of the time limit are lost, too. If a team has any money left, it proceeds to the next round where the process is repeated. Yet, teams losing all their money in one particular round are out of the game and leave the show empty-handed. Therefore, the contestants have a huge incentive to retain as much of their money as possible. Figure 6-1 illustrates the basic structure of the main game.

Figure 6-1: Flow chart of the main game



Note: Pairs can also consist of two male or two female candidates. Source: Own illustration.

Episodes usually last between one and two hours and include two to four teams competing one after another. Most recently, the show's format was significantly modified, with teams of four candidates playing the game. In addition to that, a lottery was introduced: Contestants now have the choice between either "cashing out" after question seven or doubling their prize money with a correctly answered question eight. Our analysis is however based on the "traditional" format as described above.

6.4 DATA, HYPOTHESES AND DESCRIPTIVE EVIDENCE

The data used for the analysis come from 45 episodes of *The Million Pound Drop* (as well as the German and Swiss equivalents), covering 94 different teams and 567 individual decisions. Collection of the data was done by watching game show episodes that were (and to some extent still are) available online on *youtube* and the respective broadcasters' websites. The structure and composition of the dataset is illustrated below.

Table 6-2: Structure of the dataset

Country	Episodes	Observation period	Teams	Decisions excluding question 8	Decisions including question 8
UK	15	October 2010 - February 2012	33	182	194
GER	8	May 2011 - March 2012	24	137	147
CH	22	July 2011 - June 2012	37	211	226
∑	**45**	**2010 - 2012**	**94**	**530**	**567**

As already mentioned in the previous section, question number eight includes only two possible answer options and does not allow any strategic splitting behavior. Due to this limited decision choice, the observations of the eighth question can only be considered in some of our models.

Apart from the advantages discussed at the beginning of this paper, the game show is well suited for the empirical testing of risk behavior and overconfidence for two further reasons: First, we can quantify the teams' splitting behavior by means of a concentration ratio which serves as a proxy for risk aversion.[96] Second, during the one-minute decision phase

[96] Of course, a large part of those questions where candidates decide to not split their stakes but instead place all of their money on one single trapdoor are hardly indicative of a particularly risk loving attitude. Instead, it appears more likely that the candidates simply know the correct answer which, in turn, would disqualify this observation as a decision under uncertainty. On the other hand, we cannot categorically rule out the possibil-

we are able to observe each candidate's behavior and can, therefore, explore whether men and women behave differently in single-gender and mixed teams. Based on the theoretical considerations and the literature reviewed in section 6.2, the following hypotheses are derived:

H_{risk}: *Male teams bet larger shares of their money on the answers they believe to be correct (i.e., they are less risk averse) than female teams which prefer to diversify their bets.*

$H_{overconfidence}$: *Men are overconfident especially when playing with women in mixed teams.*

Table 6-3 provides summary statistics for our sample of 45 game show episodes aired in Germany, Switzerland and the United Kingdom in the years 2010 to 2012. For each team we collected, among others, the team's composition[97], each candidate's age, the remaining amount of money prior to each decision, the proportion of money bet on the preferred answer (and whether the preferred answer is correct or not), the time that is needed for a decision, the question number (which also controls for the number of possible answer options) as well as the number of answer options chosen. Furthermore, information on the decision-makers of the categories and the quiz questions, last-minute money shifts and the success- and failure-rates in previous rounds was collected. Finally, we calculated an inequality measure in order to capture the splitting and risk behavior of male, female and mixed teams. We use the Gini coefficient which is however slightly adjusted in order to control for the varying number of answer options. The adjusted Gini coefficient ranges between zero and one, with a value of zero indicating that the money is split equally, whereas a value of one implies that the entire amount is bet on one answer option.

We use the following formula to calculate the Gini coefficient:

$$G= \sum_i^n h_i \frac{2i-n-1}{n} \text{ with } 0 \leq G \leq 1-\frac{1}{n}, \quad (1)$$

ity that especially in the later rounds (and with less money at stake) candidates stake everything "on one card" despite being uncertain about the correct answer.

[97] Here, we focused on the gender of both candidates. We also looked at team diversity with respect to age, relationship and occupational status of the candidates to examine whether, for example, a mother and her daughter or an employer and his employee behave differently than e.g. siblings or colleagues. Yet, none of these diversity measures turned out to significantly affect the decision-making behavior. This might be due to the fact that teams are indeed rather homogeneous regarding the latter characteristics.

where G represents the Gini coefficient,
i indicates the respective answer option on which the money is bet,
h_i is the relative share of money bet on an answer option, and
n stands for the number of splitting options.

To control for the varying number of answer (and splitting) options, we divide *G* by the inverse of the number of splitting options:

$$\text{Adj. Gini} = {}^{G}\!/_{\frac{1}{n}} \quad (2)$$

Table 6-3: Summary statistics

Variable	Operationalization	Obs.	Mean	SD	Min	Max
Adj. Gini	Adjusted Gini coefficient, tending towards zero if money is split equally (and towards one otherwise)	530	0.70	0.34	0	1
Team composition						
Mixed team	Dummy taking the value of 1 if team consists of a male and a female, zero otherwise	567	0.55	-	0	1
Male team	Dummy taking the value of 1 if team consists of two males, zero otherwise	567	0.30	-	0	1
Female team	Dummy taking the value of 1 if team consists of two females, zero otherwise	567	0.15	-	0	1
Country						
Country (UK)	Dummy taking the value of 1 if game show was broadcasted in the UK, zero otherwise	567	0.34	-	0	1
Country (GER)	Dummy taking the value of 1 if game show was broadcasted in Germany, zero otherwise	567	0.26	-	0	1
Country (CH)	Dummy taking the value of 1 if game show was broadcasted in Switzerland, zero otherwise	567	0.40	-	0	1
Age	Candidate's age in years	567	35.9	10.6	20	93
CumAge	Teams' cumulative age	567	71.8	21.2	40	162
Question no.	Question number (or round of the game)	567	3.88	2.15	1	8
Options chosen	Number of answer options bet on	530	1.70	0.72	1	3
Proportion preferred	Proportion of money bet on preferred answer	530	0.80	0.21	0.33	1
Preferred correct	Dummy indicating whether the preferred answer is correct (1) or not (0)	530	0.75	-	0	1
Time	Time needed to physically move the money onto the trapdoors related to the answer options (in seconds)	567	55.11	11.12	2	60

Timer stopped	Dummy indicating whether the timer was stopped (1) or not (0)	567	0.21	-	0	1
Balance	Remaining amount of money at the beginning of each round	567	589,242	387,779	25,000	1,000,000
Failure	Counter variable indicating the number of consecutive failures (i.e., substantial losses due to an incorrect preferred answer option) in previous rounds	567	0.19	0.49	0	4
Success	Counter variable indicating the number of consecutive successes (i.e., correct preferred answer option) in previous rounds	567	1.61	1.68	0	7
Decider category	Categorical variable denoting the decision-maker regarding the category	567	-	-	1	7
Decider question	Categorical variable denoting the decision-maker regarding the question	567	-	-	1	7
Last-minute en-largement	Categorical variable denoting who shifted money from a small to a large stack in the last seconds of the one-minute decision phase	567	-	-	0	7
Last-minute reduc-tion	Categorical variable denoting who shifted money from a large to a small stack in the last seconds of the one-minute decision phase	567	-	-	0	7
Questions survived	Number of questions "survived" without losing all money	94	5.34	2.36	0	8
Amount won	Overall amount of money won after having correctly answered question number eight	94	29,521	72,491	0	500,000

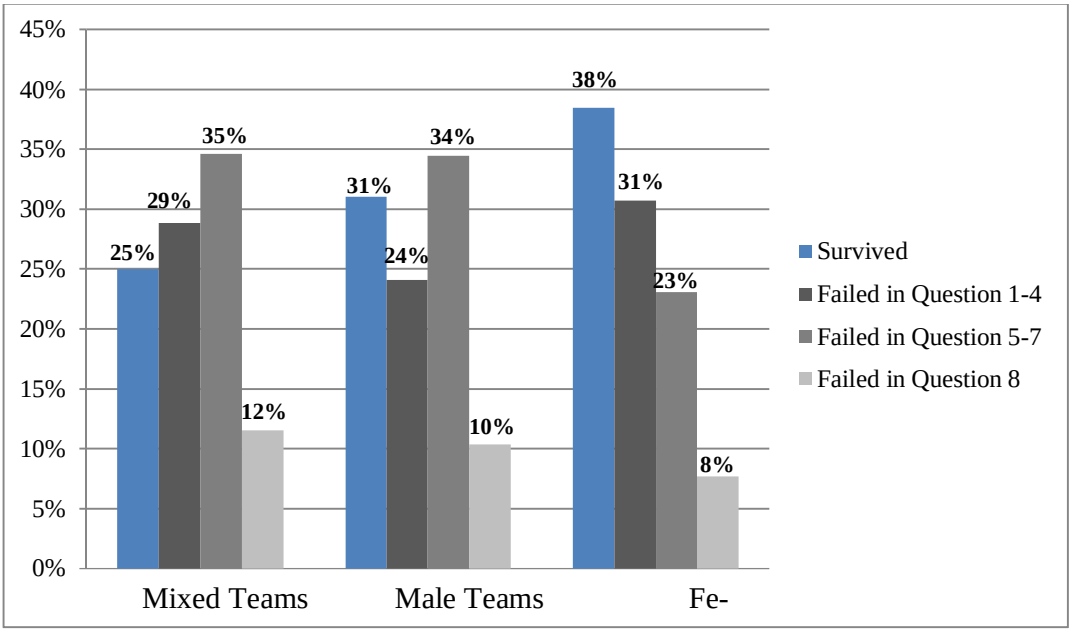
Note: Due to the limited decision choice in the last round (recall that the splitting of stakes is not allowed) these observations are omitted for a*dj. Gini*. The categorical variable *decider category* indicates whether only one male (=1) or female (=2) in a mixed team, or only one person in a single-gender male (=3) or female team (=4) has taken the decision. In these cases we refer to a non-consensual decision. If both members of a female (=5), male (=6) or mixed team (=7) agree upon a category, we refer to a consensual decision. We use the same pattern for the *decider question* and last-minute-change variables, with the latter including the possibility that no changes are made (=0).

In the following, we provide descriptive evidence on the success rates of teams, their splitting behavior as well as the decision-making process of single- and mixed-gender teams. Hereby, we particularly focus on the one-minute decision phase where candidates physically move their money onto the trapdoors related to the respective answer options.

In the first step, the teams' success rates are compared. More specifically, we examine how many of the male, female and mixed teams are able to "survive" all eight questions and whether the team composition has an effect on team performance.[98]

[98] The term "team" might be somewhat misleading as most of the literature analyzing the effect of gender diversity on team outcomes focuses on teams of three or more members (see e.g. Kashy and Kenny 2000, Horwitz and Horwitz 2007 or Bell et al. 2011 for comprehensive reviews of this literature). Therefore, "team" explicitly refers to the *pairs* (or *dyads*) of candidates competing in the quiz show.

Figure 6-2: Percentage of "surviving" and "failing" teams

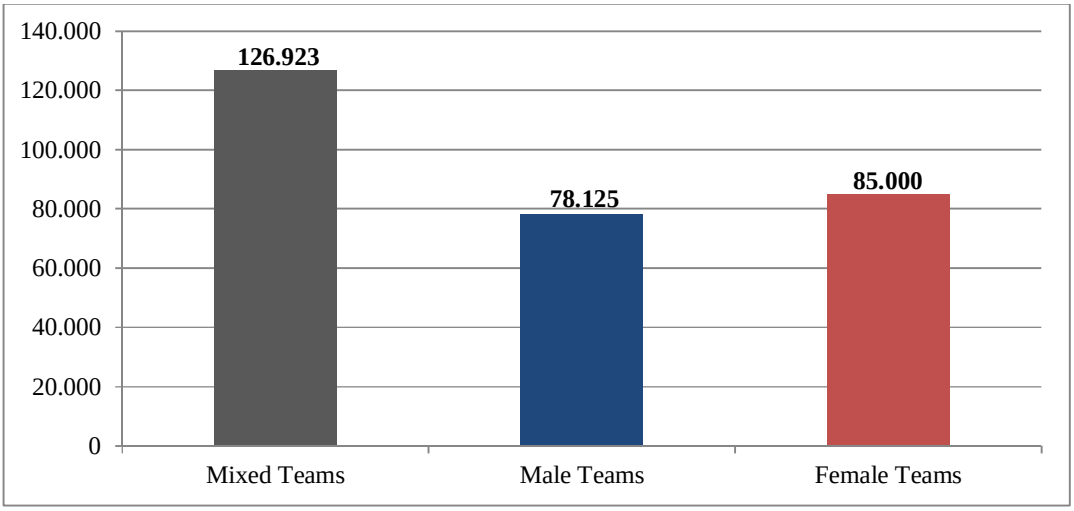


Source: Own illustration.

It appears from Figure 6-2 that female teams perform considerably better than male teams, with 38 percent of all female teams successfully completing all eight questions (as opposed to 31 percent of all male teams). Mixed teams seem to perform worst, with only every fourth team "surviving the game". Most of the mixed and male teams appear to fail in rounds 5-7, whereas the majority of female teams apparently fail in the first four rounds. Conducting a chi-squared test suggests, however, that the observed differences are statistically insignificant. The multivariate analyses discussed in the next section (we estimate a semi-parametric survival model as well as an ordered probit regression) will shed additional light on the determinants of team success.

Interestingly, in terms of earnings mixed teams fare significantly better than the previous results would have led one to suspect: Out of those teams surviving all eight questions mixed teams win on average 1.5 times the amount of female teams and 1.6 times the amount of male teams (see Figure 6-3).

Figure 6-3: Average earnings of "surviving" teams

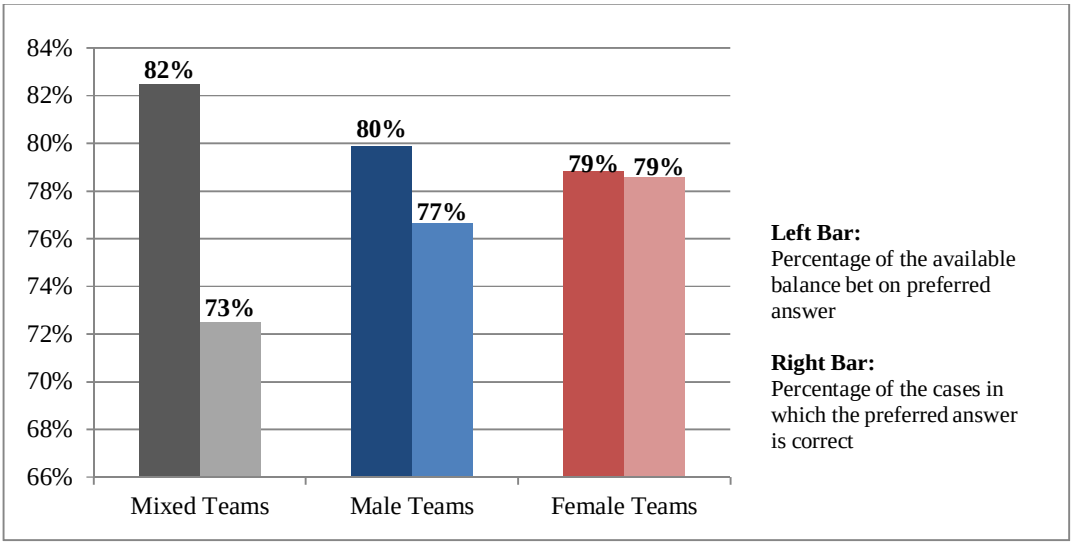


Source: Own illustration.

The results of a two-group mean-comparison test suggest that the differences in prize money between mixed and single-gender teams are statistically significant, whereas no statistically significant differences are found between male and female teams. Taking both the "survival" rates (of all) and the average earnings of (only surviving) teams into account, it seems that mixed teams split their stakes to a lesser extent (are less risk averse) than single-gender teams. In consequence, mixed pairs exit the game earlier but win higher amounts of money if they manage to survive all eight questions. The average earnings of all teams (including teams exiting the game prematurely) do not differ significantly – in a statistical sense – between the different gender compositions.

In the next step, we aim to shed light on the teams' betting patterns and success rates. Figure 6-4 below informs about how much of the available balance is bet on the preferred answer and how often this preferred answer turned out to be correct. Hence, this descriptive analysis is a good measure of the contestants' confidence. While the first column in the diagram indicates how confident the teams are of their answers, the second column shows if the teams are indeed good at evaluating themselves or not. As an illustration, mixed teams put on average 82 percent of the available balance on the preferred answer which, however, turns out to be correct in 73 percent of the cases only.

Figure 6-4: Betting patterns and success rates



Source: Own illustration.

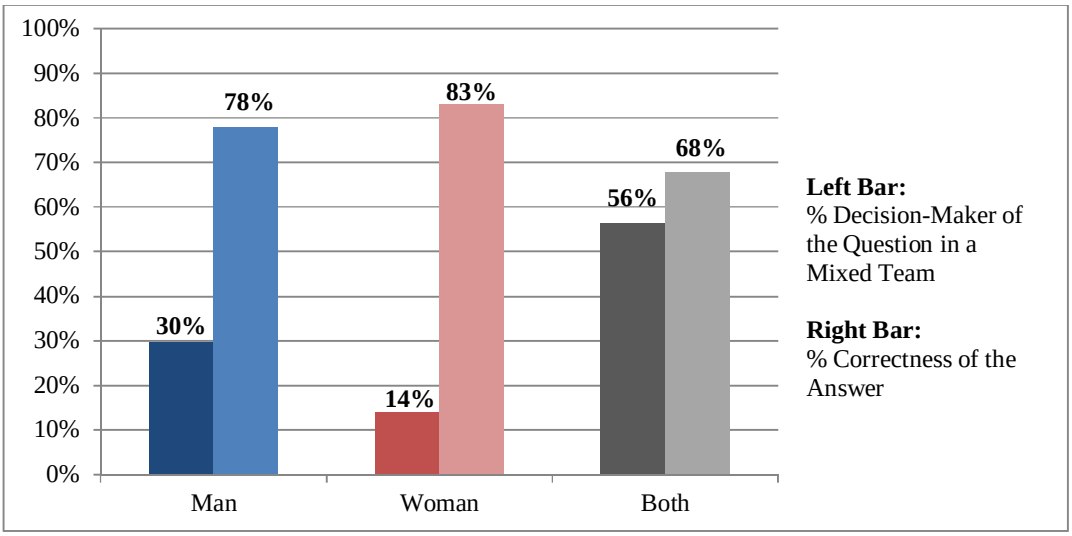
In contrast, male teams put an average portion of 80 percent of their available balance on the preferred answer which turns out to be correct in 77 percent of the cases. Female teams appear to have the best self-assessment as the percentage of money bet on the preferred answer is almost equal to their actual success rate. Although none of these differences are statistically significant, the descriptive evidence suggests that mixed teams tend to be over-confident, while single-gender teams – and particularly female teams – evaluate themselves more realistically.

In the last step, we attempt to look into the "black box" of the decision-making process. This can be done by examining the *individual* decisions within teams (rather than interpreting decisions as *team decisions*). Therefore, we closely observe the behavior of the candidates during the one-minute decision phase[99], where typical scenarios are as follows: First, both candidates appear to have a similar knowledge regarding the question (i.e., they either know the answer with a given probability or they have no idea) and decide upon their bets on a consensual basis. Second, both candidates have different opinions but nevertheless reach a consensus. Third and most interestingly, only one of the two team members takes

[99] We collect similar data regarding the selection of the question category as well as potential last-minute money shifts. However, here we are unable to find systematic (gender) differences between different team compositions.

the decision while the other person is not involved in the decision-making process.[100] The following three Figures illustrate the individual decision-making process in mixed, male and female teams.

Figure 6-5: Decision-makers in mixed teams



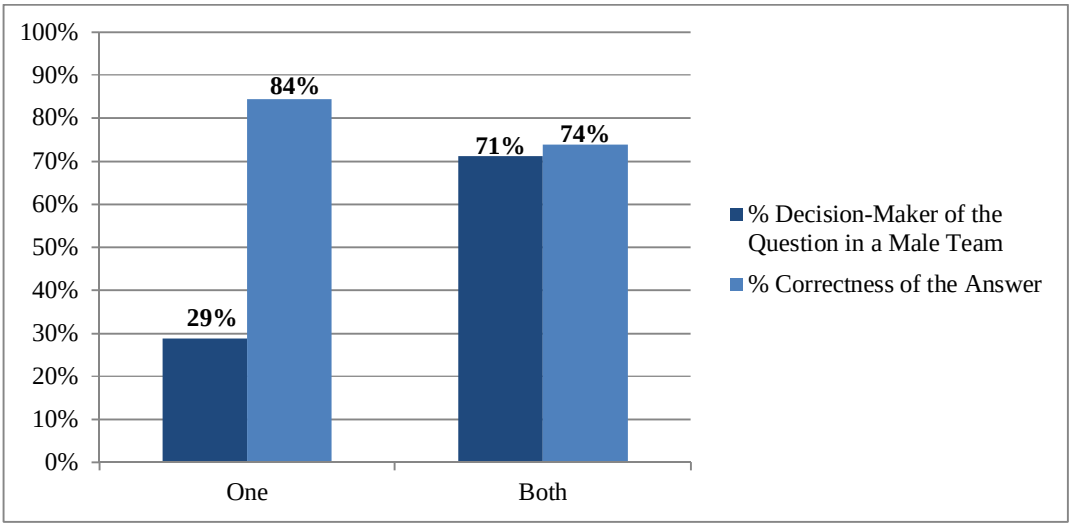
Source: Own illustration.

Regarding mixed teams, it appears that 56 percent of the questions are answered on a consensual basis. In contrast, in 44 percent of the cases only one team member makes the decision. Here, it is noticeable that men act as sole decision-makers on average twice as often as women do, even though their average success rate is slightly lower than that of women (albeit not significantly different from one another in a statistical sense). Interestingly, the likelihood of a preferred answer being correct is higher when only one team member makes the decision. This is mainly due to the fact that the majority of difficult questions in the later rounds, when both candidates are uncertain about the correct answer, are answered consensually. Given the frequency of male "solo runs" and their success rates compared to female sole decision-makers, it appears that men are overconfident when playing with a female partner.

[100] Candidates in fact communicate a lot at the beginning of the decision phase, stating how knowledgeable and confident they are. Only if one candidate clearly signals that she is confident and consequently takes the initiative, with the other candidate remaining passive, we classify this as a non-consensual decision. Note that we do not observe situations in which both candidates have totally different opinions and do not reach some form of consensus.

On the other hand, the individual choice behavior in single-gender teams is somewhat different. In male teams, 71 percent of the questions are answered on a consensual basis, whereas only 29 percent of the questions are answered non-consensually (implying that each individual solely decides in 14.5 percent of all cases).

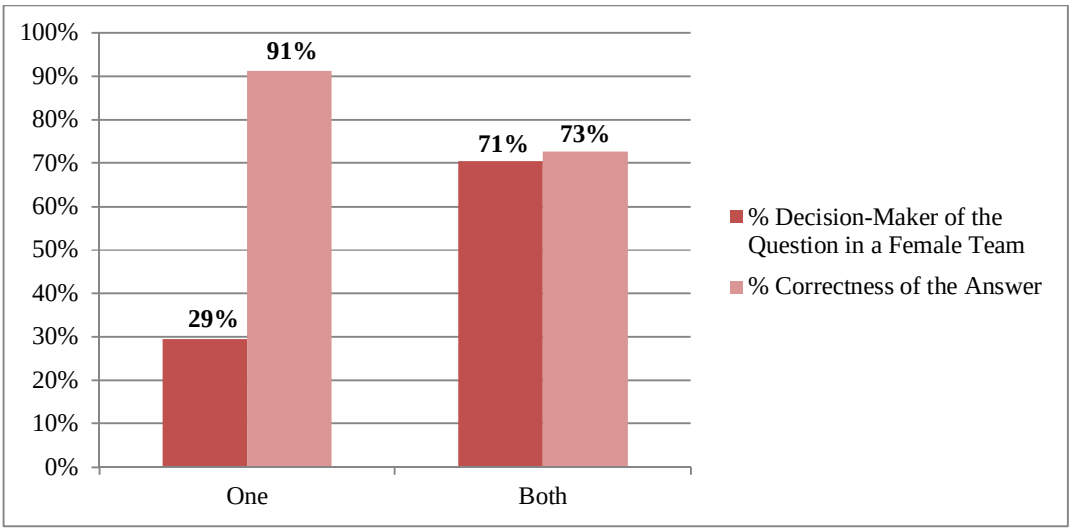
Figure 6-6: Decision-makers in male teams



Source: Own illustration.

An almost identical pattern can be observed for female teams, with exactly the same proportion of questions being answered by a sole decision-maker (and only slightly different success rates).

Figure 6-7: Decision-makers in female teams



Source: Own illustration.

When comparing the individual choice behavior in single-gender and mixed teams, a striking result is that males and females exhibit exactly the same behavior when playing in gender-homogeneous teams. In gender-heterogeneous teams, however, men act as sole decision-makers twice as often as women. This might to some extent explain why mixed teams appear to be less-risk averse in terms of their splitting behavior: Overconfident (and presumably more risk loving) men tend to influence the decision-making process more frequently than women (who do not behave differently than in female-only teams). These are of course preliminary results which have to be tested econometrically by means of multivariate analyses, as will be done in the subsequent section.

6.5 ECONOMETRIC EVIDENCE

In order to test our hypotheses that

(i) *male teams bet larger shares of their money on the answers they believe to be correct (i.e., they are less risk averse) than female teams which prefer to diversify their bets (Hrisk),*

and (ii) *men are overconfident especially when playing with women in mixed teams ($H_{overconfidence}$),*

we estimate an OLS regression as well as a probit regression, reporting marginal effects for the latter for ease of presentation. In addition to that, we estimate a semi-parametric proportional hazard model (see Cox 1972) and an ordered probit regression to analyze whether the "survival" of teams is affected by the teams' composition. The OLS model has the following functional form:

$$\text{Adj. Gini} = \beta_0 + \beta_1 \text{team composition} + \beta_2 \text{country} + \beta_3 \text{question no.} + \beta_4 \text{options chosen} + \beta_5 \text{preferred correct} + \beta_6 \text{timer stopped} + \beta_7 \text{balance} + \beta_8 \text{CumAge} + \varepsilon_i.$$

Table 6-4: Estimation results OLS regression

	Dependent variable: Adj. Gini			
Covariates	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Female team	*Reference Category*	*Reference Category*	*Reference Category*	*Reference Category*
Male team	0.064** [0.032]	0.053* [0.057]	0.056* [0.049]	0.075** [0.017]
Mixed team	0.045* [0.063]	0.043* [0.063]	0.045* [0.051]	0.068** [0.014]
Country (UK)	*Reference Category*	*Reference Category*	---	*Reference Category*
Country (GER)	-0.032 [0.289]	-0.022 [0.345]	---	-0.036 [0.392]
Country (CH)	-0.023 [0.307]	-0.019 [0.363]	---	-0.029 [0.339]
Question no.	-0.050*** [0.000]	-0.051*** [0.000]	-0.051*** [0.000]	---
Options chosen	-0.326*** [0.000]	-0.320*** [0.000]	-0.318*** [0.000]	---
Preferred correct	0.061** [0.011]	0.067*** [0.006]	0.066*** [0.006]	0.230*** [0.000]
Timer stopped	0.037** [0.033]	0.040** [0.025]	0.045** [0.011]	0.263*** [0.000]
Balance (in 10,000)	0.005 [0.193]	0.005 [0.320]	0.004 [0.440]	0.022*** [0.000]
CumAge	0.001 [0.341]	---	---	---
Constant	1.327*** [0.000]	1.317*** [0.000]	1.310*** [0.000]	0.279*** [0.000]
Observations	530	530	530	530
F	131.6	149.6	189.1	49.3
R^2	0.708	0.694	0.693	0.317

Note: *** p<0.01, ** p<0.05, * p<0.1. P-values are reported in brackets. All model specifications are estimated using robust standard errors clustered on team level. Model 1 includes all variables, whereas some controls are omitted in models 2-4.

Most of the above results are straightforward and rather unsurprising. For example, the further the candidates advance in the show (and, thus, the more difficult the questions become), the stronger the diversification of their bets. Teams deciding to stop the timer before the time limit is up seem to be quite certain about the answer and tend to bet more money on the preferred answer option. Those candidates who know the correct answer for sure most likely do not split their stakes at all. We can, to some extent, control for such

cases by means of the covariate *preferred correct*. Although we observe a strong (and inverse) correlation of -0.75 between the teams' splitting behavior (*adj. Gini*) and the number of answer options they bet on (*options chosen*) – which explains the relatively high value of R^2 – we include both controls as the latter serves as an additional proxy for risk aversion. A particularly interesting result, on the other hand, is that male teams appear to split their money to a lesser extent than female teams, all other things being equal. We can thus support our hypothesis that male teams are less risk averse than female teams. Moreover, it appears that even mixed teams are significantly more risk loving than female teams (while mixed and male teams do not significantly differ from each other). These results are robust across all model specifications.

Against the backdrop of the previous finding that males in mixed teams appear to act as sole decision-makers twice as often as in single-gender teams, a plausible explanation could be that male "risk taking" outweighs female "safeguarding" in heterogeneous teams. In the next step, we aim to assess the costs and benefits of such behavior. Among other things, we test whether the dominant behavior of men in mixed teams is advantageous for the teams' success or whether men are simply overconfident. Therefore, we estimate a probit regression of the following form:

$$\text{Preferred correct} = \beta_0 + \beta_1 \text{non-consensual: man in mixed team} + \beta_2 \text{non-consensual: woman in mixed team} + \beta_3 \text{non-consensual: woman in single-gender team} + \beta_4 \text{non-consensual: man in single-gender team} + \beta_5 \text{team composition} + \beta_6 \text{question no.} + \beta_7 \text{balance} + \beta_8 \text{failure} + \beta_9 \text{success} + \beta_{10} \text{timer stopped} + \beta_{11} \text{CumAge} + \varepsilon_i,$$

where *preferred correct* is our binary dependent variable reflecting a team's success in a given round. The regression results can be found in the following Table 6-5. For interpretation purposes, we report marginal effects at the means of the covariates which show a discrete change from the base level for categorical variables and an infinitesimal change for continuous variables.

Table 6-5: Marginal effects after probit regression

	Dependent variable: Preferred correct	
Covariates	**Model 1**	**Model 2**
Non-consensual: Man in mixed team	0.093* [0.061]	0.094* [0.061]
Non-consensual: Woman in mixed team	0.118* [0.051]	0.119* [0.053]
Non-consensual: Woman in single-gender team	0.117** [0.044]	0.130** [0.013]
Non-consensual: Man in single-gender team	0.024 [0.678]	0.016 [0.784]
Consensual decision	*Reference Category*	
Male team	0.070 [0.112]	0.076* [0.086]
Female team	0.083* [0.075]	0.084* [0.067]
Mixed team	*Reference Category*	
Question no.	-0.055*** [0.005]	-0.054*** [0.007]
Balance (in 10,000)	-0.002 [0.100]	-0.002 [0.126]
Failure	-0.164*** [0.006]	-0.164*** [0.005]
Success	0.003 [0.875]	0.003 [0.861]
Timer stopped	0.227*** [0.000]	0.226*** [0.000]
CumAge	0.001 [0.184]	---
Observations	567	567
Wald Chi^2	74.20***	70.78***
Pseudo R^2	0.142	0.140

Note: *** p<0.01, ** p<0.05, * p<0.1. P-values are reported in brackets. Both model specifications are estimated using robust standard errors clustered on team level.

Similar to the previous estimations, most of the coefficients have the expected direction: Teams which decide to stop the timer give significantly more accurate predictions than teams exploiting the one-minute time limit. This does of course not imply that stopping the timer increases the likelihood of correctly answering the question per se. It rather seems that the causality works in the opposite direction and that *timer stopped* is a signal of a

team's certainty. Perhaps surprisingly, the number of correctly answered questions in the previous rounds does not affect the accuracy of the answer in the present round. *Failure*, on the other hand, has a statistically significant and negative effect, suggesting that teams who have lost the lion's share of their money in the previous round(s) are more likely to (continue to) incur losses than successful teams. Again, with the increasing difficulty of questions in the later rounds candidates face decreasing chances of correctly answering questions.

Quite remarkably, single-gender teams seem to have an eight percentage points (PPS) higher likelihood of a correct preferred answer than mixed teams. With regard to individual decision making within the different team compositions it seems that if women act as sole decision-makers they have a higher success rate than men. This difference amounts to almost three PPS, corresponding to up to £30,000 difference in earnings (depending on the remaining budget and the splitting behavior). It should be noted, however, that part of these gender differences could be due to selection effects: Since women are found to "step forward" less frequently than men it could well be that only those women who are very certain about their answer take the initiative. Despite that, it seems that men are overconfident when playing with a female partner as their success rate does not justify the frequency of their "solo runs".

In the last step, we analyze whether the "survival" of a team is affected by its composition and/or the individual decision making. We estimate a semi-parametric Cox proportional hazard model to test whether any of these covariates has an effect on a team's probability of "surviving" the current round, conditional on having "survived" the previous round. The ordered probit model, on the other hand, analyzes the teams' "survival" in a more general sense: Here, the dependent variable reflects the number of questions "survived" at the end of the game. This variable is ordinal in nature and ranges from 0 (a team fails in the first round) to 8 (identifying the winning teams). However, irrespective of the estimation technique, we are unable to find statistically significant differences between mixed, male and female teams. Nor do the different scenarios in the decision phase (e.g. male decision-maker in mixed teams, consensual decision in female teams, etc.) have a statistically significant impact on team success.[101] With respect to these results, it should be noted that due to the limited number of observations at the aggregate level (n = 94 teams) we cannot in-

[101] In the interest of brevity, the results of these estimations are not included in this paper but are available from the authors upon request.

clude all covariates and controls. Particularly some of the categorical variables are omitted due to an insufficient number of observations.

Summarizing the results it can be stated that although male and mixed teams appear to be less risk averse than female teams – while at the individual level men tend to be overconfident when playing with a female team partner – these differences do not *systematically* influence the teams' overall success. Yet, with respect to the descriptive statistics presented in the previous section, there is at least weak evidence that female teams more often "survive" all eight questions (although their splitting behavior impedes higher earnings). At the same time, the opposite seems to be true for mixed teams, which are found to fail more often but generate, on average, higher earnings *if* they manage to successfully complete all eight rounds. Generally speaking, it appears that as soon as a man is part of the team, larger shares of the team's budget are bet on the preferred answer, indicating that men are more inclined towards risk than women.

6.6 CONCLUSIONS AND MANAGERIAL IMPLICATIONS

Drawing on data from a TV quiz show, this paper provides empirical evidence for gender differences in high-stakes risk taking and overconfidence. More specifically, among the pairs competing for a prize money of up to one million pounds female teams tend to split their stakes more often and thus appear to be more risk averse than teams consisting of at least one male candidate. Looking at individual decision making within teams it seems that men and women do not significantly differ from each other as long as they play in single-gender teams. In mixed teams, however, men are found to take the initiative twice as often as in single-gender teams (while women do not behave differently than men in homogeneous teams). Since male "solo-runs" are associated with a lower success rate compared to female decision-makers in mixed teams, we attribute this result to the existence of male overconfidence. The finding that men tend to overestimate their own ability when facing a female partner is in line with Gerdes and Gränsmark (2010) who show that even among expert chess players, men ceteris paribus prefer more aggressive (but often suboptimal) opening strategies when playing against a female opponent.

Of course, one has to be careful regarding the generalization of these results because of a potential sample selection bias: Game show contestants most likely differ in their characteristics from the general population. Although we have no information on the criteria upon which the candidates are selected and invited to the show, it is reasonable to assume that

these individuals are, on average, more extrovert, confident and risk loving than individuals in a randomly chosen sample. On the other hand, since we are able to find persistent gender differences even among highly selected (and presumably less diverse) individuals, we rather underestimate the effects with respect to the remaining population. A drawback of this study is the limited number of observations at the aggregate (i.e., team) level. In order to conduct a more conclusive survival analysis with all variables of interest, more observations are needed. In particular, a larger number of female teams would enrich the dataset and allow for a comprehensive analysis of the determinants of team success. However, considering the already mentioned rule changes of the game show (see chapter 6.3) it is no longer possible to make additional *and comparable* observations of the "traditional" format. Nevertheless, the new format (which features teams of four candidates and a lottery after the penultimate question) opens avenues for future research. In particular, one could analyze the behavior of both men and women in single-gender, (fe)male-dominated as well as equally represented teams.

Our findings have important managerial implications as they highlight situations in which individual decisions might lead to outcomes that are suboptimal for the group. One mechanism to reduce overconfidence and to encourage less confident individuals to actively engage in the decision-making process in the business context is to measure and evaluate the impact of individual decisions. Objective (and relative) performance measures that are made accessible to all employees could serve as an incentive for inherently risk averse but indeed productive agents. This would of course presuppose manageable monitoring and measurement costs that need to be amortized through efficiency gains.

7 SUMMARY AND OUTLOOK

This dissertation provides an economic analysis of individual behavior in competitive environments, with a particular focus on real-life tournaments. Drawing on data from various sports contests as well as a high-stakes TV quiz show, the present work makes several (innovative) contributions to the existing research by examining very specific and so far neglected (or even completely ignored) labor markets and peculiar competitive situations. In the first part of this work (chapters 2 and 3), the career determinants of professional ski jumpers and – at best semi-professional – referees who are active in the highest divisions of German association football are analyzed. Part 2 (chapters 4-6) focuses on gender differences in competitive environments. These competitive environments range from long-distance endurance races to professional European team sports as well as situations in which pairs of candidates have to answer knowledge-based questions in a limited amount of time and in the presence of exceptionally high stakes.

The empirical evidence presented in the respective chapters is mostly in line with the related literature. Nevertheless, some of the results are rather surprising and might be attributed to the peculiarities of the competition. Professional ski jumpers, for example, seem to have, on average, similar career lengths as professional athletes in other – "equally demanding" – sports. This is surprising insofar as ski jumping is still considered a "niche" sport with comparably little "money" involved. On the face of it, one would expect that only the top athletes, who regularly perform well in the competitions and benefit from both on-site prize money and additional income from lucrative endorsement contracts, are able to "survive" on the circuit over a longer period of time. Less talented athletes, on the other hand, should be better off falling back on their "outside options". That is, rational individuals who fail to succeed in a sports career should choose a more secure and permanent employment (perhaps in a different industry), concomitant with a guaranteed income. However, the opposite seems to be true. The fact that even less successful athletes tend to stay in the "business" for an extended period of time is most probably due to the peculiar nature of the sports industry, which will be elaborated in the following.

One of the distinctive features of professional sports competitions is that these usually represent winner-takes-all markets where marginal differences in talent translate into considerable differences in earnings (Rosen 1981). This so-called superstar theory has been complemented by Adler (1985) who identifies an individual's popularity (rather than mere tal-

ent) as an integral component of success. The "popularity factor", in turn, fosters the already mentioned additional earnings opportunities in the form of endorsement contracts, television appearances and the like. Hence, unlike "traditional" labor markets where marginal differences in talent, ability or productivity lead to significant differences in earnings at the top level (e.g. CEOs earn significantly more than employees at the top management level) that are, however, limited to direct income such as fixed salaries and bonus payments, the labor market for individual professional athletes appears to offer *two* (complementary) "winner's prizes": a) a significantly higher regular income from e.g. prize money and, b) additional income from the commercial exploitation of the individual athlete's popularity. These positive network externalities of popularity have been facilitated by the ongoing professionalization and medialization of sports (see, inter alia, Borghans and Groot 1998; Franck and Nüesch 2012; Deutscher et al. 2012).[102] Another peculiarity of the sports industry is that professional athletes have usually made investments into very specific human capital which prohibits them from selecting into an alternative and perhaps related field of work. Thus, the transferability of skills appears to be relatively low among professional athletes. As an example, a professional ski jumper is likely to fail or at least face considerable income losses when attempting to pursue a career as a downhill or cross-country skier. An engineer, on the other hand, may even benefit from a professional change into e.g. the field of financial analysis.

This might explain why even less talented individuals who are active in a niche sport strive to pursue their sporting career for as long as possible: The option value of a potential future career breakthrough heavily outweighs the expected income from a supposedly more "secure" outside option. This is particularly true as (former) "superstars" in the ski jumping World Cup are found to be, ceteris paribus, less threatened by competitive pressure and, thus, can expect significantly longer careers than athletes who have failed to win a major individual title so far. Another idiosyncratic feature of the ski jumping World Cup is that the level of competitive pressure varies dramatically from one national federation to the other and is, at the same time, easily observable and quantifiable for the econometrician. On the basis of these data, it was possible to examine individual professional careers in the same industry but at different "employers" (i.e., national federations). Although it is difficult to generalize the obtained results to other industries, they are of pivotal importance for

[102] It should be noted that the superstar phenomenon is not exclusively "reserved" for professional athletes. Among others, it is particularly the music industry that offers very similar conditions for "superstardom", as prominently indicated by e.g. Krueger (2005) as well as Connolly and Krueger (2006).

the actors in this specific market. Athletes holding a dual citizenship, for example, should carefully weigh the presumably better training opportunities in a rather strong national federation against the potential career setbacks resulting from the increased competitive pressure. After all, it could be more beneficial for the individual athlete to start for a weaker national federation.[103]

Another very specific institutional environment that has been investigated in the course of this work is the labor market for referees in German association football. A perhaps unique characteristic of referees active in any major sports is that despite their comparably low (and often solely match-based) income, their performance is frequently monitored by tens of thousands of spectators in the stadia and, not to forget, an almost unrestricted number of TV viewers. Given this close to perfect monitoring, a referee's reputation is likely to be heavily impacted by his performance in one direction or the other. These reputation effects, in turn, might serve as an ideal substitute for the provision of performance based short-term incentives which, in effect, appear to be nonexistent in the case of German football referees. On the other hand, carefully designed long-term incentives in the form of individual promotions for the best performing referees are found to be a potent measure to foster consistently good refereeing performance. Along these lines, the hypotheses and economic predictions derived from tournament theory are perfectly suited to explain the nexus between incentives, performance and individual career outcomes even in a very unique competitive environment.

The second part of this work has focused on gender differences in (again) very specific competitive environments. Although the distinctive chapters shed light on different aspects in the field of gender economics, the empirical evidence broadly supports the notion that gender differences in competition cannot be (solely) explained by inherent differences but are influenced by the institutional setting. As has been demonstrated on the example of long-distance endurance athletes, changing socio-cultural conditions may well lead to a reduction of the performance gender gap over time. These results are in line with Frick

[103] Note that only one athlete in our sample in fact switches citizenship *during* his active career. It may well be that more athletes change their nationality *before* turning professional. Since we only observe athletes who have managed to win World Cup points, information on "early switchers" is not available. Yet, similar considerations are likely to be made by e.g. professional football players possessing a dual nationality. In particular "marginal" players (i.e., players who have uncertain prospects of qualifying for a strong national team) could improve their chances by choosing to play for the presumably weaker national team. Since a player's market value is positively correlated with the number of international caps (see e.g. Frick 2007 as well as Battré and Höhmann 2011), opting for the weaker nation could maximize the expected individual income in the long term.

(2011a, 2011b). What is new in the present context is that particularly less talented, recreational athletes who finish well beyond the "money ranks" seem to respond to changing socio-cultural conditions, while (direct) monetary incentives cannot explain this behavior.

Yet, looking at professional team sports, incentives do seem to play a major role in attracting (female) talent. As the underlying research indicates, typically male-dominated organizations such as professional football teams appear to be relatively homogeneous in the case of men's leagues. In contrast, women's leagues are characterized by a considerably more uneven distribution of talent. Hence, women's football leagues are less balanced and supposedly less appealing to the fans than men's leagues. Using Rottenberg's (1956) uncertainty of outcome hypothesis to explain the enormous discrepancy between the demands for men's football on the one hand and for women's football on the other, would be evidently too simplistic and misleading. Nevertheless, the obvious disparity in demand causes a tremendous gender pay gap that deters talented females to pursue a professional career in this specific environment. It will be interesting to see whether professional female football players become relatively more competitive in response to perhaps increasing incentives in the not too distant future. Subsequent research should also focus on other, more gender equal sports. A first attempt has been made in this work by focusing on the German handball league, with the results pointing in the expected direction: A less pronounced gender pay gap is associated with similarly balanced men's and women's competitions. Given the limited scope of the analysis, it is (as of yet) not possible to generalize these results. Therefore, further research is required to test the robustness of the results.

Finally, the empirical evidence on gender differences in high-stakes decision making suggests that team composition has a statistically significant impact on individual behavior. Most notably, men appear to be overly prone to act as sole decision-makers when "cooperating" with a female team partner. Controlling for the outcomes of each decision, it has been argued that male overconfidence is likely to explain the observed gender differences in mixed teams. Moreover, in single-gender environments women are found to be more risk averse than men (albeit these differences do not seem to affect the teams' expected profit). These results support the general tenor that men and women (still) differ in their competitive behavior. An interesting insight emerging from the present research is that the commonly acknowledged behavioral differences seem to hold true even for highly selected individuals who are presumably less diverse with respect to personality traits such as self-awareness, self-confidence, risk behavior and the like.

In conclusion, the evidence and implications discussed in this work provide exciting insights in the fields of personnel, labor and sports economics and point out avenues for future research. Given the peculiar nature of the institutional settings examined in the respective analyses, a generalization of the results is undeniably difficult. Hence, one could object that the present investigations suffer from the same shortcomings as "insider econometric" studies (see chapter 1). On the other hand, this type of research offers an innovative and complementary rather than substitutable approach that allows examining whether generally accepted results and theoretical considerations can also explain individual behavior and economic relationships in very specific environments.

REFERENCES

Adler, M. (1985): Stardom and Talent, American Economic Review, 75(1), 208-212.

Agnew, J. R., L. R. Anderson, J. R. Gerlach and L. R. Szykman (2008): Who Chooses Annuities? An Experimental Investigation of the Role of Gender, Framing and Defaults, American Economic Association, 98(2), 418-422.

Akaike, H. (1974): A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, 19(6), 716-723.

Akerlof, G. A. (1980): A Theory of Social Custom, of Which Unemployment May Be One Consequence, The Quarterly Journal of Economics, 94(4), 749-775.

Alchian, A. A. (1988): Promotions, Elections and Other Contests: Comment, Journal of Institutional and Theoretical Economics, 144, 91-93.

Alicke, M. D. (1985): Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives, Journal of Personality and Social Psychology, 49(6), 1621-1630.

Andersen, S., E. Bulte, U. Gneezy and J. A. List (2008): Do Women Supply More Public Goods than Men? Preliminary Experimental Evidence from Matrilineal and Patriarchal Societies, American Economic Review, 98(2), 376-381.

Andersen, S., S. Ertac, U. Gneezy, J. A. List and S. Maximiano (2013): Gender, Competitiveness and Socialization at a Young Age: Evidence from a Matrilineal and a Patriarchal Society, Review of Economics and Statistics, 95(4), 1483-1443.

Andreoni, J. and A. A. Payne (2011): Is Crowding Out Due Entirely to Fundraising? Evidence from a Panel of Charities, Journal of Public Economics, 95(5), 334-343.

Andrews, D. and A. Leigh (2009): More Inequality, Less Social Mobility, Applied Economics Letters, 16(15), 1489-1492.

Antonioni, P. and J. Cubbin (2000): The Bosman Ruling and the Emergence of a Single Market in Soccer Talent, European Journal of Law and Economics, 9(2), 157-173.

Antonovics, K., P. Arcidiacono and R. Walsh (2005): Games and Discrimination Lessons from the Weakest Link, Journal of Human Resources, 40(4), 918-947.

Antonovics, K., P. Arcidiacono and R. Walsh (2009): The Effects of Gender Interactions in the Lab and in the Field, Review of Economics and Statistics, 91(1), 152-162.

Arch, E. (1993): Risk-Taking: A Motivational Basis for Sex Differences, Psychological Reports, 73(3), 6-11.

Åslund, O. and O. Nordström Skans (2012): Do Anonymous Job Application Procedures Level the Playing Field? Industrial and Labor Relations Review, 65(1), 82-107.

Atkinson, S. and J. Tschirhart (1986): Flexible Modelling of Time to Failure in Risky Careers, Review of Economics and Statistics, 68, 558-566.

Attali, Y., Z. Neeman and A. Schlosser (2011): Rise to the Challenge or not Give a Damn: Differential Performance in High vs. Low Stakes Tests, Discussion Paper 5693, Institute for the Study of Labor, Bonn.

Balafoutas, L. and M. Sutter (2010): Gender, Competition and the Efficiency of Policy Interventions, Discussion Paper 4955, Institute for the Study of Labor, Bonn.

Balfour, A. and P. Porter (1991): The Reserve Clause in Professional Sports: Legality and Effect on Competitive Balance, Labor Law Journal, 42, 8-18.

Balmer, N. J., A. M. Nevill and A. M. Lane (2005): Do Judges Enhance Home Advantage in European Championship Boxing? Journal of Sports Sciences, 23(4), 409-416.

Balmer, N. J., A. M. Nevill and A. M. Williams (2001): Home Advantage in the Winter Olympics (1908-1998), Journal of Sports Sciences, 19(2), 129-139.

Balmer, N. J., A. M. Nevill and A. M. Williams (2003): Modelling Home Advantage in the Summer Olympic Games. Journal of Sports Sciences, 21(6), 469-478.

Bandiera, O., I. Barankay and I. Rasul (2005): Social Preferences and the Response to Incentives: Evidence from Personnel Data, Quarterly Journal of Economics, 120(3), 917-962.

Bandiera, O., I. Barankay and I. Rasul (2007): Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment, Quarterly Journal of Economics, 122(2), 729-773.

Bandiera, O., I. Barankay and I. Rasul (2009): Social Connections and Incentives in the Workplace: Evidence from Personnel Data, Econometrica, 77(4), 1047-1094.

Bartlett, M. S. (1937): Properties of Sufficiency and Statistical Tests, Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences, 160, 268-282.

Battré, M. and A. Höhmann (2011): Die Bedeutung der letzten Karrierestation für die Entlohnung von Fußballspielern, Sport and Society, 2, 124-153.

Becker, G. S. (1962): Investment in Human Capital: A Theoretical Analysis, Journal of Political Economy, 70(5), 9-49.

Becker, G. S. (1968): Crime and Punishment: An Economic Approach, Journal of Political Economy, 76(2), 169-217.

Becker, G. S. (1971): The Economics of Discrimination, 2nd ed., Chicago: University of Chicago Press.

Becker, G. S. (1993): Nobel Lecture: The Economic Way of Looking at Behavior, Journal of Political Economy, 101(3), 385-409.

Becker, G. S. (2009): Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, 3rd Edition, University of Chicago Press.

Becker, G. S. and K. M. Murphy (2000): Social Economics: Market Behaviour in a Social Environment, Harvard University Press, Cambridge, MA.

Beetsma, R. M. W. J. and P. C. Schotman (2001): Measuring Risk Attitudes in a Natural Experiment: Data from the Television Game Show Lingo, The Economic Journal, 111, 821-848.

Bell, S. T., A. J. Villado, M. A. Lukasik, L. Belau and A. L. Briggs (2011): Getting Specific about Demographic Diversity Variable and Team Performance Relationships: A Meta-analysis, Journal of Management, 37(3), 709-743.

Benoît, J. P. and J. Dubra (2011): Apparent Overconfidence, Econometrica, 79(5), 1591-1625.

Berentsen, A. (2002): The Economics of Doping, European Journal of Political Economy, 18(1), 109-127.

Berkson, J. and R. P. Gage (1950): Calculation of Survival Rates for Cancer. Proceedings of the Staff Meetings, Mayo Clinic, 25(11), 270-286.

Bernheim, B. D. (1994): A Theory of Conformity, Journal of Political Economy, 841-877.

Berrebi, C. (2007): "Evidence about the Link Between Education, Poverty and Terrorism among Palestinians", Peace Economics, Peace Science and Public Policy, 13(1), Article 2.

Bird, E. J. and G. G. Wagner (1997): Sport as a Common Property Resource: A Solution to the Dilemmas of Doping, Journal of Conflict Resolution, 41(6), 749-766.

Blodget, H. (2009): Jack Welch: How to Kick Ass in These Tough Times, Business Insider. [Online] Available at: http://www.businessinsider.com/henry-blodget-jack-welch-how-to-kick-ass-when-times-are-tough-2009-5.

Bloom, N. and J. Van Reenen (2011): Human Resource Management and Productivity, Handbook of Labor Economics, 4, Part B, 1697-1767.

Blossfeld, H. P. and G. Rohwer (2002): Techniques of Event History Modeling: New Approaches to Causal Analysis, Mahwah, NJ: Lawrence Erlbaum Associates.

Boeri, T. and B. Severgnini (2011): Match Rigging and the Career Concerns of Referees, Labour Economics, 18(3), 349-359.

Bombardini, M. and F. Trebbi (2012): Risk Aversion and Expected Utility Theory: An Experiment with Large and Small Stakes, Journal of the European Economic Association, 10(6), 1348-1399.

Booth, A. and P. Nolen (2012): Choosing to Compete: How Different Are Girls and Boys? Journal of Economic Behavior and Organization, 81(2), 542-555.

Borenstein, S. (1989): Hubs and High Fares: Dominance and Market Power in the U.S. Airline Industry, Rand Journal of Economics, 20, 344-365.

Borenstein, S. and N. Rose (1994): Competition and Price Dispersion in the U.S. Airline Industry, Journal of Political Economy, 102(4), 653-683.

Borghans, L. and L. Groot (1998): Superstardom and Monopolistic Power: Why Media Stars Earn More than Their Marginal Contribution to Welfare, Journal of Institutional and Theoretical Economics, 154(3), 546-572.

Borooah, V. K. (2002): Logit and Probit: Ordered and Multinomial Models, Thousand Oaks, CA: Sage.

Boyden, N. B. and J. R. Carey (2010): From One-and-Done to Seasoned Veterans: A Demographic Analysis of Individual Career Length in Major League Soccer, Journal of Quantitative Analysis in Sports, 6(4), 82-98.

Boyko, R. H., A. R. Boyko and M. G. Boyko (2007): Referee Bias Contributes to Home Advantage in English Premiership Football, Journal of Sports Sciences, 25(11), 1185-1194.

Bray, S. and A. Carron (1993): The Home Advantage in Alpine Skiing, Australian Journal of Science and Medicine in Sport, 25, 76-81.

Brinig, M. F. (1995): Does Mediation Systematically Disadvantage Women? William and Mary Journal of Women and the Law, 2(1), 1-34.

Brocas, I. and J. D. Carrillo (2004): Do the "Three-Point Victory" and "Golden Goal" Rules Make Soccer More Exciting? Journal of Sports Economics, 5, 169-185.

Brody, L. R. (1993): On Understanding Gender Differences in the Expression of Emotion, in: Ablon, S. L., D. Brown, E. J. Khantzian and J. E. Mack (Eds.): Human Feelings: Explorations in Affect Development and Meaning, Hillsdale, NJ: Analytic Press, 87-121.

Brooks, R., R. Faff, D. Mulino and R. Scheelings (2009): Deal or No Deal, That is the Question: The Impact of Increasing Stakes and Framing Effects on Decision-Making under Risk, International Review of Finance, 9(1-2), 27-50.

Brown, M. B. and A. B. Forsythe (1974): Robust Tests for the Equality of Variances, Journal of the American Statistical Association, 69(346), 364-367.

Bruni, L. and R. Sugden (2007): The Road not Taken: How Psychology was Removed from Economics and How it Might Be Brought Back, The Economic Journal, 117(516), 146-173.

Bryson, A., B. Buraimo and R. Simmons (2011): Do Salaries Improve Worker Performance? Labour Economics, 18(4), 424-433.

Buraimo, B. and R. Simmons (2008): Do Sports Fans Really Value Uncertainty of Outcome? Evidence from the English Premier League, International Journal of Sport Finance, 3(3), 146-155.

Buraimo, B., D. Forrest and R. Simmons (2010): The 12th Man? Refereeing Bias in English and German Soccer, Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(2), 431-449.

Buraimo, B., G. Migali and R. Simmons (2012a): Corruption Does not Pay: An Analysis of Consumer Response to Italy's Calciopoli Scandal, Conference Paper for the XXIV Conferenza Economia Informale, Evasione Fiscale e Corruzione, Pavia.

Buraimo, B., R. Simmons and M. Maciaszczyk (2012b): Favoritism and Referee Bias in European Soccer: Evidence from the Spanish League and the UEFA Champions League, Contemporary Economic Policy, 30(3), 329-343.

Butler, M. R. (1995): Competitive Balance in Major League Baseball, American Economist, 39, 46-52.

Buzzacchi, L., S. Szymanski and T. M. Valetti (2001): Static versus Dynamic Competitive Balance: Do Teams Win More in Europe or in the USA? Economics Group Discussion Paper Series, 03/ 2001, London: Imperial College Management School.

Buzzacchi, L., S. Szymanski and T. M. Valetti (2003): Equality of Opportunity and Equality of Outcome: Open Leagues, Closed Leagues and Competitive Balance, Journal of Industry, Competition and Trade, 3, 167-186.

Cairns, J., N. Jennett and P. J. Sloane (1986): The Economics of Professional Team Sports: A Survey of Theory and Evidence, Journal of Economic Studies, 13(1), 1-80.

Camerer, C. and D. Lovallo (1999): Overconfidence and Excess Entry: An Experimental Approach, American Economic Review, 89(1), 306-318.

Cameron, A. C. and P. K. Trivedi (2009): Microeconomics Using Stata, College Station, TX: Stata Press.

Cárdenas, J. C., A. Dreber, E. von Essen and E. Ranehill (2012): Gender Differences in Competitiveness and Risk Taking: Comparing Children in Colombia and Sweden, Journal of Economic Behavior and Organization, 83(1), 11-23.

Cason, T. N., W. A. Masters and R. M. Sheremeta (2010): Entry into Winner-Take-All and Proportional-Prize Contests: An Experimental Study, Journal of Public Economics, 94(9-10), 604-611.

Cazorla, G., L. Léger, T. Olds and G. Tomkinson (2006): Worldwide Variation in the Performance of Children and Adolescents: An Analysis of 109 studies of the 20-m Shuttle Run Test in 37 Countries, Journal of Sports Sciences, 24(10), 1025-1038.

Charness, G. and U. Gneezy (2012): Strong Evidence for Gender Differences in Risk Taking, i98uJournal of Economic Behavior and Organization, 83, 50-58.

Cheuvront, S. N., R. Carter, K. C. DeRuisseau and R. J. Moffatt (2005): Running Performance Differences between Men and Women: An Update, Sports Medicine, 35(12), 1017-1024.

Cleves, M. A., W. W. Gould and R. G. Gutierrez (2008): An Introduction to Survival Analysis Using Stata, College Station, TX: Stata Press.

Coate, D. and D. Robbins (2001): The Tournament Careers of Top-Ranked Men and Women Tennis Professionals: Are the Gentlemen More Committed than the Ladies? Journal of Labor Research, 22(1), 185-193.

Coffey, B. and M. T. Maloney (2010): The Thrill of Victory: Measuring the Incentive to Win, Journal of Labor Economics, 28(1), 87-112.

Connelly, B. L., L. Tihanyi, T. R. Crook and K. A. Gangloff (2014): Tournament Theory: Thirty Years of Contests and Competitions, Journal of Management, 40(1), 16-47.

Connolly, M. and A. B. Krueger (2006): Rockonomics: The Economics of Popular Music, Handbook of the Economics of Art and Culture, 1, 667-719.

Cotton, C., F. McIntyre and J. Price (2010): The Gender Gap Cracks under Pressure: A Detailed Look at Male and Female Performance Differences during Competitions, Working Paper 16436, National Bureau of Economic Research, Cambridge, MA.

Courneya, K. S. and A. V. Carron (1992): The Home Advantage in Sport Competitions: A Literature Review, Journal of Sport and Exercise Psychology, 14, 13-27.

Cox, D. R. (1972): Regression Models and Life-Tables, Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220.

Cox, D. R. and D. Oakes (1984): Analysis of Survival Data, London, UK: Chapman and Hall/CRC Press.

Croson, R. and U. Gneezy (2009): Gender Differences in Preferences, Journal of Economic Literature, 47(2), 448-474.

Cutler, S. J. and F. Ederer (1958): Maximum Utilization of the Life Table Method in Analyzing Survival, Journal of Chronic Diseases, 8(6), 699-712.

D'Agostino, R. B., A. Belanger and R. B. D'Agostino Jr. (1990): A Suggestion for Using Powerful and Informative Tests of Normality, The American Statistician, 44(4), 316-321.

Daghofer, F. (2007): Financial Risk-Taking on “Who Wants to Be a Millionaire”: A Comparison between Austria, Germany and Slovenia, International Journal of Psychology, 42(5), 317-330.

Dawson, P. and S. Dobson (2010): The Influence of Social Pressure and Nationality on Individual Decisions: Evidence from the Behaviour of Referees, Journal of Economic Psychology, 31(2), 181-191.

Dawson, P., S. Dobson, J. Goddard and J. Wilson (2007): Are Football Referees Really Biased and Inconsistent? Evidence on the Incidence of Disciplinary Sanction in the English Premier League, Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(1), 231-250.

De Roos, N. and Y. Sarafidis (2010): Decision Making under Risk in Deal or No Deal, Journal of Applied Econometrics, 25(6), 987-1027.

Deaner, R. O. (2006a): More Males Run Relatively Fast in U.S. Road Races: Further Evidence of a Sex Difference in Competitiveness, Evolutionary Psychology, 4, 303-314.

Deaner, R. O. (2006b): More Males Run Fast: A Stable Sex Difference in Competitiveness in U.S. Distance Runners, Evolution and Human Behavior, 27(1), 63-84.

Deaner, R. O. (2012): Distance Running as an Ideal Domain for Showing a Sex Difference in Competitiveness, Archives of Sexual Behavior, doi:10.1007/s10508-012-9965-z.

Deck, C., J. Lee, J. Reyes and C. Rosen (2008): Measuring Risk Attitudes Controlling for Personality Traits, University of Arkansas, Working Paper Series.

Delfgaauw, J., R. Dur, J. Sol and W. Verbeke (2009): Tournament Incentives in the Field: Gender Differences in the Workplace, Discussion Paper 4395, Institute for the Study of Labor, Bonn.

Demmert, H. H. (1973): The Economics of Professional Team Sports, Lexington, KY: D. C. Heath.

Depken, C. (1999): Free-Agency and the Competitiveness of Major League Baseball, Review of Industrial Organization, 14, 205-217.

Deutsche Fußball-Liga (2013): Report 2013: Die wirtschaftliche Situation im Lizenzfußball, Frankfurt/Main.

Deutscher, C., J. Prinz and D. Weimar (2012): Einkommensdeterminanten von Spitzensportlern - Eine Superstar-Ökonomische Untersuchung unter direkter Berücksichtigung von Netzwerkeffekten, in: Arbeitskreis Sportökonomie e.V. (Ed.): Ökonomie der Sportspiele - Symposiumsband der Jahrestagung 2011, Hofmann, 113-132.

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

Dietl, H. M., M. Lang and S. Werner (2010): Corruption in Professional Sumo: An Update of the Study of Duggan and Levitt, Journal of Sports Economics, 11(4), 383-396.

Dilger, A., B. Frick and F. Tolsdorf (2007): Are Athletes Doped? Some Theoretical Arguments and Empirical Evidence, Contemporary Economic Policy, 25(4), 604-615.

Distaso, W., L. Leonida, D. Maimone Ansaldo Patti and P. Navarra (2012): Corruption and Referee Bias in Football: The Case of Calciopoli, Working Paper, Available at SSRN 2004385.

Dobson, S. and J. Goddard (2001): The Economics of Football, Cambridge, UK: Cambridge University Press.

Dobson, S., J. Goddard and C. Ramlogan (2001): Revenue Convergence in the English Soccer League, Journal of Sports Economics, 2(3), 257-274.

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

Dreber, A., E. von Essen and E. Ranehill (2011): Outrunning the Gender Gap – Boys and Girls Compete Equally, SSE/EFI Working Paper Series in Economics and Finance No. 709, Department of Economics, Stockholm University.

Duggan, M. and S. D. Levitt (2002): Winning isn't Everything: Corruption in Sumo Wrestling, American Economic Review, 92, 1594-1605.

Eckard, E. W. (2001a): Baseball's Blue Ribbon Economic Report: Solutions in Search of a Problem, Journal of Sports Economics, 2, 213-227.

Eckard, E. W. (2001b): Free Agency, Competitive Balance, and Diminishing Returns to Pennant Contention, Economic Inquiry, 39, 430-443.

Eckel, C. C. and P. J. Grossman (2002): Sex Differences and Statistical Stereotyping in Attitudes toward Financial Risk, Evolution and Human Behavior, 23(4), 281-295.

Eckel, C. C. and P. J. Grossman (2008): Men, Women and Risk Aversion: Experimental Evidence, in: Plott, C. and V. Smith (Eds.): Handbook of Experimental Economics Results, New York: Elsevier, 1, 1061-1073.

Efron, B. (1979): Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7(1), 1-26.

Ehrenberg, R. G. and M. L. Bognanno (1990a): Do Tournaments Have Incentive Effects? Journal of Political Economy, 98(6), 1307-1324.

Ehrenberg, R. G. and M. L. Bognanno (1990b): The Incentive Effects of Tournaments Revisited: Evidence from the European PGA Tour, Industrial and Labor Relations Review, 43(3), 74-88.

Eichenberger, E., B. Knechtle, R. Lepers, T. Rosemann and C. A. Rüst (2012): Participation and Running Times in Women and Men Master Mountain Ultra-Marathoners, Open Access Journal of Sports Medicine, 3, 73-80.

El-Hodiri, M. and J. Quirk (1971): An Economic Model of a Professional Sports League, Journal of Political Economy, 70, 1302-1319.

El-Hodiri, M. and J. Quirk (1974): The Economic Theory of a Professional Sports League, in: R. G. Noll (Ed.): Government and the Sports Business, Washington, DC: Brookings Institution.

Falk, A. and M. Kosfeld (2006): The Hidden Costs of Control, American Economic Review, 96(5), 1611-1630.

Fama, E. F. (1970): Efficient Capital Markets: A Review of Theory and Empirical Work, The Journal of Finance, 25(2), 383-417.

Fama, E. F. (1980): Agency Problems and the Theory of the Firm, Journal of Political Economy, 88(2), 288-307.

Feddersen, A. (2006): Economic Consequences of the UEFA Champions League for National Championships – The Case of Germany, Hamburg Working Paper Series in Economic Policy, 01/2006, University of Hamburg, Faculty of Economics.

Feddersen, A. and W. Maennig (2005): Trends in Competitive Balance: Is there Evidence for Growing Imbalance in Professional Sport Leagues? Hamburg Contemporary Economic Discussions, 01/2005, Hamburg: University of Hamburg, Faculty Economics and Social Science.

Feess, E. and G. Muehlheusser (2002): Economic Consequences of Transfer Fee Regulations in European Football, European Journal of Law and Economics, 13(3), 221-237.

Feess, E. and G. Muehlheusser (2003a): The Impact of Transfer Fees on Professional Sports: An Analysis of the New Transfer System for European Football, Scandinavian Journal of Economics, 105(1), 139-154.

Feess, E. and G. Muehlheusser (2003b): Transfer Fee Regulations in European Football, European Economic Review, 47(4), 645-668.

Feess, E., B. Frick and G. Muehlheusser (2004): Legal Restrictions on Buyout Fees: Theory and Evidence from German Soccer, Institute for the Study of Labor, IZA Discussion Paper Series 1180.

Feidakis, A. and A. Tsaoussi (2009): Competitiveness, Gender and Ethics in Legal Negotiations: Some Empirical Evidence, International Negotiation: A Journal of Theory and Practice, 14(3), 537-570.

Fernandez-Cantelli, E. and G. Meeden (2003): An Improved Award System for Soccer, Chance Magazine, 16, 23-29.

Findlay, L. C. and D. M. Ste-Marie (2004): A Reputation Bias in Figure Skating Judging, Journal of Sport and Exercise Psychology, 26(1), 154-166.

Flinn, C. and J. Heckman (1982a): Models for the Analysis of Labor Force Dynamics, in: Basmann, R. and G. Rhodes (Eds.): Advances in Econometrics, Vol. 1, Greenwich, CT: JAI Press.

Flinn, C. and J. Heckman (1982b): New Methods for Analyzing Structural Models of Labor Force Dynamics, Journal of Econometrics, 18, 115-168.

Forrest, D., J. Goddard and R. Simmons (2005): Odds-Setters as Forecasters: The Case of English Football, International Journal of Forecasting, 21(3), 551-564.

Fort, R. (2000): European and North American Sports Differences (?), Scottish Journal of Political Economy, 47(4), 431-455.

Fort, R. (2001): Revenue Disparity and Competitive Balance in Major League Baseball, in: Baseball's Revenue Gap: Pennant for Sale? Hearing before the Subcommittee on Antitrust, Business Rights, and Competition of the Committee on the Judiciary (pp. 42-52), U.S. Senate, 106th Congress, 2nd Session, November 21, 2000.

Fort, R. and J. Maxcy (2003): Competitive Balance in Sports Leagues: An Introduction, Journal of Sports Economics, 4, 154-160.

Fort, R. and J. Quirk (1995): Cross-Subsidization, Incentives, and Outcomes in Professional Team Sports Leagues, Journal of Economic Literature, 33, 1265-1299.

Fort, R. and J. Quirk (2011): Optimal Competitive Balance in a Season Ticket League, Economic Inquiry, 49, 464-473.

Franck, E. and S. Nüesch (2012): Talent and/or Popularity: What Does it Take to be a Superstar? Economic Inquiry, 50(1), 202-216.

Frey, B. S. and L. Goette (1999): Does Pay Motivate Volunteers? Working Paper, Institute for Empirical Research in Economics, University of Zurich.

Frick, B. (1998): Lohn und Leistung im professionellen Sport: Das Beispiel Stadt-Marathon, Konjunkturpolitik, 44, 114-140.

Frick, B. (2004): Warum laufen die denn so schnell? Die Anreizwirkungen von Prämien bei professionellen Marathonläufern, in: Jütting, D. H. (Ed.): Die Laufbewegung in Deutschland – interdisziplinär betrachtet, Münster: Waxmann, 11, 33-48.

Frick, B. (2007): The Football Players' Labor Market: Empirical Evidence from the Major European Leagues, Scottish Journal of Political Economy, 54(3), 422-446.

Frick, B. (2009): Globalization and Factor Mobility: The Impact of the "Bosman-Ruling" on Player Migration in Professional Soccer, Journal of Sports Economics, 10(1), 88-106.

Frick, B. (2011): Performance, Salaries, and Contract Length: Empirical Evidence from German Soccer, International Journal of Sport Finance, 6(2), 87-12.

Frick, B. (2011a): Gender Differences in Competitiveness: Empirical Evidence from Professional Distance Running, Labour Economics, 18(3), 389-398.

Frick, B. (2011b): Gender Differences in Competitive Orientations: Empirical Evidence from Ultramarathon Running, Journal of Sports Economics, 12(3), 317-340.

Frick, B. (2012): Career Duration in Professional Football: The Case of German Soccer Referees, in: Kahane, L. H. and S. Shmanske (Eds.): The Oxford Handbook of Sports Economics: The Economics of Sports (Vol. 1), Oxford University Press, 487-500.

Frick, B. and O. Fabel (2013): Special Issue "Insider Econometrics", Journal of Business Economics, 83(2), 99-100.

Frick, B. and J. Prinz (2006): Crisis? What Crisis? Football in Germany, Journal of Sports Economics, 7(1), 60-75.

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

Frick, B. and F. Scheel (2013): Gender Differences in Competitiveness – Empirical Evidence from 100m Races, in: Marikova Leeds, E. and M. A. Leeds (Eds.): Handbook on the Economics of Women in Sports, Northampton: Edward Elgar, 293-318.

Frick, B., O. Gürtler and J. Prinz (2009a): Men in Black: Monitoring and Performance of German Soccer Referees, in: Dietl, H., E. Franck and H. Kempf (Eds.): Football – Economics of a Passion, Schorndorf: Hofmann, 309-321.

Frick, B., G. Pietzner and J. Prinz (2007): Career Duration in a Competitive Environment: The Labor Market for Soccer Players in Germany, Eastern Economic Journal, 33(3), 429-442.

Frick, B., G. Pietzner and J. Prinz (2009): Team Performance and Individual Career Duration: Evidence from the German "Bundesliga", in: Andersson, P., P. Ayton and C. Schmidt (Eds.): Myths and Facts about Football: The Economics and Psychology of the World's Greatest Sport, Cambridge: Cambridge Scholars Press, 327-348.

Frick, B., O. Gürtler, J. Prinz and A. Wiendl (2009b): Einkommens- oder Reputationsmaximierung? Eine empirische Untersuchung der Vergütung und Leistung von Bundesliga-Schiedsrichtern, Die Betriebswirtschaft, 69, 69-83.

Fujita, F., E. Diener and E. Sandvik (1991): Gender Differences in Negative Affect and Well-Being: The Case for Emotional Intensity, Journal of Personality and Social Psychology, 61(3), 427-434.

Fullenkamp, C., R. Tenorio and R. Battalio (2003): Assessing Individual Risk Attitudes Using Field Data from Lottery Games, Review of Economics and Statistics, 85(1), 218-226.

Garcia, J. and P. Rodriguez (2002): The Determinants of Football Match Attendance Revisited: Empirical Evidence from the Spanish Football League, Journal of Sports Economics 3(1), 18-38.

Garcia, D., F. Sangiorgi and B. Urosevic (2007): Overconfidence and Market Efficiency with Heterogeneous Agents, Economic Theory, 30(2), 313-336.

Garcia-Gallego, A., N. Georgantzis and A. Jaramillo-Gutierrez (2012): Gender Differences in Ultimatum Games: Despite Rather than Due to Risk Attitudes, Journal of Economic Behavior and Organization, 83(1), 42-49.

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

Garratt, R. J., C. Weinberger and N. Johnson (2011): The State Street Mile: Age and Gender Differences in Competition Aversion in the Field, Economic Inquiry, doi: 10.1111/j.1465-7295.2011.00370.x.

Gerardi, K. S. and A. H. Shapiro (2009): Does Competition Reduce Price Dispersion? New Evidence from the Airline Industry, Journal of Political Economy, 117(1), 1-37.

Gerdes, C. and P. Gränsmark (2010): Strategic Behavior across Gender: A Comparison of Female and Male Expert Chess Players, Labour Economics, 17(5), 766-775.

Gertner, R. (1993): Game Shows and Economic Behavior: Risk-Taking on "Card Sharks", The Quarterly Journal of Economics, 108(2), 507-521.

Gibbons, R. (1987): Piece-Rate Incentive Schemes, Journal of Labor Economics, 5(4), 413-429.

Gibbs, B. G., J. A. Jarvis and M. J. Dufur (2012): The Rise of the Underdog? The Relative Age Effect Reversal among Canadian-Born NHL Hockey Players: A Reply to Nolan and Howell, International Review for the Sociology of Sport, 47(5), 644-649.

Gilbert, R. A. (1984): Bank Market Structure and Competition: A Survey, Journal of Money, Credit and Banking, 16, 617-645.

Gneezy, U. and A. Rustichini (2000): A Fine is a Price, Journal of Legal Studies, 29(1), 1-17.

Gneezy, U. and A. Rustichini (2004): Gender and Competition at a Young Age, American Economic Review, 94(2), 377-381.

Gneezy, U., K. L. Leonard and J. A. List (2009): Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society, Econometrica, 77(5), 1637-1664.

Gneezy, U., S. Meier and P. Rey-Biel (2011): When and Why Incentives (Don't) Work to Modify Behavior, The Journal of Economic Perspectives, 25(4), 191-209.

Goldin, C. and C. Rouse (2000): Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians, American Economic Review, 90(4), 715-741.

Gong, B. and C. L. Yang (2012): Gender Differences in Risk Attitudes: Field Experiments on the Matrilineal Mosuo and the Patriarchal Yi, Journal of Economic Behavior and Organization, 83(1), 59-65.

Goumas, C. (2012): Home Advantage and Referee Bias in European Football, European Journal of Sport Science, 14(1), 243-249.

Gränsmark, P. (2012): Masters of our Time: Impatience and Self-Control in High-Level Chess Games, Journal of Economic Behavior and Organization, 82(1), 179-191.

Greene, W. H. (2000): Econometric analysis, 4th ed., Englewood Cliffs, NJ: Prentice Hall.

Groot, L. F. M. (2008): Economics, Uncertainty and European Football: Trends in Competitive Balance, Cheltenham, UK: Edward Elgar.

Groothuis, P. A. and J. R. Hill (2004): Exit Discrimination in the NBA: A Duration Analysis of Career Length, Economic Inquiry, 42(2), 341-349.

Groothuis, P. A. and J. R. Hill (2008): Exit Discrimination in Major League Baseball: 1990-2004, Southern Economic Journal, 75(2), 574-590.

Groothuis, P. A. and J. R. Hill (2013): Pay Discrimination, Exit Discrimination or Both? Another Look at an Old Issue Using NBA Data, Journal of Sports Economics, 14(2), 171-185.

Gujarati, D. N. and D. C. Porter (2009): Basic Econometrics, 5th ed., Boston: McGraw-Hill Irwin.

Gupta, N. D., A. Poulsen and M. C. Villeval (2011): Gender Matching and Competitiveness: Experimental Evidence, Economic Inquiry, doi:10.1111/j.1465-7295.2011.00378.x.

Gysler, M., J. B. Kruse and R. Schubert (2002): Ambiguity and Gender Differences in Financial Decision Making: An Experimental Examination of Competence and Confidence Effects, Working Paper, Center for Economic Research, Swiss Federal Institute of Technology.

Hadley, L., J. Ciecka and A. C. Krautmann (2005): Competitive Balance in the Aftermath of 1994 Players' Strike, Journal of Sports Economics, 6, 379-389.

Halko, M. L., M. Kaustia and E. Alanko (2012): The Gender Effect in Risky Asset Holdings, Journal of Economic Behavior and Organization, 83(1), 66-81.

Hall, S., S. Szymanski and A. S. Zimbalist (2002): Testing Causality between Team Performance and Payroll: The Cases of Major League Baseball and English Soccer, Journal of Sports Economics, 3, 149-168.

Harbaugh, W. T., K. Krause and L. Vesterlund (2002): Risk Attitudes of Children and Adults: Choices over Small and Large Probability Gains and Losses, Experimental Economics, 5(1), 53-84.

Harrison, G. W. and J. A. List (2004): Field Experiments, Journal of Economic Literature, 1009-1055.

Hartley, R., G. Lanot and I. Walker (2013): Who Really Wants to Be a Millionaire? Estimates of Risk Aversion from Gameshow Data, Journal of Applied Econometrics, DOI: 10.1002/jae.2353.

Haugen, K. K. (2008): Point Score Systems and Competitive Imbalance in Professional Soccer, Journal of Sports Economics, 9, 191-210.

Healy, A. and J. Pate (2011): Can Teams Help to Close the Gender Competition Gap? The Economic Journal, 121(555), 1192-1204.

Heckman, J. J. (1979): Sample Selection Bias as a Specification Error, Econometrica: Journal of the Econometric Society, 153-161.

Heß, M., D. von Scheve, J. Schupp and G. G. Wagner (2013): Members of German Federal Parliament More Risk-Loving than General Population, SOEP paper No. 546, Available at SSRN: http://ssrn.com/abstract=2253836 or http://dx.doi.org/10.2139/ssrn.2253836.

Hilary, G. and L. Menzley (2006): Does Past Success Lead Analysts to Become Overconfident? Management Science, 52(4), 489-500.

Hill, D. (2009): To Fix or not to Fix? How Corruptors Decide to Fix Football Matches, Global Crime, 10(3), 157-177.

Hoang, H. and D. Rascher (1999): The NBA, Exit Discrimination, and Career Earnings, Industrial Relations, 38(1), 69-91.

Hoehn, T. and S. Szymanski (1999): The Americanization of European Football, Economic Policy, 28, 205-240.

Hoelzl, E. and A. Rustichini (2005): Overconfident: Do You Put Your Money on it? The Economic Journal, 115 (April), 305-318.

Hoffman, M., U. Gneezy and J. A. List (2011): Nurture Affects Gender Differences in Spatial Abilities, Proceedings of the National Academy of Sciences of the United States of America, 108(36), 14786-14788.

Hogarth, R. M., N. Karelaia and C. A. Trujillo (2012): When Should I Quit? Gender Differences in Exiting Competitions, Journal of Economic Behavior and Organization, 83, 136-150.

Hölmstrom, B. (1979): Moral Hazard and Observability, The Bell Journal of Economics, 10(1), 74-91.

Hölmstrom, B. (1982): Moral Hazard in Teams, The Bell Journal of Economics, 13(2), 324-340.

Hölmstrom, B. and P. Milgrom (1991): Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design, Journal of Law, Economics and Organization, 7, 24-52.

Hölmstrom, B. and P. Milgrom (1994): The Firm as an Incentive System, American Economic Review, 84(4), 972-991.

Holt, C. A. and S. K. Laury (2002): Risk Aversion and Incentive Effects, American Economic Review, 92(3), 1644-1655.

Horwitz, S. K. and I. B. Horwitz (2007): The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography, Journal of Management, 33(6), 987-1015.

Humphreys, B. R. (2002): Alternative Measures of Competitive Balance in Sports Leagues, Journal of Sports Economics, 3, 133-148.

Ichniowski, C. and K. L. Shaw (2009): Insider Econometrics: Empirical Studies of How Management Matters, Working Paper No. 15618, National Bureau of Economic Research, Cambridge.

Ichniowski, C., K. Shaw and G. Prennushi (1997): The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines, American Economic Review, 87(3), 291-313.

Ivanova-Stenzel, R. and D. Kübler (2011): Gender Differences in Team Work and Team Competition, Journal of Economic Psychology, 32(5), 797-808.

Jennett, N. I. (1984): Attendances, Uncertainty of Outcome and Policy in Scottish League Football, Scottish Journal of Political Economy, 31(2), 176-198.

Jensen, M. C. and W. H. Meckling (1976): Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure, Journal of Financial Economics, 3(4), 305-360.

Johnson, D. K. and T. R. Gleason (2009): Who Really Wants to Be a Millionaire? Gender Differences in Game Show Contestant Behavior under Risk, Social Science Quarterly, 90(2), 243-261.

Johnson, J. E. V. and P. L. Powell (1994): Decision Making, Risk and Gender: Are Managers Different? British Journal of Management, 5, 123-138.

Johnston, R. (2008): On Referee Bias, Crowd Size and Home Advantage in the English Soccer Premiership, Journal of Sports Sciences, 26(6), 563-568.

Jones, M. V., G. C. Paull and J. Erskine (2002): The Impact of a Team's Aggressive Reputation on the Decisions of Association Football Referees, Journal of Sports Sciences, 20(12), 991-1000.

Jurajda, S. and D. Münich (2008): Gender Gap in Admission Performance under Competitive Pressure, Working Paper 371, Center for Economic Research and Graduate Education, Charles University Prague.

Jurajda, S. and D. Münich (2011): Gender Gap in Performance under Competitive Pressure, American Economic Review, 101(3), 514-518.

Kahn, L. M. (2000): The Sports Business as a Labor Market Laboratory, The Journal of Economic Perspectives, 14(3), 75-94.

Kahnemann, D. and A. Tversky (1979): An Analysis of Decision under Risk, Econometrica, 47(2), 263-292.

Kalter, F. (1999): Ethnische Kundenpräferenzen im Professionellen Sport: Der Fall der Fußball-Bundesliga, Zeitschrift für Soziologie, 28 (3), 219-234.

Kamas, L. and A. Preston (2009): Social Preferences, Competitiveness and Compensation: Are there Gender Differences? Working Paper, Santa Clara University.

Kaplan, E. L. and P. Meier (1958): Nonparametric Estimation from Incomplete Observations, Journal of the American Statistical Association, 53, 457-481.

Kashy, D. A. and D. A. Kenny (2000): The Analysis of Data from Dyads and Groups, Handbook of Research Methods in Social and Personality Psychology, 451-477.

Kesenne, S. (1996): League Management in Professional Team Sports with Win Maximizing Clubs, European Journal for Sport Management, 2(2), 14-22.

Kesenne, S. (1999): Player Market Regulation and Competitive Balance in a Win Maximizing Scenario, in: C. Jeanrenaud and S. Kesenne (Eds.): Competition Policy in Professional Sports: Europe after the Bosman Case, Antwerp, Belgium: Standaard Editions Ltd.

Kesenne, S. (2000): Revenue Sharing and Competitive Balance in Professional Team Sports, Journal of Sports Economics, 1, 56-65.

Kiefer, N. M. (1988): Economic Duration Data and Hazard Functions, Journal of Economic Literature, 26(2), 646-679.

Kiefer, N. M. and G. R. Neumann (1979): An Empirical Job Search Model with a Test of the Constant Reservation Wage Hypothesis, Journal of Political Economy, 87(1), 89-107.

Knechtle, B., R. Lepers, C. A. Rüst and P. J. Stapley (2012): Relative Improvements in Endurance Performance with Age: Evidence from 25 Years of Hawaii Ironman Racing, Age (Dordr), doi: 10.1007/s11357-012-9392-z.

Kolle, A. (2014): Gender and Ethnic Discrimination in Hiring: Evidence from Field Experiments in the German Labor Market, Dissertation, University of Paderborn.

Koning, R. H. (2005): Home Advantage in Speed Skating: Evidence from Individual Data, Journal of Sports Sciences, 23(4), 417-427.

Koszegi, B. (2006): Ego Utility, Overconfidence and Task Choice, Journal of the European Economic Association, 4(4), 673-707.

Kräkel, M. (2007): Doping and Cheating in Contest-like Situations, European Journal of Political Economy, 23(4), 988-1006.

Krautmann, A., Y. H. Lee and K. Quinn (2010): Playoff Uncertainty and Pennant Races, Journal of Sports Economics, Online First, DOI: 10.1177/1527002510388944.

Krueger, A. B. (2005): The Economics of Real Superstars: The Market for Rock Concerts in the Material World, Journal of Labor Economics, 23(1), 1-30.

Krueger, A. B. and J. Maleckova (2003): Education, Poverty and Terrorism: Is There a Causal Connection? The Journal of Economic Perspectives, 17(4), 119-144.

Lallemand, T., R. Plasman and F. Rycx (2008): Women and Competition in Elimination Tournaments: Evidence from Professional Tennis Data, Journal of Sports Economics, 9(1), 3-19.

Lancaster, T. (1979): Econometric Methods for Duration of Unemployment, Econometrica, 47, 939-956.

Larkin, J. E. and H. A. Pines (2003): Gender and Risk in Public Performance, Sex Roles, 49(5-6), 197-210.

Larwood, L. and W. Whittaker (1977): Managerial Myopia: Self-Serving Biases in Organizational Planning, Journal of Applied Psychology 62, 94-198.

Lavy, V. (2008): Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments among Teachers, Working Paper 14338, National Bureau of Economic Research, Cambridge, MA.

Lavy, V. (2012): Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments among Teachers, Economic Journal, doi: 10.1111/j.1468-0297.2012.02542.x.

Lazear, E. P. (1987): Incentive Contracts, in Eatwell, J., M. Milgate and P. Newman (Eds.): The New Palgrave: A Dictionary of Economics, Vol. 2, London: The Macmillan Press Limited, 744-748.

Lazear, E. P. and S. Rosen (1981): Rank-Order Tournaments as Optimum Labor Contracts, Journal of Political Economy, 841-864.

Lazear, E. P. (2000): Performance Pay and Productivity, American Economic Review, 90(5), 1346-1361.

Leeds, M. and P. von Allmen (2008): The Economics of Sports, 3, Boston: Pearson.

Leuven, E., H. Oosterbeek, J. Sonnemans and B. van der Klaauw (2011): Incentives versus Sorting in Tournaments: Evidence from a Field Experiment, Journal of Labor Economics, 29(3), 637-658.

Levitt, S. D. (2003): Testing Theories of Discrimination: Evidence from "Weakest Link", National Bureau of Economic Research.

Liao, T. F. (1994): Interpreting Probability Models: Logit, Probit and Other Generalized Linear Models, Newbury Park, CA: Sage.

Long, J. S. and J. Freese (2006): Regression Models for Categorical Dependent Variables Using Stata, 2nd ed., College Station, TX: Stata Press.

Lucey, B. and D. Power (2004): Do Soccer Referees Display Home Team Favouritism? Mimeo, Trinity College Dublin, Dublin.

Lynch, J. G. and J. S. Zax (2000): The Rewards to Running: Prize Structure and Performance in Professional Road Racing, Journal of Sports Economics, 1(4), 323-340.

MacDonald, J. M. (1987): Competition and Rail Rates for the Shipment of Corn, Soybeans, and Wheat, Rand Journal of Economics, 18, 151-163.

Malmendier, U. and G. Tate (2005): CEO Overconfidence and Corporate Investment, Journal of Finance, 60(6), 2661-2700.

Malmendier, U. and G. Tate (2008): Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction, Journal of Financial Economics, 89, 20-43.

Maloney, M. T. and R. E. McCormick (2000): The Response of Workers to Wages in Tournaments: Evidence from Foot Races, Journal of Sports Economics, 1(2), 99-123.

Malul, M., D. Shapira and A. Shoham (2013): Practical Modified Gini Index, Applied Economics Letters, 20(4), 324-327.

Manning, A. and F. Saidi (2010): Understanding the Gender Pay Gap: What's Competition Got to Do with It? Industrial and Labor Relations Review, 63(4), 681-698.

Marburger, D. R. (2002): Property Rights and Unilateral Player Transfers in a Multiconference Sports League, Journal of Sports Economics, 3, 122-132.

Markowski, C. A. and E. P. Markowski (1990): Conditions for the Effectiveness of a Preliminary Test of Variance, The American Statistician, 44(4), 322-326.

Mas, A. and E. Moretti (2009): Peers at Work, American Economic Review, 99(1), 112-145.

Matheson, V. A. and J. Congdon-Hohman (2013): International Women's Soccer and Gender Inequality: Revisited, in: Marikova Leeds, E. and M. A. Leeds (Eds.): Handbook on the Economics of Women in Sports, Northampton: Edward Elgar, 345-364.

Matheson, V. A., J. Treber and R. Levy (2013): Gender Differences in Competitive Balance in Intercollegiate Basketball, in: Marikova Leeds, E. and M. A. Leeds (Eds.): Handbook on the Economics of Women in Sports, Northampton: Edward Elgar, 251-268.

Matthews, P. H., P. Sommers and F. Peschiera (2007): Incentives and Superstars on the LPGA Tour, Applied Economics, 39(1), 87-94.

Maxcy, J. G. (2002): Rethinking Restrictions on Player Mobility in Major League Baseball, Contemporary Economic Policy, 20, 145-159.

Mellström, C. and M. Johannesson (2008): Crowding Out in Blood Donation: Was Titmuss Right? Journal of the European Economic Association, 6(4), 845-863.

Menkhoff, L., U. Schmidt and T. Brozynski (2006): The Impact of Experience on Risk Taking, Overconfidence and Herding of Fund Managers: Complementary Survey Evidence, European Economic Review, 50(7), 1753-1766.

Metrick, A. (1995): A Natural Experiment in "Jeopardy!", American Economic Review, 85(1), 240-253.

Michie, J. and C. Oughton (2004): Competitive Balance in Football: Trends and Effects, London: The Sports Nexus.

Michie, J. and C. Oughton (2005): Competitive Balance in Football: An Update, London: The Sports Nexus.

Milgrom, P. and J. Roberts (1988): An Economic Approach to Influence Activities in Organizations, American Journal of Sociology, 94(1), 154-179.

Mills, B. and R. Fort (2011): League Level Attendance and Outcome Uncertainty in the NBA, NFL and NHL, Working Paper, University of Michigan.

Mills, B. and R. Fort (2014): League-Level Attendance and Outcome Uncertainty in US Pro Sports Leagues, Economic Inquiry, 52(1), 205-218.

Mincer, J. (1958): Investment in Human Capital and Personal Income Distribution, Journal of Political Economy, 66(4), 281-302.

Mincer, J. (1970): The Distribution of Labor Incomes: A Survey with Special Reference to the Human Capital Approach, Journal of Economic Literature, 8(1), 1-26.

Mincer, J. (1974): Schooling, Experience, and Earnings, New York: Columbia University Press.

Mulino, D., R. Scheelings, R. Brooks and R. Faff (2009): Does Risk Aversion Vary with Decision-Frame? An Empirical Test Using Recent Game Show Data, Review of Behavioral Finance, 1(1-2), 44-61.

Myers, T. and N. Balmer (2012): The Impact of Crowd Noise on Officiating in Muay Thai: Achieving External Validity in an Experimental Setting, Frontiers in Psychology, 3, 1-7.

Myers, T., A. Nevill and Y. Al-Nakeeb (2012): The Influence of Crowd Noise upon Judging Decisions in Muay Thai, Advances in Physical Education, 2, 148-152.

Nalebuff, B. J. and J. E. Stiglitz (1983): Prizes and Incentives: Towards a General Theory of Compensation and Competition, The Bell Journal of Economics, 14(1), 21-43.

Neale, W. C. (1964): The Peculiar Economics of Professional Sports, Quarterly Journal of Economics, 78, 1-14.

Neelakantan, U. (2010): Estimation and Impact of Gender Differences in Risk Tolerance, Economic Inquiry, 48(1), 228-233.

Nekby, L., P. Skogman Thoursie and L. Vahtrik (2008): Gender and Self-Selection into a Competitive Environment: Are Women More Overconfident than Men? Economics Letters 100, 405-407.

Nekby, L., P. S. Thoursie and L. Vahtrik (2008): Gender and Self-Selection into a Competitive Environment: Are Women More Overconfident than Men? Economics Letters, 100(3), 405-407.

Nevill, A. M. and R. L. Holder (1999): Home Advantage in Sport, Sports Medicine, 28(4), 221-236.

Nevill, A. M., N. J. Balmer and A. Mark Williams (2002): The Influence of Crowd Noise and Experience upon Refereeing Decisions in Football, Psychology of Sport and Exercise, 3(4), 261-272.

Nevill, A. M., S. M. Newell and S. Gale (1996): Factors Associated with Home Advantage in English and Scottish Soccer Matches, Journal of Sports Sciences, 14(2), 181-186.

Nevill, A., T. Webb and A. Watts (2013): Improved Training of Football Referees and the Decline in Home Advantage Post WW2, Psychology of Sport and Exercise, 14(2), 220-227.

Nevill, A. M., R. L. Holder, A. Bardsley, H. Calvert and S. Jones (1997): Identifying Home Advantage in International Tennis and Golf Tournaments, Journal of Sports Sciences, 15(4), 437-443.

Nickell, S. (1979): Estimating the Probability of Leaving Unemployment, Econometrica, 47, 1249-1266.

Niederle, M. and L. Vesterlund (2007): Do Women Shy Away from Competition? Do Men Compete too Much? Quarterly Journal of Economics, 122(3), 1067-1101.

Niederle, M. and L. Vesterlund (2011): Gender and Competition, Annual Review of Economics, 3(1), 601-630.

Niederle, M., C. Segal and L. Vesterlund (2013): How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness, Management Science, 59(1), 1-16.

Noll, R. G. (2002): The Economics of Promotion and Relegation in Sports Leagues: The Case of English Football, Journal of Sports Economics, 3, 168-203.

Nöth, M. and M. Weber (2003): Information Aggregation with Random Ordering: Cascades and Overconfidence, Economic Journal, 113(484), 166-189.

Ohkusa, Y. (1999): Additional Evidence for the Career Concern Hypothesis with Uncertainty of the Quit Period: The Case of Professional Baseball Players in Japan, Applied Economics, 31, 1481-1487.

Ohkusa, Y. (2001): An Empirical Examination of the Quit Behavior of Professional Baseball Players in Japan, Journal of Sports Economics, 2(1), 80-88.

Ors, E., F. Palomino and E. Peyrache (2008): Performance Gender-Gap: Does Competition Matter? Discussion Paper No. 6891, Centre for Economic Policy Research, London.

Ors, E., F. Palomino and E. Peyrache (2013): Performance Gender Gap: Does Competition Matter? Journal of Labor Economics, 31(3), 443-499.

Osterloh, M. and B. S. Frey (2002): Does Pay for Performance Really Motivate Employees? In: Neely, A. D. (Ed.): Business Performance Measurement: Theory and Practice, Cambridge, UK: Cambridge University Press, 107-122.

Parsons, C. A., J. Sulaeman, M. C. Yates and D. S. Hamermesh (2011): Strike Three: Discrimination, Incentives and Evaluation, American Economic Review, 101(4), 1410-1435.

Paserman, M. D. (2007): Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players, Discussion Paper 2834, Institute for the Study of Labor, Bonn.

Paserman, M. D. (2010): Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players, Working Paper, Boston University and Hebrew University.

Pawlowski, T., C. Breuer and A. Hovemann (2010): Top Clubs' Performance and the Competitive Situation in European Domestic Football Competitions, Journal of Sports Economics, 11(2), 186-202.

Perline, M. M. and G. C. Stoldt (2007a): Competitive Balance and the Big 12, The SMART Journal, 4(1), 47-58.

Perline, M. M. and G. C. Stoldt (2007b): Competitive Balance in Men's and Women's Basketball: The Cast of the Missouri Valley Conference, The Sport Journal, 10(4), October.

Pettersson-Lidbom, P. and M. Priks (2010): Behavior under Social Pressure: Empty Italian Stadiums and Referee Bias, Economics Letters, 108(2), 212-214.

Plessner, H. and T. Betsch (2001): Sequential Effects in Important Referee Decisions: The Case of Penalties in Soccer, Journal of Sport and Exercise Psychology, 23(3), 254-259.

Post, T., M. J. van den Assem, G. Baltussen and R. H. Thaler (2008): Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show, American Economic Review, 98(1), 38-71.

Powell, M. and D. Ansic (1997): Gender Differences in Risk Behaviour in Financial Decision-Making: An Experimental Analysis, Journal of Economic Psychology, 18, 605-628.

Prendergast, C. (1999): The Provision of Incentives in Firms, Journal of Economic Literature, 37(1), 7-63.

Price, J. (2008): Gender Differences in the Response to Competition, Industrial and Labor Relations Review, 61(3), 320-333.

Price, C. R. (2010): Do Women Shy Away from Competition? Do Men Compete too Much? A (Failed) Replication, Working Paper, University of Southern Indiana.

Price, C. R. (2012): Gender, Competition, and Managerial Decisions, Management Science, 58(1), 114-122.

Price, J. and J. Wolfers (2010): Racial Discrimination among NBA Referees, The Quarterly Journal of Economics, 125(4), 1859-1887.

Quirk, J. and R. D. Fort (1992): Pay Dirt: The Business of Professional Team Sports, Princeton, NJ: Princeton University Press.

Quirk, J. and R. D. Fort (1997): Competitive Balance in Sports Leagues, in: Quirk, J. and R. D. Fort (Eds.): Pay Dirt: The Business of Professional Team Sports, Princeton, NJ: Princeton University Press, 240-293.

Reilly, B. and R. Witt (2013): Red Cards, Referee Home Bias and Social Pressure: Evidence from English Premiership Soccer, Applied Economics Letters, 20(7), 710-714.

Reuben, E., P. Rey-Biel, P. Sapienza and L. Zingales (2012): The Emergence of Male Leadership in Competitive Environments, Journal of Economic Behavior and Organization, 83(1), 111-117.

Rickman, N. and R. Witt (2008): Favouritism and Financial Incentives: A Natural Experiment, Economica, 75(298), 296-309.

Ridder, G., J. S. Cramer and P. Hopstaken (1994): Down to Ten: Estimating the Effect of a Red Card in Soccer, Journal of the American Statistical Association, 89(427), 1124-1127.

Rocha, B., F. Sanches, I. Souza and J. Carlos Domingos da Silva (2013): Does Monitoring Affect Corruption? Career Concerns and Home Bias in Football Refereeing, Applied Economics Letters, 20(8), 728-731.

Rosen, S. (1981): The Economics of Superstars, American Economic Review, 71(5), 845-858.

Rosen, S. (1986): Prizes and Incentives in Elimination Tournaments, American Economic Review, 76(4), 701-715.

Rosen, S. and A. Sanderson (2001): Labour Markets in Professional Sports, The Economic Journal, 111(469), 47-68.

Rottenberg, S. (1956): The Baseball Players' Labor Market, Journal of Political Economy, 64, 242-258.

Royston, P. (1991a): Tests for Departure from Normality, Stata Technical Bulletin 2: 16–17, Reprinted in Stata Technical Bulletin Reprints, Vol. 1, 101–104, College Station, TX: Stata Press.

Sanderson, A. R. (2002): The Many Dimensions of Competitive Balance, Journal of Sports Economics, 3, 203-228.

Säve-Söderbergh, J. and G. S. Lindquist (2011): "Girls Will Be Girls", Especially among Boys: Risk-Taking in the "Daily Double" on Jeopardy, Economic Letters, 112, 158-160.

Schmidt, M. B. (2001): Competition in Major League Baseball: The Impact of Expansion, Applied Economic Letters, 8, 21-26.

Schmidt, M. B. and D. J. Berri (2001): Competitive Balance and Attendance: The Case of Major League Baseball, Journal of Sports Economics, 2, 145-167.

Schneider, M., and F. Bauhoff (2013): "Sekretaerin des Vorstandes" gesucht: Stellenanzeigen und die expressive Funktion des AGG (Hiring "Female Secretary to the Board of Directors": Job Adverts and the Expressive Function of the General Act on Equal Treatment), Industrielle Beziehungen (The German Journal of Industrial Relations), 20(1), 54-76.

Schubert, R., M. Brown, M. Gysler and H. W. Brachinger (1999): Financial Decision-Making: Are Women Really More Risk-Averse? American Economic Review (Papers and Proceedings), 89(2), 381-385.

Schultz, T. W. (1961): Investment in Human Capital, American Economic Review, 51(1), 1-17.

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.

Scully, G. W. (1989): The Business of Major League Baseball, Chicago: University of Chicago Press.

Sealy, R. and S. Vinnicombe (2013): The Female FTSE Board Report 2012: False Dawn of Progress for Women on Boards? Bedford: Cranfield University Management School Press.

Shearer, B. (2004): Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment, Review of Economic Studies, 71(2), 513-534.

Shurchkov, O. (2008): Performance in Competitive Environments: Are Women Really Different? Discussion Paper, Department of Economics, Wellesley College.

Shurchkov, O. (2012): Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints, Journal of the European Economic Association, 10(5), 1189-1213.

Simmons, R. (1997): Implications of the Bosman Ruling for Football Transfer Markets, Economic Affairs, 17(3), 13-18.

Simon, H. (1955): A Behavioural Model of Rational Choice, Quarterly Journal of Economics, 69(1), 99-118.

Simon, H. A. and C. P. Bonini (1958): The Size Distribution of Business Firms, American Economic Review, 48(4), 607-617.

Sitkin, S. B. and L. R. Weingart (1995): Determinants of Risky Decision-Making Behavior: A Test of the Mediating Role of Risk Perceptions and Propensity, Academy of Management Journal, 38(6), 1573-1592.

Sloane, P. J. (1971): The Economics of Professional Football: The Football Club as a Utility Maximiser, Scottish Journal of Political Economy, 18(2), 121-146.

Solt, F. (2009): Standardizing the World Income Inequality Database, Social Science Quarterly, 90(2), 231-242.

Spurr, S. J. and W. Barber (1994): The Effect of Performance on a Worker's Career: Evidence from Minor League Baseball, Industrial and Labor Relations Review, 47(4), 692-708.

Staw, B. M. and H. Hoang (1995): Sunk Costs in the NBA: Why Draft Order Affects Playing Time and Survival in Professional Basketball, Administrative Science Quarterly, 40, 474-494.

Sullivan, D. (1985): Testing Hypotheses about Firm Behavior in the Cigarette Industry, Journal of Political Economy, 93, 586-598.

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 M. G. Kocher (2004): Favoritism of Agents – The Case of Referees' Home Bias, Journal of Economic Psychology, 25(4), 461-469.

Sutter, M. and D. Rützler (2010): Gender Differences in Competition Emerge Early in Life, Discussion Paper 5015, Institute for the Study of Labor, Bonn.

Svensson, O. (1981): Are We All Less Risky and More Skillful than Our Fellow Drivers? Acta Pschologica, 47(2), 143-148.

Szymanski, S. (2001): Income Inequality, Competitive Balance and the Attractiveness of Team Sports: Some Evidence and a Natural Experiment from English Soccer, The Economic Journal, 111, 69-84.

Szymanski, S. and T. Kuypers (1999): Winners and Losers: The Business Strategy of Football, London: Viking, Penguin.

Szymanski, S. and R. Smith (1997): The English Football Industry: Profit, Performance and Industrial Structure, International Review of Applied Economics, 11(1), 135-153.

Thibault, V., M. Guillaume, G. Berthelot, N. El Helou, […], J. F. Toussaint (2010): Women and Men in Sport Performance: The Gender Gap has not Evolved Since 1983, Journal of Sports Science and Medicine, 9, 214-223.

Tversky, A. and D. Kahnemann (1991): Loss Aversion in Riskless Choice: A Reference Dependent Model, Quarterly Journal of Economics, 106(4), 1039-1061.

Tversky, A. and D. Kahnemann (1992): Advances in Prospect Theory: Cumulative Representation of Uncertainty, Journal of Risk and Uncertainty, 5(4), 297-323.

Van den Steen, E. (2004): Rational Overoptimism (and other Biases), American Economic Review, 94(4), 1141-1151.

Vandegrift, D. and A. Yavas (2009): Men, Women, and Competition: An Experimental Test of Behavior, Journal of Economic Behavior and Organization, 72(1), 554-570.

Von Neumann, J. and O. Morgenstern (1944): Theory of Games and Economic Behavior, Princeton, NJ: Princeton University Press.

Vrooman, J. (1995): A General Theory of Professional Sports Leagues, Southern Economic Journal, 61, 971-990.

Waldman, M. (1990): Up-or-Out Contracts: A Signaling Perspective, Journal of Labor Economics, 8(2), 230-250.

Weibull, W. (1951): A Statistical Distribution Function of Wide Applicability, Journal of Applied Mechanics, 18(3), 293-297.

Witnauer, W. D., R. G. Rogers and J. M. Saint Onge (2007): Major League Baseball Career Length in the 20th Century, Population Research and Policy Review, 26(4), 371-386.

Wooldridge, J. M. (2013): Introductory Econometrics – A Modern Approach, 5th ed., South-Western: Cengage Learning.

Wozniak, D. (2012): Gender Differences in a Market with Relative Performance Feedback: Professional Tennis Players, Journal of Economic Behavior and Organization, 83(1), 158-171.

Wozniak, D., W. T. Harbaugh and U. Mayr (2010): Choices about Competition: Differences by Gender and Hormonal Fluctuations, and the Role of Relative Performance Feedback, Working Paper, University of Oregon.

Zimbalist, A. S. (2002): Competitive Balance in Sports Leagues: An Introduction, Journal of Sports Economics, 3, 111-121.

Zimbalist, A. S. (2003): Competitive Balance Conundrums: Response to Fort and Maxcy's Comment, Journal of Sports Economics, 4, 161-163.

Zitzewitz, E. (2006): Nationalism in Winter Sports Judging and its Lessons for Organizational Decision Making, Journal of Economics and Management Strategy, 15(1), 67-99.

INTERNET SOURCES

Association of Road Racing Statisticians
http://www.arrs.net

Betexplorer.com
http://www.betexplorer.com

Bundesligainfo.de
http://www.bundesligainfo.de

DFB annual financial report 2012/13
http://www.dfb.de/uploads/media/Saisonreport_3.Liga_1213_01.pdf

DFB 2013
http://www.dfb.de

German Football Association (DFB)
http://www.dfb.de

Handball-World.com
http://www.handball-world.com

International Ski Federation
http://www.fisskijumping.com

Ironman World Championship
http://www.ironman.com
http://www.ironmanworldchampionship.com

Kicker Online
http://www.kicker.de

Ski-Jumping World Cup Results on Wikipedia
http://de.wikipedia.org/wiki/Skisprung-Weltcup

SoccerWay
http://www.women.soccerway.com

Statto.com
http://www.statto.com

Swiss Alpine Marathon
http://www.swissalpine.ch

TennisDigital.com
http://www.tennisdigital.com

The “Rec.Sport.Soccer Statistics Foundation”
http://www.rsssf.com

Vasaloppet
http://www.vasaloppet.se

Weltfussball.de
http://www.weltfussball.de/schiedsrichter/bundesliga

Women's Tennis Association
http://www.wtatennis.com

Worldloppet
http://www.worldloppet.com

EIDESSTATTLICHE ERKLÄRUNG

Hiermit versichere ich, Friedrich Scheel, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben. Alle Stellen, die wörtlich oder sinngemäß veröffentlichtem oder unveröffentlichtem Schrifttum entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Paderborn, 07.10.2014 Friedrich Scheel