IN OMNIBUS VERITAS

UNIVERSITÄT MANNHEIM

Essays in Applied Microeconomics

**Inauguraldissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim**

Johannes Dittrich

Herbst-/Wintersemester 2022

Abteilungssprecher:	Prof. Klaus Adam, Ph.D.
Vorsitzender des Prüfungsausschusses:	Prof. Nicolas Schutz, Ph.D.
Referent:	Prof. Dr. Martin Peitz
Korreferent:	Prof. Helena Perrone, Ph.D.
Tag der mündlichen Prüfung:	15.09.2022

Eidesstattliche Erklärung

Hiermit erkläre ich, Johannes Dittrich, dass ich die vorliegende Dissertation selbstständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben habe.

München, 30. Juni 2022 Johannes Dittrich

Acknowledgments

First and foremost, I would like to thank my supervisor Martin Peitz for his continuous support and guidance during the my doctoral studies in Mannheim. I am very grateful for discussions about my research and ideas, the constructive criticism, the highly valuable advice on the chapters in this dissertation and his incredibly helpful career related advice. I am also very grateful to Helena Perrone and Kathleen Nosal for their support and the highly useful comments on my research.

I would also like to thank the audience and organizers of the IO and theory seminar as well as the CDSE seminar at the University of Mannheim for regularly giving me the opportunity to present my research. Thank you for your valuable comments and suggestions. I have also benefitted greatly from the opportunity to present parts of my research at local MaCCI conferences, which I am very grateful for.

A very special thanks goes to my fellow doctoral students in Mannheim for the enjoyable time we had together. Many of them have become good friends over the years, which I am very thankful for. A special mention goes to Johannes Poeschl and Matthias Hölzlein for making our exchange year at Berkeley such an enjoyable experience.

I would also like to thank the GESS and CDSE administration, especially Marion Lehnert and former center manager Sandro Holzheimer for assisting with the administrative efforts. You have saved me a lot of time during my studies. Moreover, I would like to thank the members of my long-time football club Heidelberger SC. The numerous training sessions and match days have been a welcome change to the daily academic routine. Even after sometimes long periods of non-appearance I have always felt welcome. You guys are not just a football club but a truly remarkable community.

I also want to thank my colleagues at CRA for supporting and encouraging me to finish my dissertation.

I also thank my parents, my siblings Theresa and Paul and my grandparents as well as my good and long-term friends back home for always encouraging and supporting me.

Last but not least, I am grateful to Katharina Momsen, who has helped me to get through the numerous struggles and times when finishing my PhD seemed like a far-fetched and truly impossible task to accomplish. This dissertation would not exist without your continuous support and encouragement. Thank you.

Contents

List of Figures

List of Tables

Preface

Over the past decades, modern economies have become increasingly information-intensive. Recent debates about the market power of platform providers, whose competitive advantages can be tied to the overarching ability to gather and combine information, demonstrate that information has become a key resource in society. The so-called “Information Age” has opened the opportunity to very efficiently generate and acquire information on nearly every subject we are interested in. Many decisions we make in our daily lives are only taken after careful consideration and assessment of the situation as well as the advantages and disadvantages of the numerous available alternatives.

While the interest in gathering information is huge, its impact on decision making depends on several factors. First and foremost the information that is most prominently presented to consumers does not necessarily need to be most reflective of the quality of goods or services. In a world where information is the main driver of purchase decisions, sellers have an inherent incentive to be strategic and selective in the presentation of information. This has been shown by researchers in many markets and settings. Firms tend to present and influence information such that it gives a favorable impression of their product, platforms have a tendency to cater to their user’s preferences or prior beliefs and consumers sometimes fail to correctly inform their decisions based on observable information.

But even if the consumer is presented all the information she could possibly base her decisions on, she may still fail to make the ex-post correct choices. In fact, most people tend to have a hard time to correctly interpret and asses the multitude of news they receive on a subject. They tend to overvalue certain signals or focus too much on a particular aspect, which leads to wrong decisions even in the presence of perfectly observable and accessible information. This not only shows that studying the impact of information on decision making is an important subject but also that interference between informational quality and observed decisions can occur on both sides of the market. When aiming to analyze the impact of informational quality on decision making it is therefore essential to carefully assess what kind of possible distortions could occur on which market side and what the potential consequences could be.

This dissertation contains three self-contained articles, each of which takes a different angle on the presentation of information and its impact on consumer decision making. As a common theme, all of these articles are concerned with the impact of information on decision making. I try to shed light on the question of informational quality in a variety of settings from an empirical angle. I provide new evidence that (i) sellers may be able to strategically affect consumer review scores through lower prices and may therefore be able to change the public image of and future demand for a product; (ii) prices in online sports betting may exhibit signs of the so-called hot hand effect or "streakiness" even though such beliefs are not supported by the underlying player performance statistics; and (iii) the degree to which individual player performance is affected by such a hot hand effect depends on their skill level and experience with a particular setting.

In all of my three chapters, I use web scraping and tracking techniques to construct novel datasets that enable me to provide empirical evidence on research topics which are difficult to approach with available standard datasets. As the internet continues to play a major role in our daily lives, an extensive amount of data generated is also made publicly available. The rise of online shopping and trading has drastically improved the observability of product prices, characteristics, recommendations in the form of reviews and in some cases even individual level purchasing information. The multitude of information that firms and consumers share voluntarily via the internet can be a valuable addition or even alternative to standard survey or commercial datasets. Although recently many industrial economists and applied econometricians have picked up on the trend to work with the internet as a new and exciting data source, I believe that much more can be done in this regard. Working with self-recorded or self-tracked data poses new and interesting challenges that go well beyond standard data collection work but in many cases provides the unique opportunity to construct very detailed datasets.

However, these methods do not only provide the possibility to exploit new data sources but also allow for construction of dynamic datasets that allow to very precisely track and study consumer and firm behavior even over short timeframes. Nowadays, the data that becomes available is not only much broader but also arrives at a higher frequency. Tracking and scraping techniques allow to record changes to data sources in self-chosen time intervals. This enables the researcher to potentially link changes in various data sources to each other and create highly dynamic datasets which offer interesting new perspectives and possibilities for empirical work.

The first chapter of this dissertation empirically analyzes the link between purchasing prices and online review scores. In the past years, online user reviews have become a major source of information for consumers. Almost all consumers consult experts or reviews before making an important purchase decision. The "Annual BrightLocal Consumer Review Survey 2022" finds that 77% of consumers read online reviews to gather information about local businesses. Consequently, it is crucial for sellers to both retain a good reputation and offer highly rated products. Since direct manipulation of reviews is risky and quality improvements are costly and time-consuming, lowering prices may be an effective and efficient channel through which sellers can affect review scores: If prices are low, the net utility of consumers is higher if they follow an evaluation approach that is based on value-for-money and reviews should be more favorable. Except for this direct effect of prices on review scores, there may also be a countervailing selection effect: If prices are low, the probability of consumers with low valuations buying the good increases and hence average review scores may decrease because those consumers inherently have a lower valuation of the product relative to consumers who would also be willing to buy at higher prices. To empirically analyze how sales prices affect review scores, I use web scraping and tracking techniques to construct a dataset that matches individual-level purchasing data and review data from a large digital distribution platform for video games. I find that the effect of discounts on review scores depends on the magnitude of the price decrease. While small discounts increase review scores by a significant margin, large discounts have a negative effect on review scores. This indicates, that there may indeed be two countervailing effects at play and that their relative strength may depend on the magnitude of the discount that is given. This suggests the existence of a reputation-maximizing discount level for sellers that should not be exceeded if the primary aim of a price decrease is to boost review scores. For consumers, the results suggest that the purchase price is an essential input for the interpretation of reviews. Consequently, consumers should benefit from disclosure of purchase prices alongside reviews and review scores.

The second chapter investigates belief formation in sports betting markets for US Open 2012 tennis matches. Using high frequency sports betting data from the world's largest betting exchange and a statistical model of tennis play, I study if odds ratios are consistent with the belief in a hot hand in professional tennis. Estimating a dynamic random effects Probit model, I test for path dependence of winning-on-service probabilities in professional tennis matches and show that treating point-winning probabilities as independent of the match's history is a good approximation. I proceed by analyzing match-winning probabilities implied by prices on the Betfair betting exchange and find that they over- and under-react to certain news events. Bettors seem to falsely believe in the so-called hot hand. The current server's probability of winning a point is believed to be higher if he was able to secure the last point. Consequently, the match-winning probabilities implied by prices on the betting exchange are sometimes too high for the point-winner and too low for the opponent. The fact that prices do not always reflect the underlying probabilities of events casts doubt on the longstanding assumption that betting markets, much like financial markets, are generally efficient.

The third chapter takes a more focused perspective on the hot hand and its interaction with experience and skill level. Most of the economic literature has followed Gilovich et al. (1985) and studied the hot hand in a number of fairly narrow settings like professional sports. I offer a more in-depth analysis of the hot hand effect in a setting that allows to investigate how the hot hand changes with skill level and experience. I use an extensive dataset from one of the most popular online games to demonstrate that the hot hand effect is high for players with low experience but decreases substantially as players play more games and familiarize themselves with the game and setting. Similarly, player skill level substantially influences the hot hand effect. These findings are robust to alternative definitions of the hot hand and are based on a large dataset that guarantees sufficient statistical power. The results presented should not be neglected in the discussion about the hot hand effect and may help explain why many people still believe in a hot hand effect based on their own experiences even though the evidence for its existence, for instance in professional sports settings, is mixed.

Chapter 1

The Effect of Prices on Online Review Scores[*]

1.1 Introduction

In recent years, the impact of online reviews on consumers' purchasing decisions has risen tremendously. While obtaining advice pre-purchase has always been a major factor in the market for experience goods, the internet has not only simplified the information-gathering process but also vastly increased the amount of information that is available to consumers. One source of information that has become increasingly important is the "digital word-of-mouth" in the form of consumer reviews. According to the "Annual BrightLocal Consumer Review Survey 2022", 77% of consumers "always" or "regularly" read online reviews to gather information about local businesses.[1] A survey by Power-Reviews reveals that 95% of consumers use online reviews as a source of information, with 86% even stating that reviews are essential for their purchase decisions and 56% of consumers even specifically visit review websites.[2] These facts make online reviews not only an important informational device for consumers, but also a powerful marketing tool for sellers. Nowadays, almost all online sales platforms provide their own review systems and put extensive efforts in improving their functionalities.

[*] I am grateful to Martin Peitz, Helena Perrone, Kathleen Nosal, Emanuele Tarantino, seminar audiences at the EEA 2018 in Cologne and the IO and Theory seminar in Mannheim for their valuable comments and suggestions.

[1] See www.brightlocal.com/learn/local-consumer-review-survey (accessed on 28/06/2022).

[2] See www.powerreviews.com/blog/survey-confirms-the-value-of-reviews (accessed on 28/06/2022).

The fact that consumer reviews have become an essential feature of online sales platforms and are the main source of information used by consumers in markets for experience goods, the presence of review systems can be a double edged-sword for sellers: On the one hand, review systems are regarded as an essential feature of platforms and are inevitably required to draw consumer attention. As a consequence, online sales platforms heavily push their review systems. On the other hand, it has been shown that negative reviews and low review scores substantially affect product sales (see e.g. M. Anderson and Magruder, 2012; Cabral and Hortaçsu, 2010; Chevalier and D. Mayzlin, 2006) and can therefore be detrimental to a seller's long-term success. Consequently, a natural incentive for sellers arises to keep review scores high.[3]

One way through which sellers could possibly generate higher review scores is being strategic about their pricing.[4] If consumers evaluate their satisfaction with purchases based on a consideration of value for money, they will likely also give better reviews when paying less. I will refer to this as the *price effect.* However, prices will not only affect consumer satisfaction with a product but will also have an effect on consumers' purchase decision. Notably, when there is heterogeneity in consumers' tastes for a product, changing prices may not only affect the evaluation of those who are already buying the product (and will continue to do so) but also the composition of the consumer population. With a price decrease, for instance, a product may become attractive to those who were not quite convinced by it before based on a value-for-money consideration, e.g. because it did not quite match their taste. Such a pricing-induced change in the selection of consumers deciding to buy a product may affect review scores beyond the value-for-money consideration. I will refer to this as the *selection effect.* Specifically, price decreases may attract consumers with on average lower valuations for the product, which in turn will be less likely to leave a positive review. This additional selection effect may counteract the direct price effect of higher value per money such that the net effect

[3] It should be noted that short-term incentives of platforms and sellers may not necessarily be aligned if they are separate players. While platforms have an interest in review scores that represent seller quality to screen good sellers and enhance consumer satisfaction, sellers primarily aim for high review scores to boost demand. If the review system functions properly, this potential conflict of interest should resolve in the long term as low quality sellers exit the market.

[4] Of course, for sellers, various tools exist to keep review scores high, most notably increasing the product's quality or manipulating scores directly. While the former is not always possible or can be too costly, especially in the short run, the latter is nowadays widely regarded as very risky and can lead to severe reputation losses in the long term. B. D. Mayzlin et al. (2014) find some evidence of review manipulation on online hotel booking sites but their results do not necessarily transfer to product reviews and are less pronounced for chains and franchises, suggesting that there exists a severe negative reputation effect, especially for large sellers. With purchasing verification systems being introduced by many online platforms (e.g. Amazon and Steam), it seems highly unlikely that sellers can effectively manipulate review scores through fake reviews.

of price changes on review scores is dampened or even reversed and the strategy to set lower prices to boost review scores may be less effective or could even backfire.[5]

There is a small but growing literature that studies the effect of price changes on review scores and ratings (see e.g. Carnehl et al., 2021a; Luca and Reshef, 2021; Zegners, 2019). I contribute to the literature by investigating the relationship between prices and review scores for video games sold on the online platform Steam. Steam frequently runs sales, granting discounts of varying degrees on games sold on the platform. To shed light on the relationship between prices and review scores, I construct a novel dataset using web tracking and data scraping techniques, which allow me to link individual level purchasing and review data for a large digital distribution platform. Relating review scores to the prices consumers have paid for the product they are reviewing enables me to isolate and estimate the effect of prices on average review scores at different price levels.

A major challenge in consistently estimating the effect of prices on review scores is disentangling the effect from other influencing factors. Most notably, product markets like the one I am studying are characterized by heterogeneity of both products as well as consumers. Specifically, products (video games in my setting) will differ with regard to their quality and consumers will differ with regard to their standards and individual reviewing behavior. Both sources of heterogeneity will lead to endogeneity problems if they are not addressed in the estimation. Differences in video game quality will naturally be correlated with review scores but also with prices if lower quality games receive larger discounts. Differences in consumer rating behavior will also be correlated with both review scores and prices e.g. if consumers buying at certain discount levels (like for instance a passionate gamer who always buys the newest games) show a specific rating behavior (for instance the same passionate gamer who is very critical with new games). I deal with both sources of heterogeneity by constructing multiple game- and reviewer-level control variables aimed at capturing those game- and reviewer-specific effects such that the remaining variation can be used to identify the effect of prices on review scores.

A further challenge in consistently estimating the effect of prices on review scores is the potential presence of a selection effect that results in reviews specifically being written only when the reviewer was more or less content with her purchase. This effect is difficult to address as it would require information on a reviewer's ex-post satisfaction with a product relative to her individual ex-ante expectations to understand when she is

[5] Carnehl et al. (2021b) illustrate the occurrence of these two effects in a dynamic pricing model.

likely to write a review. Although I do not address this issue econometrically, I argue that it is of second order in my analysis as I specifically select a sample of frequent reviewers, i.e. consumers who inherently reveal a high probability to review and therefore should be less driven by negative and/or positive experiences in their reviewing behavior. Notably, a focus on very active reviewers should not diminish the relevance of my results as such since every seller-strategy aimed at enhancing review scores would anyways affect outcomes mainly via the very active reviewers on a platform.

This chapter contributes to the literature in two ways. Firstly, I construct and use a novel dataset that allows me to link purchasing and review data in a setting with relatively high price fluctuation. Prices on most online sales platforms are very volatile and sales can be short-lived, especially if sellers are capacity-constrained. Hence, even if reviews were written very close to the date of purchase, it is oftentimes difficult to determine at which price reviewed products were bought. The dataset I use allows to establish this link directly for individual transactions and therefore enables me to provide robust evidence on the link between prices and review scores.[6] Secondly, the fact that I observe a high degree of price variation allows me to study the effect of prices on review scores for a wide range of discount levels. Carnehl et al. (2021a) as well as Luca and Reshef (2021) find a negative relationship between price and review score whereas Zegners (2019) finds a positive relationship. However, it is not clear that their results are comparable to each other. Notably, both Carnehl et al. (2021a) and Luca and Reshef (2021) obtain their results for variation in positive (non-zero) prices, while the result in Zegners (2019) originates from comparing normally priced to entirely free ebooks.[7] In contrast, the dataset I use allows me to analyze a wide range of discount levels. This is relevant because a possible explanation for the seemingly contradictory results in Carnehl et al. (2021a), Luca and Reshef (2021) and Zegners (2019) might be that the direct price effect based on a value-for-money consideration dominates the consumer selection effect for “normal” discount levels, whereas the opposite is true for higher discount levels. The dataset I use allows me to test this hypothesis in the same setting.

Indeed, I find that the relationship between price discounts and review scores on Steam is non-linear. While the probability of receiving a positive review score increases for low discount levels, it decreases again for very high discount levels. Higher discounts

[6] This is e.g. in contrast to Carnehl et al. (2021a) who are only able to match monthly aggregated review scores to pricing data.

[7] The exact range of price changes underlying the analyses in Carnehl et al. (2021a) and Luca and Reshef (2021) is not clear. However, it can be expected that the normal range of discounts is relatively narrow given the settings analyzed.

lead to higher review scores for low to moderate discount levels, indicating that the selection effect is either positive or at least outweighed by the price effect. In contrast, for high discount levels the positive effect of discounts on review scores is lower and ultimately negative for high discounts (i.e. very low prices). This implies that there may be a “reputation-maximizing discount level” from a seller’s perspective. In my setting, I estimate that the probability of obtaining a positive review score is maximized at a discount level of around 43%. At this discount level, the probability of receiving a positive review score is increased by about 8.45 percentage points. This indicates that the dynamic pricing problem a seller faces is much more complex in the presence of review systems through a potential second-order feedback effect of prices on demand through review scores.

The relevance of the relationship between prices and review scores goes beyond the dynamic pricing considerations of sellers. Indeed, this information is potentially helpful to consumers as well. If consumers use reviews as a source of information about a product[8], the price at which the reviewed product was bought is a valuable piece of information on which inference could be based. When price level and product quality are substitutes in the consumer’s utility function, it can be very hard for her to infer both from review scores, especially when the product price was volatile and reviews contain no or very little information about purchasing prices. Whether a good rating was based on a relatively low price or high product quality can thus be very difficult to assess for consumers. If the main goal of a review system is to provide helpful information to consumers (e.g. to increase consumer satisfaction with a platform in general or reduce return rates), releasing price information related to reviews could improve consumer decision making if review scores actually varied with prices. While some platforms have started to tag reviews that were written based on free copies (e.g. by professional reviewers or after free giveaways), it often remains uncertain at which price other reviewers have bought the product. My results point towards a strong and non-linear relationship between prices and review scores, suggesting that information on purchase prices would be highly valuable for consumers in interpreting reviews.

The remainder of this chapter is structured as follows: Section 1.2 provides a brief review of the literature on review and rating systems, Section 1.3 develops the theoretical predictions that I will test empirically. Section 1.4 discusses data collection and data matching, while Section 1.5 provides a descriptive discussion of the dataset. Section 1.6

[8] The idea that reviews can be utilized as a means to transmit information about a product has been discussed recently by Carnehl et al. (2021b) and Martin and Shelegia (2021).

discussed the empirical model and the estimation procedure. Section 1.7 presents and discusses the empirical results and Section 1.8 concludes.

1.2 Literature

This chapter relates to several strands of the literature on rating and review systems.[9] In recent years, a number of papers have empirically investigated the link between prices and reviews or ratings. Zegners (2019) studies the effect of free giveaways on review scores using data from an online sales platform for ebooks. In his setting, semi-professional authors aim to build up a reputation by increasing the number and score of ratings via offering their ebooks for free. Comparing reviews for the same ebooks, he finds that review scores are lower for the free version which points towards the dominance of a (negative) selection effect. Carnehl et al. (2021a) study the relationship between prices and ratings on Airbnb and find that higher prices are associated with lower ratings in most rating dimensions (except the location rating), suggesting the dominance of a (value-for-money) price effect over the selection effect in their data. The authors also find evidence that entrants on the platform receive better value-for-money ratings and more bookings if they price relatively low, which then translates into the ability to charge higher prices and revenues later on. Luca and Reshef (2021) analyze the relationship between prices and ratings using data from Yelp. Similarly to Carnehl et al. (2021a), they find a significantly negative relationship between prices and online ratings, suggesting that a value-for-money consideration dominates customers' rating behavior.

Apart from the difference in the settings analyzed by Zegners (2019), Carnehl et al. (2021a) and Luca and Reshef (2021), the main distinction between the studies is with regard to the results found by these authors. While Zegners (2019) finds a positive relationship between prices and ratings (in the sense that lower prices result in worse ratings), both Carnehl et al. (2021a) and Luca and Reshef (2021) find a negative relationship. Notably, the degree of price variation analyzed is also very different. The result in Zegners (2019) hinges on comparing a price of zero to the full price of a product, i.e. estimating the effect for an extreme discount of 100%. In contrast, in both Carnehl et al. (2021a) as well as Luca and Reshef (2021) the results are obtained based on a continuous variation of prices in a presumably more normal range. This raises the question whether the relative strength of the (value-per-money) price effect and the selection effect, and therefore the direction of the net effect of prices on rating scores, depends on the price

[9] For surveys of this literature see e.g. Dranove and Jin (2010) and Cabral (2012).

level itself. I contribute to the discussion by analyzing a wide range of discounts, which allows me to investigate the strength and direction of the effect for different discount levels. Indeed, I find that the effect of prices on review scores on Steam is non-linear in the discount level, with a negative correlation arising for low discount levels and a positive correlation arising for high discount levels. This is consistent with both the findings in Zegners (2019) as well as Carnehl et al. (2021a) and Luca and Reshef (2021), suggesting that beyond the difference in the settings analyzed, the degree of price variation also plays an important role for analyzing the relationship between prices and rating or review scores.[10]

Recent theoretical contributions have also established the importance of reviews for firm pricing in a dynamic setting. Martin and Shelegia (2021) analyze how introductory pricing may increase future profits via inducing high ratings upon launch of a new product. In their two-period model firms can induce lower expectations about product quality by setting low prices in the first period, which leads to more favorable reviews and higher profit in the second period but lower profits in the first period. Carnehl et al. (2021b) analyze the dynamic pricing strategy of a firm when information is transmitted across periods via ratings and consumer inference is partially based on aggregate ratings. In their multi-period model a similar trade-off between current and future profits leads to firms pricing lower or higher compared to the myopic optimum depending on the underlying assumptions on consumers' behavior. In both papers the results crucially depend on the interaction of rating scores and prices, which is *a priori* ambiguous.

There is also a related stream of literature that has focused on the effects of lower prices on feedback and review scores more broadly. In a laboratory setting, L. (Li and Xiao (2014) show that rebates, which buyers receive from sellers for committing to giving feedback, increase the probability of purchase as well as the probability of feedback in a buyer-seller-trust game (and hence market efficiency). Yet, this observation cannot be confirmed in a real-world setting. Cabral and L. Li (2015) conduct a field experiment on eBay by selling a homogenous product (USB-stick) while varying transactional quality. Offering buyers a payment in exchange for their feedback, they find that higher monetary rewards neither increase the likelihood of feedback nor the bidding behavior itself. Yet, when offering a rebate, the likelihood of negative feedback decreases if transaction quality is low. This finding suggests that buyers indeed tend to rate products relative to the quality they receive.

[10] In my analysis I omit free giveaways as the effect of those could still be very different – even when compared to (very) high discount levels. See e.g. Shampanier et al. (2007) for evidence that consumers' evaluation of free products is very different.

Beyond the contributions discussed above, there is a large and growing literature that studies the impact of consumer reviews and ratings on demand, highlighting the economic importance of consumer review systems in a wide variety of settings. Chevalier and D. Mayzlin (2006) compare book sales ranks on Amazon and Barnes & Noble and show that more positive reviews and a higher number of reviews lead to higher sales ranks. Cabral and Hortaçsu (2010) collect data from eBay, and find that sellers' weekly sales rate grows by 5% until they first receive negative feedback, which not only leads to a decrease of sales by 8%, but also increases the rate at which they receive negative feedback and therefore increases their likelihood of exit.[11] Utilizing data on reviews from Yelp and data on restaurant reservations, M. Anderson and Magruder (2012) estimate the effect of an additional half-star on the rate at which restaurants are booked out to be 19 percentage points. In a related study, Luca (2016) estimates that an additional one-star on Yelp corresponds to an increase in restaurant revenues by 5-9%.[12] Closely related to these articles is the empirical literature on social learning and peer effects in markets for experience goods.[13] In a field experiment, Cai et al. (2009) find that demand for dishes in a restaurant increases by 13 to 20% if consumers are given ranking information. Moretti (2011) confirms these results using real-world data. He shows that for movies of unexpectedly high quality, social learning accounts for 32% higher box-office revenues.

A parallel stream of the literature establishes very similar results for expert reviews. Reinstein and Snyder (2005) find some evidence that more favorable expert reviews of newly released movies result in higher box office revenues.[14] Hilger et al. (2011) confirm the positive effect of expert reviews on demand using a field experiment. They find that wines with an expert rating label have 25% higher demand but, more importantly, this increase can be attributed to wines of a higher quality and thus higher ratings, while demand for wines with low ratings actually decreases. The common theme of these

[11] Similar results on the relationship between bidding behavior on eBay and seller reputation have been established by other authors, e.g. Melnik and Alm (2002), Jin and Kato (2006), Resnick et al. (2006), Lucking-Reiley et al. (2007) and more recently Klein et al. (2016). For an early survey of this literature see Bajari and Hortaçsu (2004).

[12] These results have also been confirmed in the marketing literature for numerous industries such as travel bookings (Dickinger and Mazanec, 2008; Vermeulen and Seegers, 2009; Ye et al., 2011), movies (Duan et al., 2008) or online sales platforms (Chen et al., 2004; Forman et al., 2008).

[13] See the seminal paper by Nelson (1970) for a discussion of the distinction between search and experience goods.

[14] Their results are very dependent on the genre and release schedule. While they find that more positive expert reviews induce 50 to 60% higher box office revenues for dramas and up to 37% higher box office revenues for narrow release schedules, there is no effect for wide release schedules and other genres.

articles is that the effect of review and rating scores on future demand is sizable and should be a major consideration for sellers.

Despite the clear relationship between review scores and demand, very few papers have studied the strategic incentives of sellers to improve review scores. There is a small literature that considers the strategic manipulation of review scores. Analyzing bidding and transaction data from an online intermediary for software development, Moreno and Terwiesch (2014) show that buyers and sellers trade off reputation and price.[15] In particular, buyers are willing to accept higher bids whenever the seller's reputation score is high. Similarly, sellers use their reputation to not only increase their probability of being selected but also their bids for projects. B. D. Mayzlin et al. (2014) find some evidence of review manipulation on online hotel booking sites. Their results show that the incentives to fake reviews are highest for independent hotels and lowest for franchises. Luca and Zervas (2016) come to a very similar conclusion in the case of Yelp reviews.[16] Whether these findings are transferable to product markets is questionable. Indeed, E. T. Anderson and Simester (2014) show that reviews that have been written by consumers who have not purchased the product are more likely to be fake reviews. Consequently, review systems that required consumers to purchase products before they can be reviewed, thereby drastically increasing the cost of manipulation, should show far less manipulated reviews.[17] Notable examples of this trend to higher standards of review fraud protection include Amazon's "verified purchase" tag and Steam's policy to only allow consumers who have purchased the product to write a review.

1.3 Model

In this section, I illustrate the potentially ambiguous effect of prices on review scores in a simple reduced form theoretical framework. As the focus of this chapter is to analyze the effect of prices on review scores, I will abstract from dynamic pricing considerations of the firm as well as market structure. Suppose that there exists a monopolistic seller (in the setting I analyze the online platform selling video games, Steam) who aims to sell an indivisible durable good (video games) to a population of buyers of mass 1. Let

[15] In this setting, sellers are developers who bid for tasks that are posted by buyers (e.g. firms or individuals).

[16] There exists a parallel computer science literature that uses content analysis to detect fake reviews. Notable examples include Ott et al. (2011), Xie et al. (2012) and Hu et al. (2012).

[17] Dellarocas (2006) also points this out in a simple theoretical model of review manipulation.

p be the price for one unit of the good. Consumer preferences are represented by the following linear utility function:

$$u_i(p, y_i, \delta, q, \gamma) = \delta q + y_i - \gamma p$$

Consumers consider whether or not to buy one unit of the good. Their utility of not buying is normalized to 0. I assume that video games are differentiated in two dimensions, where q represents the quality (vertical characteristic) of the game and y_i (horizontal characteristic) represents the type of variety. Consumers are homogenous in terms of their taste for quality,[18] but heterogenous with respect to their preferred variety, y_i. For simplicity, I assume that tastes for variety are uniformly distributed, i.e. $y_i \sim U\left[\underline{y}, \bar{y}\right]$. I also assume that the quality of the product is public information. This ignores effects through price signaling, which in principle could also play a role as they affect the population of consumers buying the product. However, introducing price signaling would only complicate the analysis while not adding much insight into the key mechanism. It is also questionable if the uncertainty about quality is very high in the market for video games. Typically, information on the quality of a game is readily available for professional reviews and reports about the game. Moreover, the quality of games typically does not change with price changes, especially if they are temporary. If anything, then price signaling should play a role in assessing the quality between differently priced games, which is not the focus of this chapter.

There are two decisions consumers face in this setting:

1. Whether or not to purchase the good: Consumers buy the video game whenever the expected utility they receive exceeds the utility from the outside option, i.e. $\mathbb{E}[u_i] = \delta q + y_i - \gamma p \geq 0$.
2. Conditional on purchasing the good, consumers decide if they want to write a positive or a negative review.

For simplicity I assume that every consumer who has purchased the good also writes a review. While this assumption is clearly restrictive and should be relaxed in a more

[18] Undoubtedly, the assumption that consumers are homogenous with respect to quality is restrictive and adopted to keep the model simple. Assuming that consumers value quality heterogeneously, i.e. $u(p, y_i, \delta_i, \gamma) = \delta_i q + y_i - \gamma p$, valuation for quality is distributed $\delta \sim U\left[\underline{\delta}, \bar{\delta}\right]$ and taste for variety and quality are uncorrelated leads to qualitatively very similar results.

complex model,[19] it fits the specific dataset I will use in the estimation. As the focus of this chapter is to estimate the probability of awarding a positive review score as a function of the purchasing price, the dataset consists of purchases for which I am able to observe a review and a price.[20] After having purchased the game, consumers evaluate their satisfaction with the product. This is given by

$$v_i(p, y_i, \delta, q, \gamma) = \delta q + y_i - \eta(p)p$$

where $\eta(p) \in [0, \gamma]$ reflects the degree to which consumers factor the price they payed into their reviewing decision. I assume that the degree to which the price is factored into the reviewing decision may depend on the price itself. Indeed, consumers may put less weight on the price when a game was cheap relative to when a game was very costly. I further assume that $\frac{\partial v_i}{\partial p} \leq 0$, i.e. price decreases always generate a net value increase for consumers. Notably, with $\eta(p) = 0$ the price does not play any role in the reviewing decision while $\eta(p) = \gamma$ corresponds to the case where consumers evaluate their net utility for the review. I assume that consumers write a positive review whenever they generate enough satisfaction from the use of the product, i.e.

$$v_i(p, y_i, \delta, q, \eta(p)) \geq \bar{v}$$

where $\bar{v}$ is the cutoff value that determines whether or not the consumer writes a positive or negative review. This is equivalent to saying that the value of v_i reflects how happy the consumer is with her purchase and that $\bar{v}$ represents a standard of satisfaction above which the consumer is willing to provide a positive review. This is a technical assumption that is necessary to generate negative reviews as otherwise everyone who purchases the

[19] In fact, in Section 1.5 I will show that the probability of writing a review is not only much smaller than 1 but also depends on the price a consumer has payed. Products which are bought on a heavy discount have a much lower probability of being used and are therefore reviewed with a much lower probability.

[20] It is important to note that consistently estimating the effect of prices on average review scores requires a much broader dataset. While I am convinced that I am able to very precisely estimate the effect of price changes on *individual* review scores, estimating the effect of prices on *average* review scores also requires knowledge of the probability of writing a review as well as the elasticity of demand. Whereas I will later on discuss estimates of the former, an important caveat is that my dataset is composed of "active reviewers" who are most likely not representative of the consumer population. Hence, these estimates do not necessarily represent the true probability of receiving a review from a representative consumer. Similarly, demand estimates based on my data could differ substantially from true demand if the purchasing behavior of my users is significantly different from the rest of the consumer population. Consequently, consistently estimating demand and the probability of receiving a review would require tracking an additional parallel dataset composed of a random sample of "active users", which goes beyond the scope of this chapter.

product will provide a positive review as $\eta(p) \leq \gamma$. However, I do not consider it to be too far from reality. Indeed, many rating systems are more nuanced than the one I use data from for the analysis in this chapter (Steam), which only allows for positive and negative rating as opposed to e.g. Amazon's which allows for 1 to 5 stars or the system used by booking.com which allows for ratings between 1 to 10 points. It is therefore likely that Steam's binary ratings do not reflect the extreme points of a more granular rating scale but would rather correspond to the respective lower and upper parts.[21] Hence, the marginal type $\hat{y}$ that is indifferent between buying and not buying the video game is characterized by:

$$\hat{y} = \gamma p - \delta q$$

The marginal type $\tilde{y}$ that is indifferent between writing a positive or negative review is given by

$$\tilde{y} = \eta(p)p - \delta q + \bar{v}.$$

I assume that $\bar{v} \geq \gamma p - \eta(p)p$, $\gamma p - \delta q \geq \underline{y}$ and $\eta(p)p - \delta q + \bar{v} \leq \bar{y}$ such that $\underline{y} \leq \hat{y} \leq \tilde{y} \leq \bar{y}$. The fraction of consumers writing a positive review after purchase, i.e. the probability of receiving a positive review score at a given price, represents the review score:

$$r = \frac{1 - F(\tilde{y})}{1 - F(\hat{y})} = \frac{\bar{y} - \tilde{y}}{\bar{y} - \hat{y}} = \frac{\bar{y} - (\eta(p)p - \delta q + \bar{v})}{\bar{y} - (\gamma p - \delta q)}$$

The effect of price changes on review scores can be separated into two potentially opposing effects: a price effect that reflects that a lower price will increase the utility of consumers purchasing the product and increase the likelihood of a review and a selection effect that reflects that the population of consumers that purchase the product also changes with a change in price.

$$\frac{\mathrm{d}r}{\mathrm{d}p} = \underbrace{\gamma \frac{\bar{y} - \tilde{y}}{(\bar{y} - \hat{y})^2}}_{\text{selection effect } \geq \, 0} - \underbrace{\left(\eta(p) + \frac{\partial \eta}{\partial p}\right) \frac{\bar{y} - \hat{y}}{(\bar{y} - \hat{y})^2}}_{\text{price effect } \geq \, 0}$$

With lower prices, for a given customer population, each consumer is more likely to write a positive review as his ex-post valuation of the product, v_i, is higher. On the other hand, lower prices also have the effect that consumers with lower valuations will purchase the

[21] An alternative interpretation would be that $\bar{v}$ represents the costs of reviewing the product.

product. A priori, these additional customers are less likely to give a positive review as they have a lower valuation of the product at any given price.

Both effects drive the change of review scores into different directions. On the one hand, the price effect will lead to higher review scores with decreasing prices for the reason that consumers ex-post valuation will be more favorable. On the other hand, the selection effect will lead to lower review scores with decreasing prices as ex-post valuations of consumers with lower taste for the specific game are generally less favorable. The price effect therefore leads to a shift of all valuations being more positive, while the selection effect leads to the overall mix of valuations being more negative. The net effect depends on the relative strength of both effects and may notably also change if the relevance of the price for the review decision changes with the price level itself.

Two special cases are worth mentioning. If the price plays no role in the review decision, i.e. $\eta = 0$, there will only be a selection effect that leads to price decreases having a negative effect on review scores. In this case the price has no direct effect on the review score for a given customer type, as customers do not value the price decrease itself for their review decision. However, the price decrease results in consumers with lower valuations purchasing the product, which has a negative effect on the average review score as these consumers are more likely to give negative reviews. If consumers evaluate the product according to their net utility, i.e. $\eta = \gamma$, the price and the selection effect cancel out such that the net effect of price change on the review score is equal to 0.

1.4 Data

I collect price, review and purchasing data directly from the Steam Store[22] and the Steam Web API[23] using web scraping and tracking techniques. Steam is a digital distribution platform[24] for video games operated and developed by the Valve Corporation. Introduced in 2003, Steam has grown steadily (see Figure 1.1) and become the leading platform with a market share of around 50-70%[25] among digital distribution platforms for PC

[22] All data was collected from the official Steam web page store.steampowered.com.

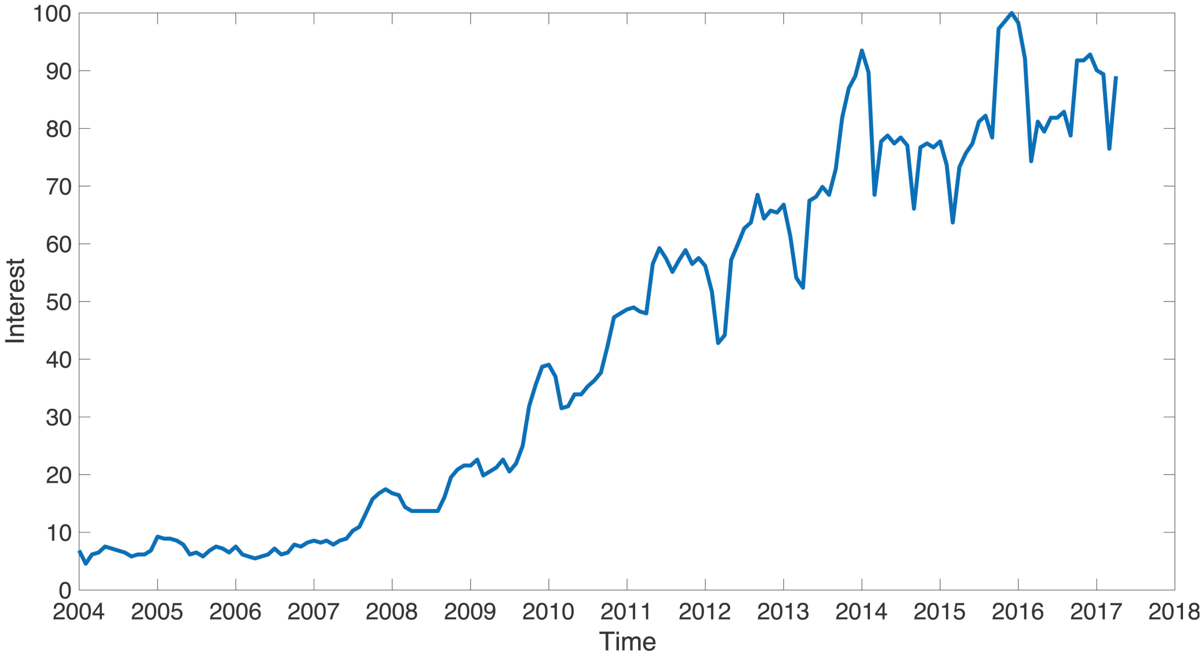
[23] The Steam Web API provides user- and game-level information from the Steam community website in an easily readable and processable format.

[24] In fact, Steam also offers other services to users such as digital rights management, social networking and online matchmaking.

[25] Measuring Steam's market share is not trivial as Valve is not very vocal about total sales.

gaming.[26] The platform had around 125 million registered accounts[27] and an estimated total revenue of around $3.5 billion in 2016.[28] To date, almost all PC video games are sold via the Steam platform either exclusively or in addition to retail or other digital distribution channels. In December 2016, a total of around 11,000 games[29] was available for purchase through Steam with around 5,245 being added in 2016 alone.[30] A unique feature of the Steam community webpages is the observability of users' games libraries, which consist of all video games that have been purchased directly through the store or

Figure 1.1: Index of frequency of google web searches for the term "steam store"



Source: Google Trends (accessed on 07/07/2017)

[26] The market for digital distribution of PC games itself has grown significantly in the last decade with the main reason being that gross margins for publishers and developers are about twice as high compared to retail channels. The total market size is estimated to be around $4 billion. All numbers according to Forbes (www.forbes.com/forbes/2011/0228/technology-gabe-newell-videogames-valve-online-mayhem.html, accessed on 28/06/2022).

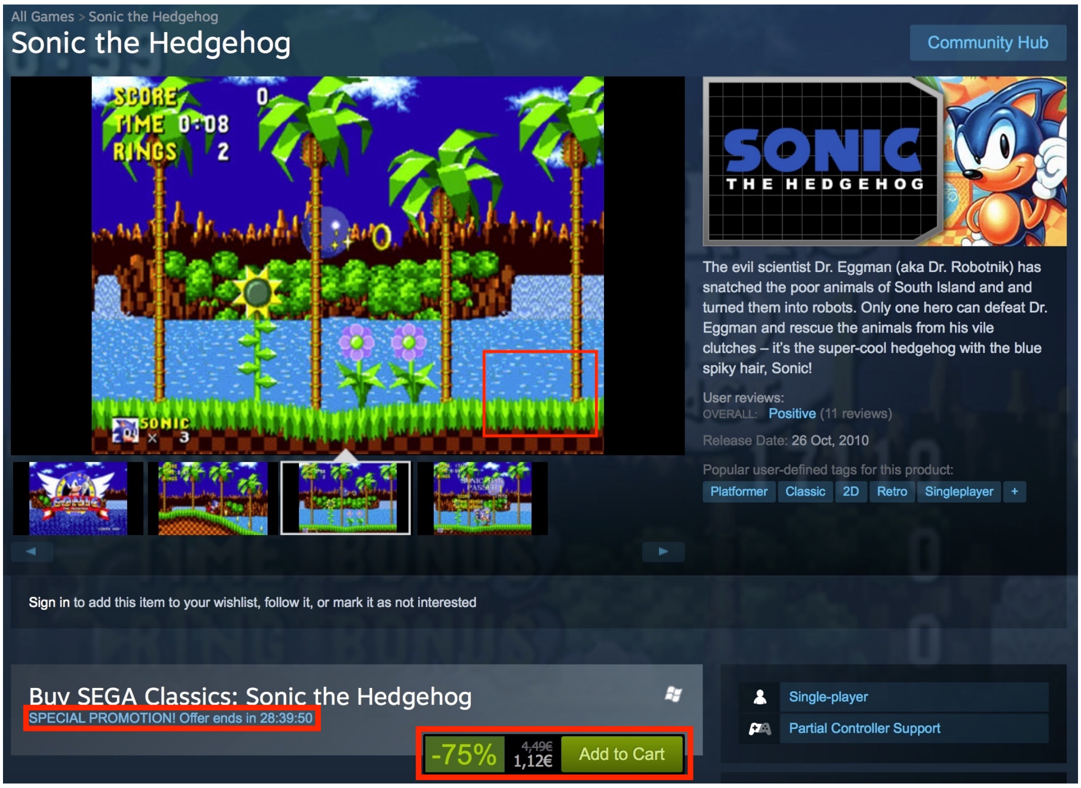
[27] This estimate is according to the website www.vg247.com/2015/02/24/steam-has-over-125-million-active-users-8-9m-concurrent-peak (accessed on 28/06/2022).

[28] This estimate is according to the Steam tracking service "Steam Spy": galyonk.in/steam-sales-in-2016-def2a8ab15f2 (accessed on 28/06/2022).

[29] See www.polygon.com/2016/12/1/13807904/steam-releases-2016-growth (accessed on 28/06/2022).

[30] A key component to the recent success of Steam is the platforms continued support of "independent developers" by offering them a well established and easy to handle digital distribution channel. As a result, Steam has become the prime release platform for small budget video games.

Figure 1.2: Example of Steam web store page



All Games > Sonic the Hedgehog

Sonic the Hedgehog

The evil scientist Dr. Eggman (aka Dr. Robotnik) has snatched the poor animals of South Island and and turned them into robots. Only one hero can defeat Dr. Eggman and rescue the animals from his vile clutches – it's the super-cool hedgehog with the blue spiky hair, Sonic!

User reviews:
OVERALL: Positive (11 reviews)

Release Date: 26 Oct, 2010

Popular user-defined tags for this product:

Sign in to add this item to your wishlist, follow it, or mark it as not interested

Buy SEGA Classics: Sonic the Hedgehog

SPECIAL PROMOTION! Offer ends in 28:39:50

-75% 4,49€ 1,12€

Source: Steam web store (accessed on 03/04/2017)

activated via a digital key.[31] It is therefore possible to access and collect games libraries from all user pages that are not explicitly privacy protected.[32] Even though the existence of privacy-protected user pages makes it impossible to track these users, the impact on the resulting dataset is rather small, as the ratio of privately to publicly observable pages is only about 10% and therefore rather low.

A supposed concern is the limited capacity of user pages that can be tracked daily. The Steam API allows up to 100,000 API calls[33] daily with a minimum difference of one second between each call which leads to a maximum of 86,000[34] games library pages that can be accessed and tracked daily. With the total number of active users estimated

[31] Many newly released video games directly rely on the services provided by Steam regardless of whether they have been purchased via Steam Store or other sellers (online or offline). These games need to be registered with a Steam account.

[32] More precisely, a user can block access to his games library either completely or grant access only to a subset of other users.

[33] Regarding users' games libraries this directly translates into a maximum of 100,000 libraries.

[34] There are some functionalities of the API that "cost" more than one call, which leads to the discrepancy between the maximum number of calls allowed and the "one-second-rule".

at around 12.5 million, it is hardly possible to track the entire user population. The capacity limitation is intensified by the fact that the average propensity of a user ever writing at least one review is as low as about 0.5%, which makes it impossible to build a sufficiently large dataset via tracking techniques if users are selected completely randomly from the population. Instead, I have built a dataset of 50,424 "active reviewers" by collecting all reviews that were written over the year 2016. I classify users of the website as "active reviewers" if they have reviewed at least two products in the past year. Moreover, users need to have set their profile status to "public" as otherwise their libraries are not observable.[35] This dataset covers only about 0.5% of the entire population on the platform but crucially includes the subpopulation which is most likely to review purchased products. It is therefore to be expected that the ability to only track a small number of users does not lead to results which are not representative. In fact, as the majority of users do not write reviews at all, the expected additional information gained by tracking more user accounts is low. This database of "active reviewers" was tracked daily from March 2017 to July 2017. The tracking protocol is described in detail in the appendix of this chapter.

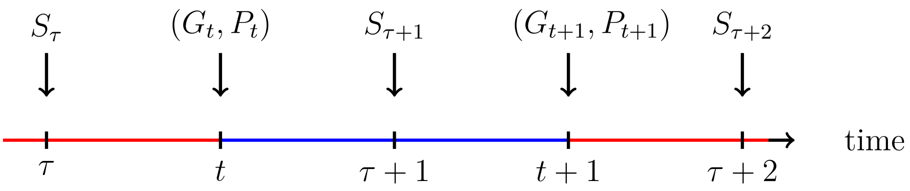
1.4.1 Matching: Prices

I identify additions to the games libraries as purchases by tracking changes in users' games inventories. As the purchasing prices are not observable directly, I need to collect price data separately and match it to the observed purchases. While this would be almost impossible for many online sales platforms as prices are fairly volatile, the pricing system on Steam possesses a number of features that enable a very robust and precise matching: Although publishers are technically free to set prices at their preferred level and change them at any point, a number of "store policies" make the prices structure fairly rigid. In fact, permanent price drops are very infrequent. Temporary price drops (sales) usually last several days. Throughout a sale period, discounts on the initial price are fixed[36] and quantities are not restricted. I assign sales prices to purchases only if they were guaranteed to happen within the sales period. Similarly, I assign full prices to purchases only if they were guaranteed to happen outside the sales period. Figure 1.3 illustrates the matching procedure.

[35] This could in principle lead to a selection effect if, for instance, more critical reviewers tend to keep their profiles private. However, only 5% of all users considered have set their profile to "private", probably due to the fact that profiles are public by default.

[36] There are very few reports about pricing errors which are typically fixed within minutes.

Figure 1.3: Matching procedure to assign prices to purchases



Let G_t be the set of games added to the library of a user at the time t of scanning the library, let P_t be the set of corresponding prices and let S_τ be the set of sale prices at the time τ of scanning. Suppose an addition to the games library is observed at $t+1$: The sale price is assigned to a game only if $p_t \in S_\tau, S_{\tau+1}, S_{\tau+2}$, while the regular price is assigned to a game only if $p_t \notin S_\tau, S_{\tau+1}, S_{\tau+2}$. Since prices are constant in the short run, purchases and prices can be matched correctly.[37]

1.4.2 Matching: Reviews

The review data was collected from the individual steam game pages alongside the purchasing data from March to April 2017 and from June to July 2017 using web scraping techniques. Each review consists of the review text, a binary recommendation (**recommend** or **not recommend**), the date at which the review was written and a rating statistic that indicates how many other users have considered the review helpful. The matching between the purchasing and the review data is done via a unique identifier that links users' activities on the platform. Upon registration with the service, Steam users are assigned a unique 64 bit Steam ID that identifies all their activities in the store as well as the community pages. Reviews on Steam are not directly linked to this ID, but to a name that is chosen freely and can be changed voluntarily by the user. Crucially, this user name is linked to the unique ID via users' profile pages. By translating user names into Steam IDs for each review, I can link reviews to the tracked purchases.

1.4.3 Matching: Metascore Data

To account for the fact that review scores are not only affected by purchasing price but also product quality, I additionally collect data about average game scores. This data is gathered from Metacritic[38] which itself collects review scores for video games from professional reviewers such as video gaming magazines or blogs. The scores of the

[37] It should be noted that pricing errors almost never happen and permanent price adjustments are very infrequent, such that prices on the platform can be treated as constant over short periods.

[38] See metacritic.com.

different professional reviewers are then converted to a normalized score between 0 and 100 points and averaged to construct a representative professional rating of a game's quality. This "Metascore" is directly collected from each game's store page at the time of purchase[39] and serves as a measure of objective game quality at the time of purchase. I will discuss below to what extent this data reflects consumers perception of game quality at the time of purchase.

1.5 Descriptive Evidence

In this section, I describe the dataset and present first evidence that review scores vary with prices. In total, I observe 1,021,483 additions to users' games libraries. Table 1.1 shows summary statistics for the complete dataset of all key activations and purchases I record. Table 1.1 shows that there is substantial variation in prices (and in discount levels equivalently) across the recorded purchases with discount rates ranging all the way from no discounts to 97%. On average, 42% of purchases are recorded as bought on a sale. However, as discussed in Section 1.4, not all of these additions can be classified as purchases via Steam because additions to games libraries contain both direct purchases and key activations. Unfortunately, it is impossible to distinguish key activations from purchases for most games by only using the tracking data.[40] Hence, purchasing prices might be overstated, while discounts might be understated if users buy video games through other stores during sales or for lower prices in general.[41] This problem resolves

Table 1.1: Summary statistics: all key activations and purchases

Variable	Obs	Mean	Std. Dev.	Min	Max	P50
Review exists	1021483	.01	.11	0	1	0
Initial Price in Dollar	1021483	15.56	13.36	.5	99.99	12.99
Final Price in Dollar	1021483	11.59	12.05	.49	99.99	7.99
Discount in %	1021483	.26	.33	0	.97	0
Sale	1021483	.42	.49	0	1	0
Metascore (0-100)	450456	77.58	8.95	20	96	79
Game was played	1021483	.5	.5	0	1	0
Days to first played	506601	5.57	15.15	0	158.94	0

[39] This is important as Metascores can change substantially due to both game updates and inflow of new professional review scores, especially during the first weeks after the release of a game.

[40] Some games can only be purchased via Steam directly. This mostly applies to small budget games, where setting up other distribution channels would not pay off to the developer/publisher. For these games additions to games libraries could be classified as purchases.

[41] Since developers/publishers set prices on Steam, the Steam price is almost always the suggested retail price and therefore on average higher compared to the actual market price.

when matching purchases to reviews, as I observe in the review data whether the game was bought via Steam other through other channels. More precisely, the filtering feature of the review section enables me to single out reviews that have been written by users who have bought the game through Steam.[42] Table 1.2 shows summary statistics for the matched dataset that only contains purchases for which a review was written.

Table 1.2: Summary statistics: purchases and key activations for which a review was written

Variable	Obs	Mean	Std. Dev.	Min	Max	P50
Recommended	12030	.76	.43	0	1	1
Playtime until Review	12030	12.2	62.57	0	2656.15	2.83
Initial Price in Dollar	12030	22.37	17.04	.99	99.99	19.99
Final Price in Dollar	12030	18.68	16.52	.49	99.99	14.99
Discount in %	12030	.19	.28	0	.97	0
Sale	12030	.42	.49	0	1	0
Metascore (0-100)	4494	78.16	8.79	34	96	79
words (in 1,000)	12030	.74	1.13	0	22.61	.34
Pct. rated helpful	11827	.7	.28	0	1	.75
Game was played	12030	1	.03	0	1	1
Reviews written	12030	63.93	173.15	1	1493	22
Days to review	12030	7.85	16.33	0	119.33	.6

In total, I am able to match 12,030 reviews to tracked purchases, which implies a conversion rate from purchases to reviews of about 1%. Comparing the summary statistics between Table 1.1 and 1.2 provides insights into the types of purchases for which a review was written. Comparing the price levels between the complete and the matched dataset, it appears that reviews are predominantly written for more expensive games. This observation is amplified by the fact that prices in the complete dataset are likely overstated for the reasons discussed above. It also appears that reviews are not more likely to be written for games that have been bought on sale. One should note, however, that both the share of games bought on sale as well as the average discount might be understated in the complete dataset, as games bought on sale from other platforms or retailers are omitted. This observation is not a concern for the analysis, which only focusses on direct purchases from Steam, for which discounts are recorded correctly.

Moreover, Table 1.2 provides a number of interesting insights into reviewing behavior. The average review score across all games is at 76% and resembles approximately the

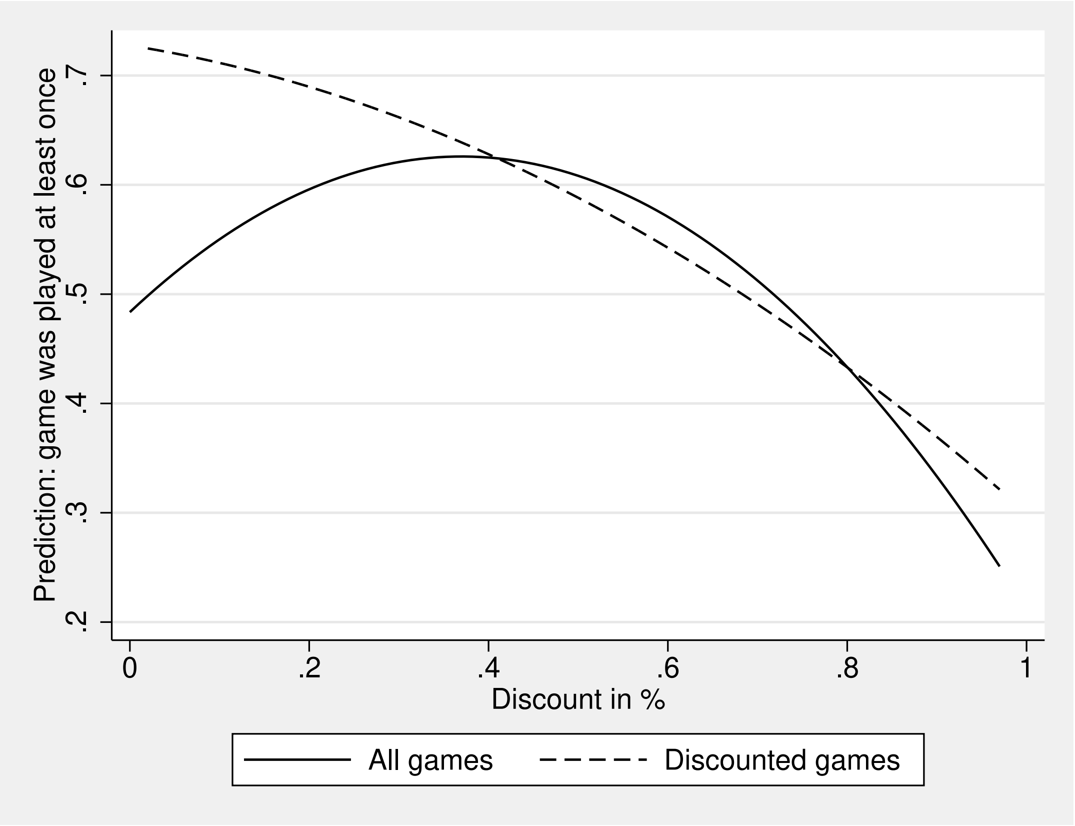
[42] This feature was originally added to notify consumers about reviews written by users that had obtained the game from other sources (potentially even free in exchange for a positive review).

average Metascore at 78.16 points. The playtime until a review was produced was 12.2 hours on average or 2.83 hours for the median reviewer. This discrepancy is due to the fact that the distribution of time until review is right-skewed. Nevertheless, a median playtime of approximately 3 hours should generally be sufficient for reviewers to make an informed decision. Similarly, the time to review is relatively short for the median reviewer. Overall, these statistics are consistent with a behavior that games are mostly played after purchase and reviewed shortly after if they are reviewed at all.

1.5.1 Playing Behavior

Before turning to the main part of the analysis, I will discuss briefly discuss the playing behavior of the tracked users with respect to newly obtained video games. Although there is no formal requirement that forces consumers to play games they want to review, in the dataset almost all games that have been reviewed have also been played. Hence, it seems important to understand which games actually get a chance to be played (and therefore reviewed) in the first place.

Figure 1.4: Estimated relationship between discount and product usage



Note: The underlying regression model is $played_i = discount_i + discount_i^2 + \varepsilon_i$.

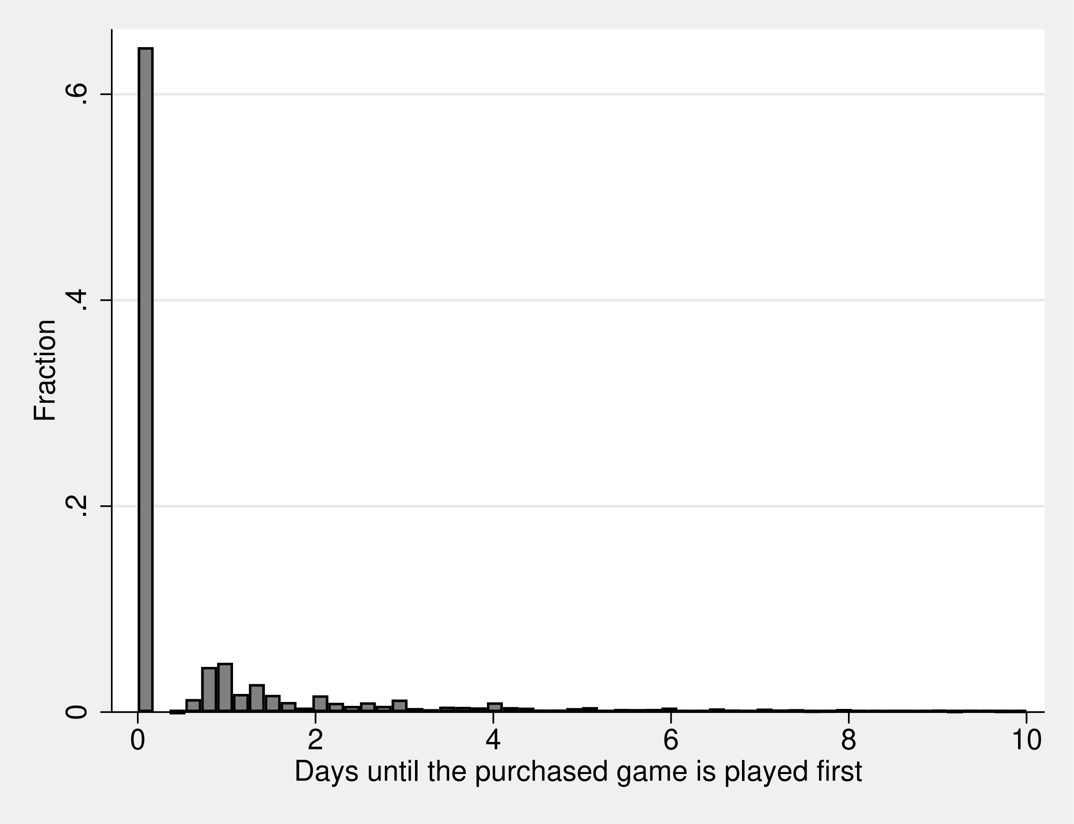
The observation that the average price of reviewed games is larger than the average price of all purchased games indicates that reviews are written predominately for games that have not been bought "too cheap" and supports the popular presumption that games are bought "on reserve". Indeed, about half of the purchases that I observe are games that have not yet been played.[43]

Figure 1.4 depicts the estimated quadratic relationship between discount and an indicator variable that is 1 when the game was played and 0 otherwise. The solid line shows the results of this estimation if all purchases are included while the model underlying the dashed line explicitly excludes all non-discounted games. The stark difference reinforces the presumption that discounts are understated in the complete dataset. In fact, whenever I observe a strictly positive discount for a purchase it should be relatively more likely that the game was bought via Steam compared to the situation when the observed discount is 0 for the following reason: Whenever a game is discounted on Steam, it is unlikely that it is available for a much lower price on other distribution channels at the same time. Generally, if games are on sale on multiple platforms simultaneously, publishers tend to offer the same discounts across distribution channels. In contrast, whenever a game is not discounted on Steam, it may very well be discounted on other (online and or retail) distribution channels. One could therefore presume that some of the purchases in the complete dataset that are recorded as non-discounted are actually discounted. If this were the case, the predicted probability of whether a non-discounted game was played would be downward biased. Nevertheless, both graphs show that the probability of playing a game decreases with its price. Consumers seem to "grab the chance" whenever they see a very cheap offer.

Interestingly, it appears that the probability of ever starting to play a game decreases drastically with time. Figure 1.5 shows the time that passes between the purchase and the first time the game is played for all games that have been played at least once. Most games are either started immediately right after the purchase or at least on the first or second day after the purchase.

[43] Here, "not played" refers to the game not being started at least once until the last time the games library of the user was scanned.

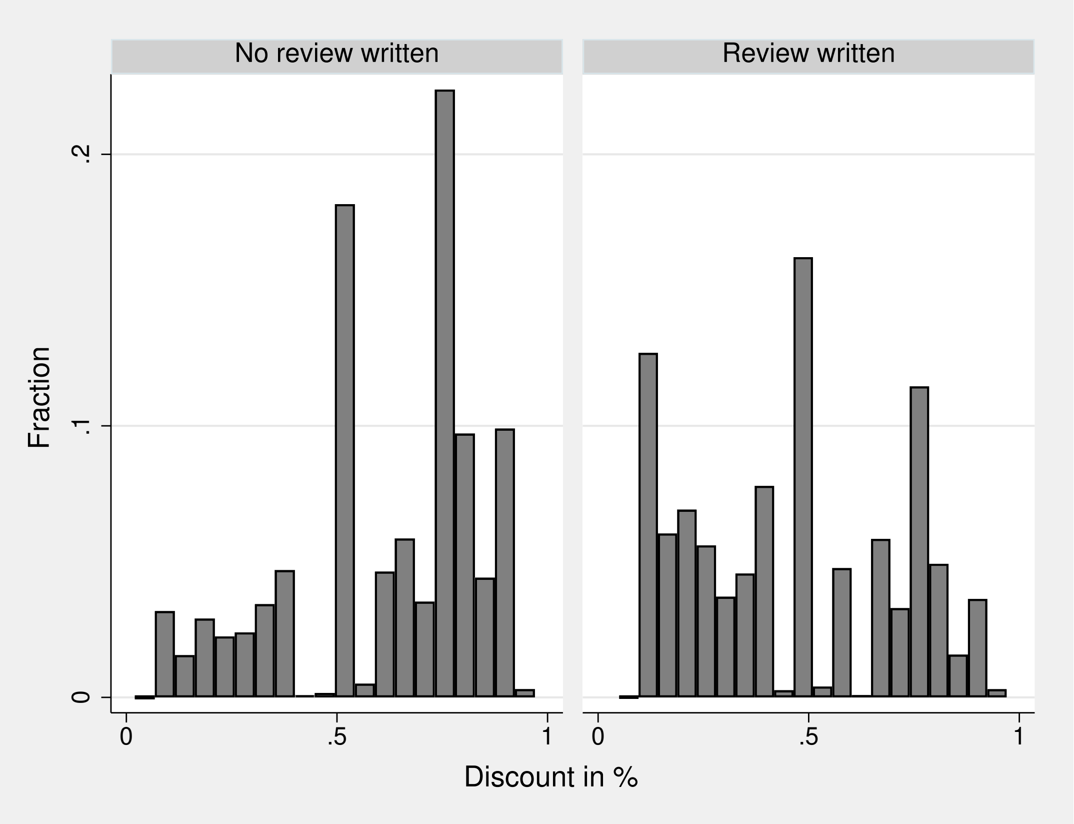
Figure 1.5: Distribution of days between purchase and first usage of a purchased game



Note: Histogram calculated for differences up to 10 days.

The observation that games which were purchased at a relatively high price get played with a higher probability translates to the distribution of discounts for reviewed and unreviewed games. Figure 1.6 depicts the distribution of discount levels for all games that were on sale during the tracking period separately for game purchases where a review if observed and where no review is observed. In comparison, the fraction of games purchased at a high discount is substantially higher when no review is observed, indicating that reviews are predominantly written for games acquired at low discount levels.

Figure 1.6: Distribution of discount level for purchases with and without review



1.6 Empirical Model

In this section, I outline the empirical model that is used for the estimation and discuss the empirical challenges, in particular endogeneity problems and ways to account for unobservable heterogeneity.

1.6.1 Identification

The general idea behind my identification strategy is to use variation in purchasing prices while controlling for game-level as well as reviewer-level effects to single out the effect of prices on review scores. The main empirical challenge lies in correcting for unobserved effects on the game and reviewer level. In particular, variation in prices only identifies the model if the price itself is not endogenous with respect to the error term.

There are two different channels through which prices could be correlated with unobserved effects: If game quality is not properly controlled for, it will be part of the error term. In case sellers tend to award higher discounts for low quality games, e.g.

to increase demand, an endogeneity issue arises, as prices will be correlated with game quality. I will discuss in detail how to correct for quality differences in Section 1.6.3. A similar problem arises if reviewers are heterogenous with respect to the price at which they buy a video game. If buyers who are ex-ante very positive towards a certain game buy at higher prices, buyer preferences as part of the error term will be correlated with prices and hence an endogeneity issue arises. I will discuss in detail how to deal with buyer-specific effects in Section 1.6.5.

1.6.2 Basic Model

In this section, I outline the basic naive estimation model. Let r_{ig} be the binary review score that is either 0 (not recommended) or 1 (recommended). Here, i denotes a specific review, while g denotes a particular game. The basic estimation equation can be written as

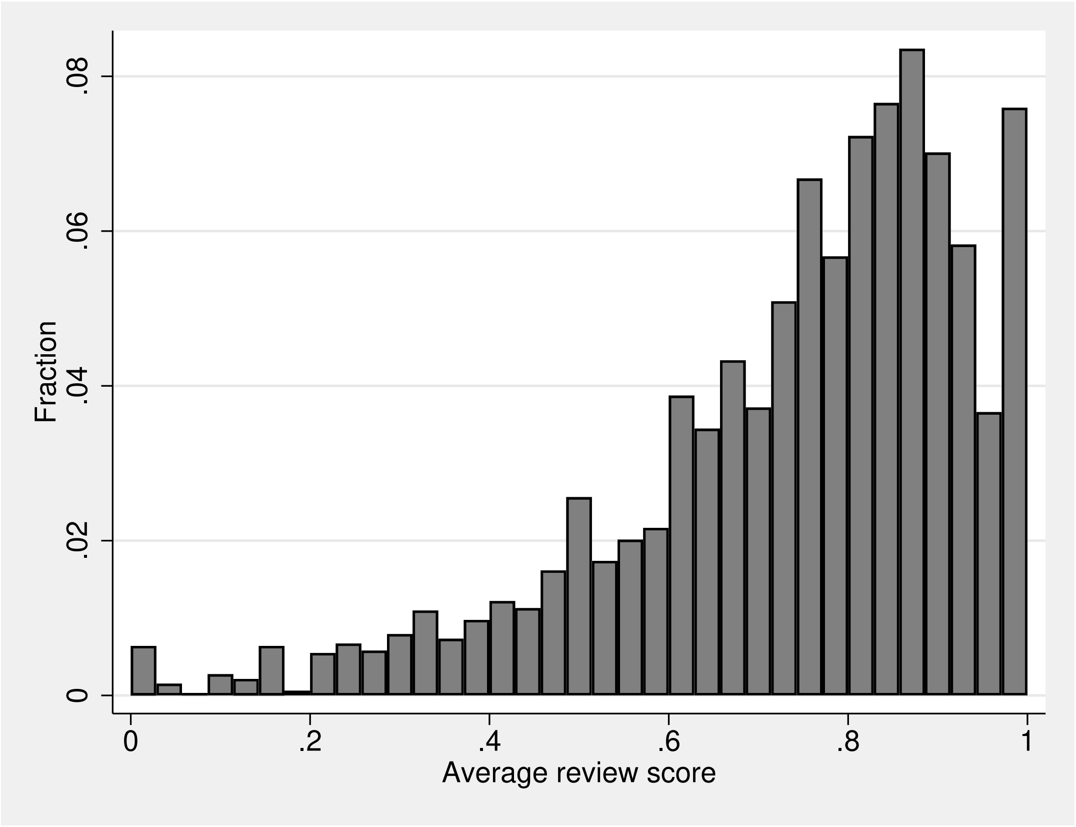
$$r_{ig} = \alpha + X_{ig}\beta + \gamma \; discount_{ig} + \varepsilon_{ig} \tag{1.1}$$

where X_{ig} is a set of controls and $discount_{ig}$ is the percentage discount that user i has payed for game g. The parameter γ is the effect of interest that captures the impact of the discount on the review score. Estimating equation (1.1) by standard OLS will yield consistent estimates only if the covariates are not correlated with the error term, which is highly unlikely in this setting. In particular, as discussed above, the discount on a game is very likely to be endogenous if game quality and reviewer specific effects are not properly controlled for. If ε_{ig} contains (parts of) the quality of a video game and discounts are on average higher for low quality video games, d_{ig} is correlated with ε_{ig} and an endogeneity issue arises. Similarly, discounts will be endogenous if ε_{ig} contains reviewer specific effects such as preferences for particular types of games which are correlated with the discount. In the subsequent sections, I propose a strategy to deal with these endogeneity problems that uses fixed-effects methods and proxy variables.

Another problem with the functional form in specification 1.1 is the binary nature of the dependent variable. While a linear-probability model could, in principle, fit the data well and allows for straightforward interpretation of the coefficients as marginal effects, issues may arise if the predicted probabilities are not within the boundaries. This is a serious concern in my setting as average ratings are very high for certain games and very low for others. Figure 1.7 shows the distribution of average ratings across games for all matched reviews. Indeed, almost 70% of games have average ratings very close to 1, while about 15% have average ratings very close to 0. To account for the dichotomous structure of the data, I additionally estimate a logistic regression model that ensures predicted probabilities within the boundaries of 0 and 1.

$$P(r_{ig} = 1) = \frac{e^{\alpha + X_{ig}\beta + \gamma\ discount_{ig}}}{1 + e^{\alpha + \beta X_{ig} + \gamma\ discount_{ig}}} \tag{1.2}$$

Figure 1.7: Distribution of average review scores across games



1.6.3 Controlling for Quality

For the reasons discussed above, controlling for game quality is important to isolate the effect of price changes on review scores. Figure 1.7 shows that the average review score differs drastically across games which gives a first indication that the range of (perceived) quality differences is indeed large across games. As outlined in Section 1.6.1 controlling for quality differences of games is crucial to identify the causal effect of discounts on review sores as discounts are very likely to be correlated with game quality. Controlling for game quality requires a suitable proxy for quality as it is perceived by reviewers. There are three potential strategies:

Average review score observed before purchase. The average review score pre purchase, i.e. the review score that a buyer sees before she buys the product, should best reflect a game's quality as it is perceived by reviewers. However, the average review score does not reflect the quality of a game net of its price but rather the average total utility reviewers have received, which may be affected by the price they pay and hence the average discount of a game over its lifecycle. Thus, the average review score pre-purchase is a function of the average discount of a game itself. A potential concern with this measure for quality could be that it will potentially bias the estimated effect of discounts on review scores, depending on the only partially observable pricing history of a game. Since I cannot observe prices for reviews written before the tracking period, I am unable to correct average review scores for prices payed or only use average review scores for discounted purchases.

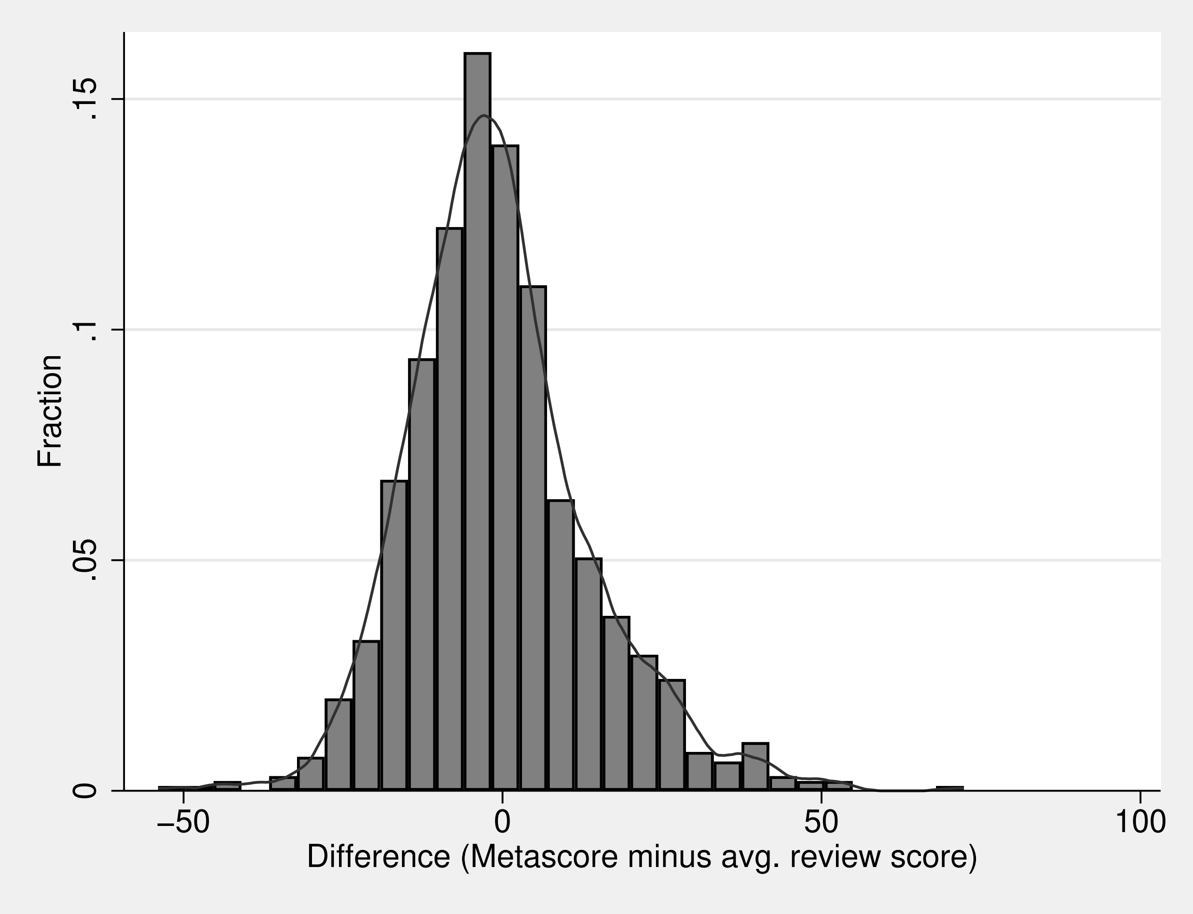
Average ratings by professional reviewers. An alternative to average review scores as a measure of game quality is the usage of professional ratings. This approach has the advantage that ratings are awarded based on the initial price of a game and therefore are independent of a game's pricing history. I use normalized, averaged professional ratings for video games gathered from Metacritic. The data is described in detail in Section 1.4.3. The drawback of using professional ratings to measure game quality as it is perceived by reviewers is that the views of both groups may not coincide. Professional reviewers may focus on different aspects than consumers when evaluating quality. Figure 1.8 shows the difference between the "Metascore" as an average rating score based on professional reviewers' opinions and the average review score[44] calculated from my dataset. Although the difference between both scores is centered around 0, the distribution appears to be slightly right-skewed. Moreover, the difference between both scores appears to be fairly large for some games, hinting that the difference may be systematic.

Table 1.3 shows the results of a regression of the score difference on game-level characteristics. The results confirm that professional review scores are indeed larger for higher priced games, indicating that, compared to professional reviewers, consumers might be more critical towards expensive "blockbuster" games but less critical towards cheaper

[44] To ensure comparability, the average review score is normalized to the interval $[0, 100]$. The Metascore ranges from 0 to 100 by default.

independently produced games.[45] This indicates that professional ratings, although they are most likely independent of the discount of a game, might be a poor predictor of game quality as it is perceived by non-professional reviewers.

Figure 1.8: Difference between "Metascore" and average review score



A second problem with professional ratings is that they are not available for all games in my dataset. As discussed in Section 1.4, the video game market has grown significantly in the recent years. Nowadays, a large fraction of newly released video games are low-budget, independently produced games, which are oftentimes only selectively rated by professional reviewers. Hence, for many low-priced games professional ratings do not exist – at least not in sufficient quantity. In contrast, almost all of the high-budget blockbuster games released by well-known publishers have sufficiently many professional ratings to construct an average rating score. This asymmetry leads to a potential selection effect if average professional review scores are used as a measure of quality. Table

[45] This is also supported by the fact that self-published games, which are oftentimes independently produced low-budget titles, are rated better by consumers than by professional reviewers.

Table 1.3: Regression results: difference between Metascore and avg. review score

	Metascore - avg. review score	
Initial Price in Dollar	0.369***	(0.044)
No. of developers	2.250*	(1.240)
No. of publishers	−1.424	(1.209)
Required age	0.212**	(0.083)
Achievement system	−7.630***	(1.127)
Game is self-published	−1.871**	(0.854)
Available on multiple platforms	1.078	(0.929)
Constant	−2.887	(2.096)
R-squared	0.15	
Observations	949	

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

1.4 reports the results of a Logit regression of an indicator variable, that is 1 whenever a professional rating score is available and 0 otherwise, on games characteristics.[46] The estimates confirm that higher priced games with more publishers, which is a typical feature of high-budget "blockbuster" projects, are more likely to be rated by professional reviewers. In contrast, self-published games – oftentimes low-budget and independently developed games – are less likely to be professionally reviewed. In my dataset, only 935 out of 3282 games for which reviews were recorded had received a professional rating score. In total, 4,409 out of 11,904 reviews were written for a game with a professional rating score. Hence, the results obtained using this measure of quality will only be valid for a particular subset consisting of higher priced "premium" games that were reviewed during the tracking period.

Fixed effects. A third approach to deal with quality differences between games is to control for game-specific unobserved heterogeneity using random or fixed effects estimators. While estimates obtained from random effects models are generally more efficient than estimates obtained via fixed effects due to the lower number of parameters that needs to be estimated, they may be biased if the unobserved effect is correlated with the covariates. As discussed above, I suspect that discounts are correlated with product quality as lower quality games may receive higher discounts in order to increase demand. For this reason, I will use a fixed rather than a random effects estimator in my baseline

[46] Only games for which at least one review was recorded were considered.

Table 1.4: Regression results: availability of Metascore measure

	Metascore available	
Initial Price in Dollar	0.054^{***}	(0.003)
No. of developers	0.280^{***}	(0.048)
No. of publishers	0.455^{***}	(0.105)
Required age	0.075^{***}	(0.007)
Achievement system	0.592^{***}	(0.054)
Game is self-published	-0.545^{***}	(0.049)
Available on multiple platforms	0.373^{***}	(0.049)
Constant	-3.226^{***}	(0.121)
R-squared	0.11	
Observations	14275	

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

model. As the fixed effects estimator uses only within-group variation to estimate the parameters, grouping on game-level should control for omitted game quality.[47]

A concern that applies to the fixed effects estimator and also to a slightly lesser extent to both measures based on average rating scores are potential quality changes over time. Although video games sometimes receive a multitude of patches and updates, especially at the start of their lifecycle, significant jumps in game quality are both costly and time intensive for game developers and publishers and therefore unlikely. Furthermore, the purpose of updates is mostly to resolve technical issues whereas content and feature extensions are oftentimes sold as add-ons and reviewed separately from the main game such that they should not affect the main game's ratings significantly.

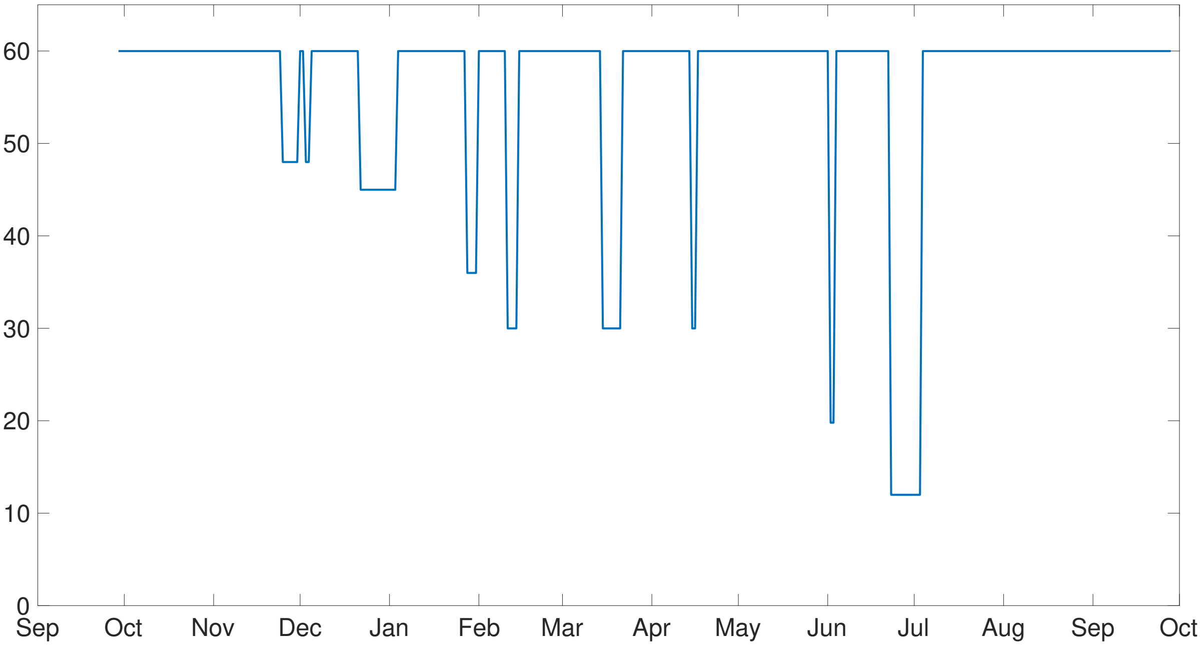
1.6.4 Controlling for Time Trends

Controlling for time trends even in my relatively short tracking dataset is crucial as not including a time trend in the model could cause sever omitted variable bias if both discounts and review scores evolve over time. Like many digital products, the average

[47] In principle, the correlation between the covariates and the random effect could be modeled out to restore the consistency of the random effects estimator. While this correlated random effects (CRE) approach offers the advantage of a potentially higher efficiency and the possibility to estimate the effect of group-level covariates, it also has some drawbacks. The most notable one is that the correlation between the random effect and the covariates still has to be modeled correctly. Consequently, the fixed effects approach still allows for more flexibility. As the focus of this chapter is not on estimating the effect of group-level variables, I think that the fixed effects approach is more appropriate. For a more detailed discussion on the CRE approach see e.g. Jeffrey M Wooldridge (2019), Jeffrey M Wooldridge (2010), Chamberlain (1982) or Mundlak (1978).

price of a video game decreases the longer it has been available on the market.[48] Since permanent price drops happen very rarely on Steam, the price decline of a game over its lifecycle is mainly driven by increasing discounts across sales. Figure 1.9 shows an example of a decreasing sales price path for the game "NBA 2k16" from its release in September 2015 to the release of its successor "NBA 2k17" in September 2016.[49] During the period displayed in Figure 1.9 and up until today, the base price of the game was never changed, but the sales priced decreased with almost every consecutive sale.

Figure 1.9: Prices for NBA 2k16 game from September 2015 to September 2016



As my dataset does not contain longer time series of pricing data for individual games, I can only verify the correlation between the discount level and the time since release cross-sectionally. Figure 1.10 shows the relationship between the time since release and average review scores as well as discounts during a sale for every purchase of an at most one year old game in my dataset.[50] While there is only a slight downward trend in review scores over time, the average discount of a game very clearly increases the longer it is on the market. I restrict attention to at most one year old games to illustrate the price path of games during their early lifecycle when they still receive a significant amount

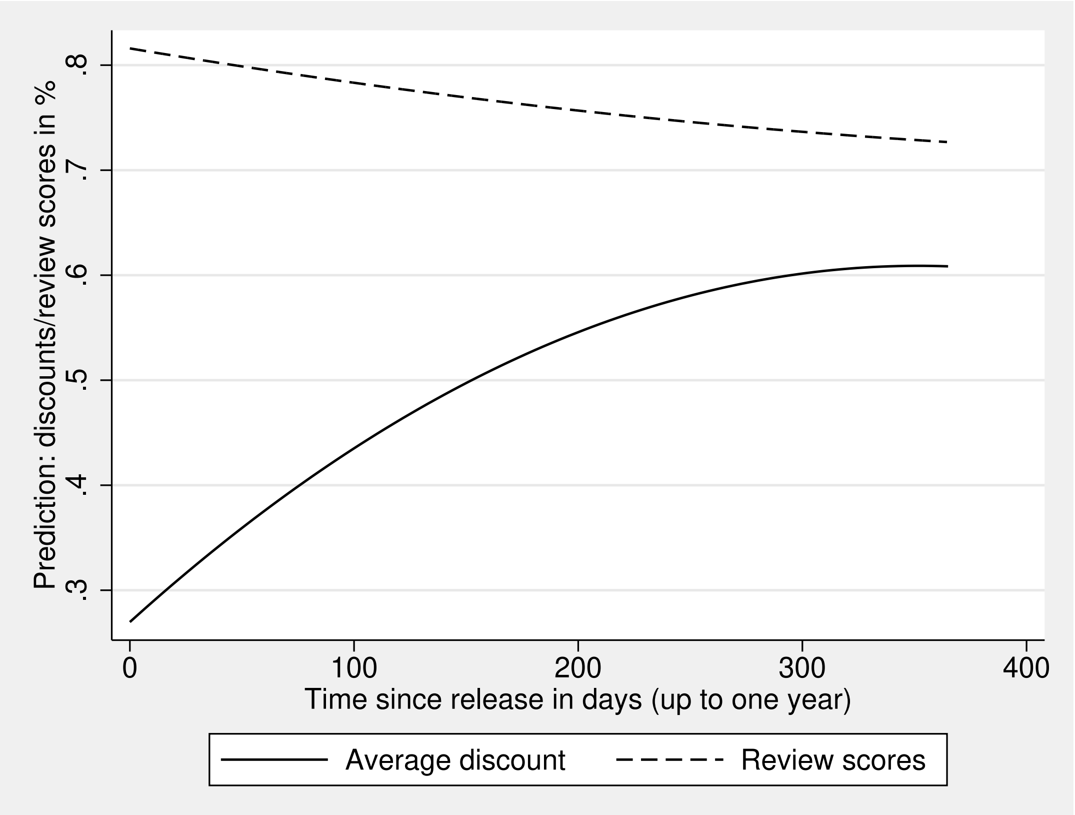
[48] This is mainly due to dynamic pricing considerations of publishers, as users have no incentive to buy multiple copies of a game.

[49] The graph is based on pricing data recorded by https://steamdb.info (accessed on 06/08/2017).

[50] The underlying regression models are $discount_i = time\ since\ release_i + time\ since\ release_i^2 + \varepsilon_i$ and $recommend_i = time\ since\ release_i + time\ since\ release_i^2 + \varepsilon_i$ respectively.

of attention. On Steam, most games hit a maximum discount level, at which they are not further reduced during subsequent sales, fairly quickly. Moreover, older games that are frequently re-released on Steam start with low discounts. This leads to an inversely U-shaped relationship between the discount level and the time since release if all games are considered. However, even if I consider all games, the correlation coefficient between the discount level and the time since release remains relatively high (0.4547). If at most one year old games are considered, the correlation coefficient is 0.5442. Hence, “time since release” should definitely be controlled for.

Figure 1.10: Relationship between “time since release” and “review scores” as well as “discount” (only sales prices)



Note: The underlying regression models are $discount_i = game\_age_i + game\_age_i^2 + \varepsilon_i$ and $recommend_i = game\_age_i + game\_age_i^2 + \eta_i$. Games considered are within the first year after their release.

Figure 1.10 provides clear evidence that review scores show a downward trend over time. This could be due to a negative selection effect caused by higher discounts but also other exogenous reasons such as broken promises about quality improvements or releases of superior games and therefore relatively lower review scores. If the latter effect is large, it could possibly even outweigh a positive selection effect. In any case, controlling for time trends seems indispensable to isolate the effect of discounts on review scores.

1.6.5 Controlling for Reviewer-Specific Heterogeneity

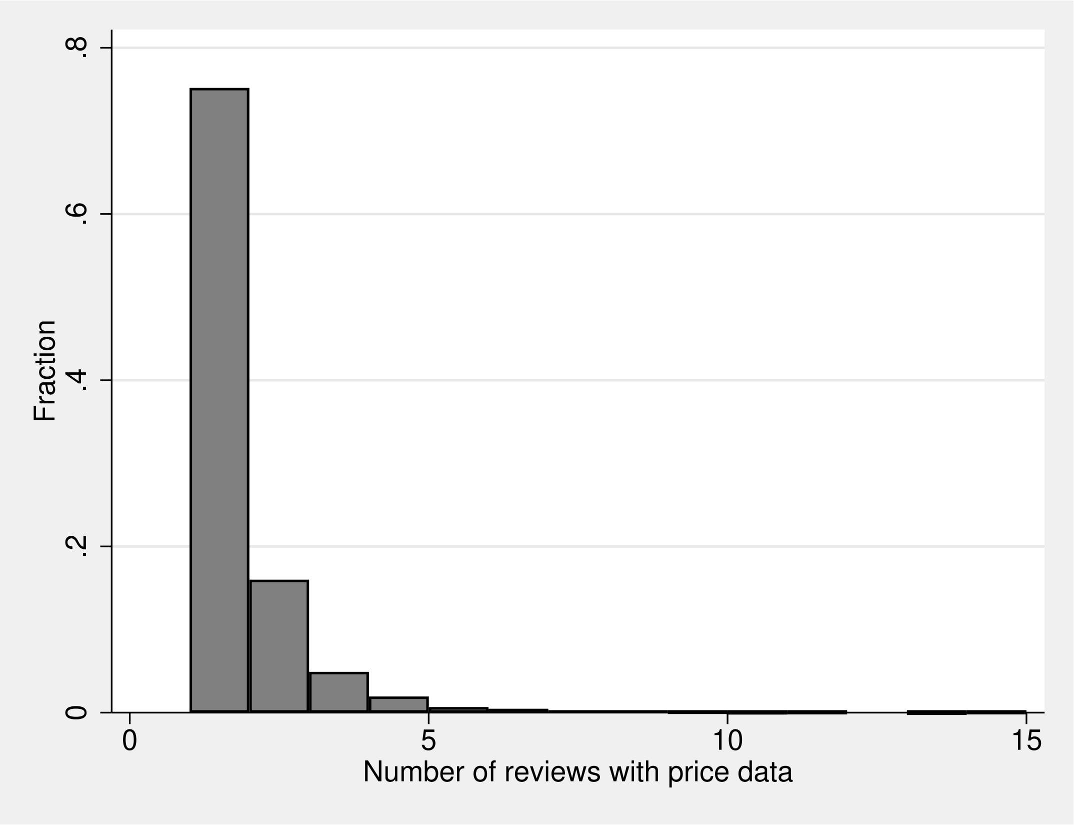
Another source of unobserved heterogeneity is the specific reviewing behavior of consumers. Even if the binary rating system in my setting is very simple, it is unlikely that all reviewers use the same standard when rating products. While some consumers may be more critical and hence more likely to award negative ratings even if they were somewhat satisfied with the product, others may tend to rate more positively, even if the product did not exactly meet their expectations. If reviewers' rating behavior is not properly controlled for, it will be part of the error term. This might cause omitted variable bias if rating behavior and discount are correlated. Imagine for instance a passionate video gamer who is very critical towards games, solely because she has very high standards with respect to game quality. At the same time, being a passionate gamer, she always buys the newest games immediately at a high price. On the contrary, imagine a very casual video gamer who buys games cheap during sales and at the same time rates very positively whenever he enjoys a game a little bit. Even in case all reviewers would rate in a similar way, they may differ in what they rate. Suppose that high priced games are always rated because consumers spent a significant share of their income on them. On the contrary, imagine sales purchases are only rated if they exceed expectation very much or fall short by a lot.

Hence, controlling for reviewer-specific rating behavior is essential to estimate the effect of discounts on review scores consistently. To achieve this, I construct a set of reviewer-specific controls to approximate rating behavior. In principle, it would be possible to use user-level fixed effects to control for reviewer-specific effects. This strategy requires, however, a sufficiently high number of observed reviews for every user in my dataset. Figure 1.11 shows the distribution of the number of reviews a user has written that can be matched with tracked price data.[51] About 75% of all reviewers had written only one review during the tracking period. To nevertheless construct proxy variables that control for rating behavior sufficiently well, I gather additional review data. For every game and reviewer in my dataset, I use web scraping techniques to download all reviews that were written on Steam during the two years prior to my tracking period. This allows me to expand the number of reviews I observe for each reviewer and con-

[51] For ease of graphical exposition I have excluded 14 reviewers that had written more than 15 reviews during the tracking period.

struct sound proxy variables for rating behavior that can be used for estimation on the tracking dataset. Figure 1.12 shows a quantile-quantile plot of the number of reviews with matched price data against the total number of reviews from June 2015 until June 2017.[52] Using this additional data, the average number of reviews that I observe per reviewer increases by 500%.

Figure 1.11: Number of reviews with price data



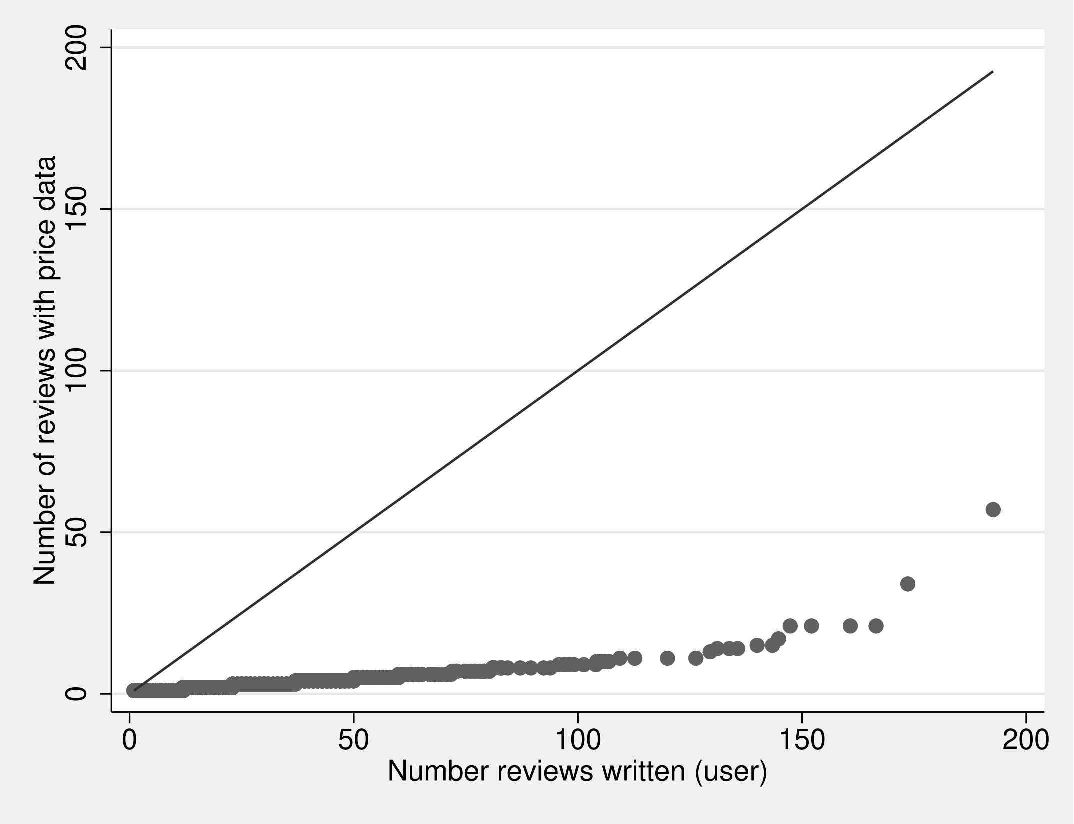
From the additional review data I construct three statistics that should describe the average rating behavior of each reviewer sufficiently well.

Average review score for other games. As discussed above, reviewers could have different degrees of "positivity" towards ratings scores. This measure is supposed to capture a reviewer's likelihood of giving a positive rating conditional on reviewing a game.

Average helpfulness rating of other reviews. Many modern review systems do not only offer the possibility to rate products but also reviews themselves. Steam offers

[52] For ease of graphical exposition I have excluded reviewers that had written more than a total of 200 reviews.

Figure 1.12: Number of reviews with price data: quantile-quantile plot



other platform users the opportunity to rate the quality of reviews as either positive or negative. Reviews that rate the product positively tend to receive more positive feedback themselves. If users who put a lot of effort into writing better reviews also tend to rate products more positively, review quality could help predicting review scores.[53] I therefore construct a “helpfulness” rating as a measure of review quality as perceived by other platform users as follows:

$$\text{helpfulness} = \frac{\text{positive ratings}}{\text{positive ratings} + \text{negative ratings}}.$$

As the aim of this metric is to account for the average quality of reviews produced by a certain reviewer, I use the average helpfulness rating over a reviewer’s other reviews.

Share of other reviewed games that were purchased on Steam. From the additional review data I can also observe the share of games that were bought on the

[53] An alternative but related explanation for this could be that the customers who were satisfied with the product are happy to put more effort into writing a good review than those who want to vent their anger. I find that in my dataset the correlation between the review score and the helpfulness rating is 0.3382.

Steam store.[54] On the one hand, one could presume that users who buy a larger share of their games elsewhere are more price sensitive, as prices on Steam are usually very high outside of sales. Hence, a user who buys a lot of games on other stores is most likely able to acquire them at a lower price. On the other hand, very price-sensitive consumer may only purchase games if they are on sale. As sale prices on the Steam store are usually very low, a high share of games bought on Steam may also indicate that consumers are very price-sensitive. If price-sensitivity is correlated with reviewing behavior, e.g. because price-sensitive consumers are more critical, the share of games purchased on Steam could be useful to predict review scores.

1.6.6 Extended Model

Before turning to the results, I will briefly outline the extended model that is used for estimation. As I have discussed in the last sections, correcting for unobserved game quality and reviewer-specific heterogeneity is crucial to get an unbiased estimate of the effect of discounts on review scores. In my baseline model, I will use a game-level fixed effects specification as well as specifications using average review scores pre-purchase and professional review scores to correct for unobserved game quality. Moreover, I will add a quadratic time trend to the model to allow review scores to evolve over time. As discussed above, not including a time trend in the model will lead to omitted variable bias and hence inaccurately estimated coefficients on the discount level if “time to purchase” is correlated with both review scores and discount levels. As only a small share of purchased and reviewed games is released during the tracking period, I additionally include an interaction term between time and the age of the game at the start of the tracking period. This is important as I expect time trends to be significantly different in the early lifecycle of a game.

The model in (1.1) thus becomes

$$r_{igt} = \alpha + X_{igt}\beta + \gamma_1 \; discount_{gt} + \gamma_2 \; discount_{gt}^2 + \delta_1 \; t + \delta_2 \; t^2 + \tau \; t \; age_g + \theta_g + \varepsilon_{igt} \tag{1.3}$$

where i indicates the reviewer, g indicates the game and t represents time measured in days since the first tracked purchase. Moreover, r_{igt} is the review score, X_{igt} is a set of reviewer-game-specific controls, $discount_{gt}$ is the percentage discount at time t, θ_g is a

[54] As discussed in Section 1.4, not necessarily all games that are registered on a user’s account need to be purchased via the store but could also be registered via a game key purchased elsewhere. From the reviews themselves I can observe is the reviewed game was purchased on Steam.

game-level fixed effect and age_g is the age of a game at the start of the tracking period. I also include a quadratic term of $discount_{gt}$. This specification allows that the direction of effect of discounts on review scores depends on the level of the discount itself. Similarly, the equivalent Logit specification becomes

$$P(r_{igt} = 1) = \frac{e^{\alpha + X_{igt}\beta + \gamma_1\ discount_{gt} + \gamma_2\ discount^2_{gt} + \delta_1\ t + \delta_2\ t^2 + \tau\ t\ age_g + \theta_g}}{1 + e^{\alpha + X_{igt}\beta + \gamma_1\ discount_{gt} + \gamma_2\ discount^2_{gt} + \delta_1\ t + \delta_2\ t^2 + \tau\ t\ age_g + \theta_g}}. \tag{1.4}$$

In specifications with explicit proxies for game quality, I replace θ_g with ηq_{gt}, where q_{gt} is the expected quality of game g at day t. Additionally, I include a set of game-level controls in all specifications that do not include a fixed effect.

1.7 Results

Before turning to the specification described in (1.3), I will show results for a baseline specification that does not account for quality differences across games. Table 1.5 shows regression results for such a model. Additionally, specification (1) in Table 1.5 presents results for the naïve basic model described in (1.1) without a quadratic term for discounts and a time trend. The estimated effect of discounts on review scores is clearly negative in this model. However, including a time trend in the model shows that the effect is likely overstated.

At first sight the magnitude and direction of the effect is very close to the results obtained by Zegners (2019) and points towards a small but negative effect of higher discounts on review scores. Ceteris paribus, a 50% discount would lead to a drop in the probability of awarding a positive review score by 3 percentage points. Still, specifications (1) and (2) do not account for potential non-linearities in the relationship between discounts and review scores. Including a quadratic term of the discount level indeed suggests that small discounts could lead to higher review scores, which is in line with the findings by Carnehl et al. (2021a) and Luca and Reshef (2021), while high discounts seem to lead to lower review scores, which is in line with the findings by Zegners (2019). I will discuss these findings below once I have established that they hold true when controlling for quality differences among games. The estimates obtained using the Logit model instead of OLS show a similar pattern in terms of significance levels.[55]

[55] In fact, the average marginal effect obtained from the Logit model is very similar to OLS parameter estimates.

Table 1.5: Regression results: basic model without game quality controls

	OLS (1)	OLS (2)	OLS (3)	Logit (1)	Logit (2)
Discount in %	−0.104***	−0.067**	0.218**	−0.508**	1.362**
	(0.033)	(0.031)	(0.090)	(0.213)	(0.628)
Discount in % × Discount in %			−0.416***		−2.661***
			(0.119)		(0.784)
Initial Price in Dollar	−0.001	−0.000	−0.000	−0.002	−0.001
	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)
Rel. Playtime	0.012***	0.011***	0.011***	0.099***	0.102***
	(0.004)	(0.004)	(0.004)	(0.037)	(0.038)
Days to first played	−0.001	−0.002	−0.002	−0.010	−0.010
	(0.001)	(0.001)	(0.001)	(0.006)	(0.006)
Share purchases on Steam	−0.046**	−0.048**	−0.050**	−0.332**	−0.348**
	(0.021)	(0.021)	(0.020)	(0.147)	(0.148)
Avg. helpfulness rating	0.020	0.023	0.021	0.151	0.143
	(0.034)	(0.035)	(0.035)	(0.253)	(0.255)
Total number ratings for reviews	0.000**	0.000	0.000	0.002*	0.002*
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Avg. ratings other games (user)	0.458***	0.449***	0.448***	2.750***	2.749***
	(0.026)	(0.026)	(0.026)	(0.162)	(0.163)
Number reviews written (user)	−0.000*	−0.000**	−0.000*	−0.001***	−0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age of account (years)	−0.002	−0.002*	−0.003*	−0.021**	−0.022**
	(0.002)	(0.001)	(0.002)	(0.011)	(0.011)
Game is self-published	0.011	0.021	0.018	0.143	0.122
	(0.017)	(0.016)	(0.016)	(0.110)	(0.109)
Available on multiple platforms	0.024	0.023	0.024	0.156	0.170
	(0.016)	(0.017)	(0.017)	(0.120)	(0.120)
No. of publishers	−0.052	−0.072*	−0.075*	−0.435**	−0.456**
	(0.038)	(0.039)	(0.038)	(0.216)	(0.208)
No. of developers	−0.008	−0.012	−0.012	−0.079	−0.079
	(0.027)	(0.027)	(0.028)	(0.177)	(0.178)
Achievement system	0.032**	0.029*	0.026	0.194*	0.172
	(0.016)	(0.016)	(0.016)	(0.113)	(0.110)
Days		0.003***	0.003***	0.020***	0.020***
		(0.001)	(0.001)	(0.006)	(0.006)
Days × Days		−0.000***	−0.000***	−0.000***	−0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Days × Age of game		0.000	0.000**	0.000	0.000**
		(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.558***	0.535***	0.531***	−0.049	−0.072
	(0.068)	(0.071)	(0.070)	(0.469)	(0.455)
adj. R^2	0.10	0.11	0.11		
pseudo R^2				0.11	0.11
Observations	5558	5526	5526	5526	5526
Number of clusters	1625	1616	1616	1616	1616

Standard errors in parentheses (clustered on the game level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.6: Regression results: fixed effects model

	FE (1)	FE (2)	FE (3)	FE Logit (1)	FE Logit (2)
Discount in %	−0.013	0.123*	0.393**	0.757*	3.265**
	(0.053)	(0.067)	(0.159)	(0.434)	(1.275)
Discount in % × Discount in %			−0.458*		−4.136**
			(0.236)		(1.981)
Initial Price in Dollar	0.004***	0.003	0.003*	0.029	0.031
	(0.001)	(0.002)	(0.002)	(0.023)	(0.023)
Rel. Playtime	0.012***	0.009**	0.009**	0.094**	0.093**
	(0.004)	(0.004)	(0.004)	(0.039)	(0.039)
Days to first played	−0.001	−0.002**	−0.002**	−0.021**	−0.021**
	(0.001)	(0.001)	(0.001)	(0.010)	(0.010)
Share purchases on Steam	−0.074***	−0.080***	−0.079***	−0.607***	−0.602***
	(0.022)	(0.022)	(0.022)	(0.205)	(0.205)
Avg. helpfulness rating	−0.007	0.008	0.005	0.143	0.094
	(0.040)	(0.041)	(0.041)	(0.311)	(0.312)
Total number ratings for reviews	0.000**	0.000**	0.000**	0.004**	0.004**
	(0.000)	(0.000)	(0.000)	(0.002)	(0.002)
Avg. ratings other games (user)	0.408***	0.392***	0.392***	2.695***	2.708***
	(0.029)	(0.028)	(0.028)	(0.192)	(0.192)
Number reviews written (user)	−0.000	−0.000	−0.000	−0.001	−0.001
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Age of account (years)	−0.003*	−0.003*	−0.003*	−0.025*	−0.024
	(0.002)	(0.002)	(0.002)	(0.015)	(0.015)
Days		0.003***	0.003***	0.012*	0.012*
		(0.001)	(0.001)	(0.006)	(0.006)
Days × Days		−0.000***	−0.000***	−0.000***	−0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Days × Age of game		0.000*	0.000**	0.000**	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.491***	0.481***	0.478***		
	(0.043)	(0.053)	(0.049)		
adj. R^2	0.08	0.10	0.10		
pseudo R^2				0.13	0.13
Observations	5558	5526	5526	3359	3359
Number of clusters	1625	1616	1616		
Number of groups	1625	1616	1616	327	327
Min. Obs. per group	1	1	1	2	2
Avg. Obs. per group	3	3	3	10	10
Max. Obs. per group	369.00	369.00	369.00	369.00	369.00
corr(θ, covariates)	−0.09	−0.11	−0.09		

Standard errors in parentheses (clustered on the game level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As I have already discussed above, the estimates in Table 1.5 are obtained from a model that neglects quality differences among games which could substantially change the results. In the last section, I have argued that including game-level fixed effects should account for quality differences sufficiently well. Table 1.6 shows estimation results for the model specifications (1.3) and (1.4) with game-level fixed effects. Comparing the results from the fixed-effects models to the results obtained via OLS, it appears that the parameter estimates for the discount level are understated in the latter. Compared to OLS, when controlling for quality differences, the effect of the discount level on review scores is larger in all specifications. This is consistent with the rather intuitive presumption that low-quality games should receive higher discounts on average. If low quality not only leads to higher discounts but also lower review scores, a significant fraction of the negative effect of quality on review scores is falsely attributed to the discount level if quality differences are not properly controlled for.

Apart from the quantitative differences, the fixed-effects estimates confirm the presumption that the relationship between discount levels and review scores is non-linear. At low to medium discount levels, review scores increase with the size of the rebate until they reach a maximum and start to decrease again towards their initial level. In the linear specification, this peak level can be computed as $-\frac{1}{2}\frac{\gamma_1}{\gamma_2}$. Based on the estimates from Table 1.6 specification (3) the review score optimizing discount level would be reached at around 43%, which is just below 50% – a discount level that most games reach within the first year of their lifecycle. This effect is economically sizable as well. At a discount level of 43%, the probability of receiving a positive review increases by about 8.45 percentage points. Finally, for very high discount levels, discounts can potentially even have a negative effect on review scores. The "break-even" discount level in terms of review scores is reached at $-\frac{\gamma_1}{\gamma_2}$ in the linear model or about 86% based on the estimates from Table 1.6 specification (3). Interestingly, at a hypothetical discount of 100%, i.e. if games had been given away for free, review scores would decrease by 0.065 on average which is very close to the estimates of Zegners (2019) obtained in the free ebooks case.[56] The main implications of the results are threefold:

1. The effect of discount levels on review scores appears to be positive for low to medium discount levels. This indicates that the selection effect is either positive or at least outweighed by the price effect for all moderate discount levels. Hence, this constitutes a potentially reinforcing effect of discount levels on demand: If both lower prices and higher review scores have an increasing effect on demand, the second order effect of discounts on review scores may substantially affect firms' dynamic pricing considerations.

2. There seems to be a "review score optimizing discount level" beyond which the selection effect causes review scores to decrease again in the discount level. This points towards a potential trade-off for firms' dynamic pricing considerations. For

[56] This prediction should, however, be taken with care as a discount level of 100% is "out of sample".

high discount levels the positive effect of prices on demand may be partially outweighed by a second order effect that leads to lower review scores compared to medium discount levels. Consequently, at high discounts prices might be "too low" if the feedback effect of discounts on review scores is not internalized.

3. Only very high discount levels may lead to lower review scores on average. This indicates that a strong enough negative selection effect occurs only for games that are "almost given away for free", which casts doubt on the presumption that free giveaways serve as a way to increase reputation.

A number of other interesting results are worth mentioning. As one would expect, the best predictor of individual review scores is the overall rating behavior of the reviewer. The more positive individual reviewers tend to rate other games, the higher is the unconditional likelihood that they award a positive rating in general. Being a good reviewer in the sense that one's other reviews are on average rated very positively by other users does not seem to have an effect on the probability of giving a positive review score. Crucially, being a more positive reviewer is strongly correlated with receiving good ratings for reviews but does not directly impact the probability of writing a positive review. Similarly, the number of reviews written does not significantly influence the probability of giving a good rating. Older accounts are, however, slightly more critical even though the effect is minimal at a 0.3 percentage points lower probability of giving a positive rating per year. Interestingly, reviewers who prefer to purchase games on Steam directly are significantly more likely to give a bad rating. When it comes to playing behavior, games that are played more extensively relative to the game's average playtime and games that are started sooner after purchase are more likely to receive higher scores.

Even though the fixed-effects approach should lead to consistent estimates, I have discussed professional review scores and average review scores from other users as alternative approaches to control for unobserved game quality. Table 1.7 shows estimation results for these specifications. Qualitatively, the effect of discount levels on review scores is very similar to the fixed-effects results in Table 1.6 specification (3). Figure 1.13 shows the predicted effect of discounts on review scores for all four models. As I have discussed in the previous section, quantitative differences can be explained by the construction and the specific properties of these quality proxies. More precisely, average review scores from other users may be functions of unobserved discounts themselves. If higher discounts lead to higher review scores, which is strongly suggested by the fixed-effects results for a wide range of discounts, the effect of discount levels on review scores is partially absorbed by average review scores.

Table 1.7: Regression results: model with proxy variables for game quality controls

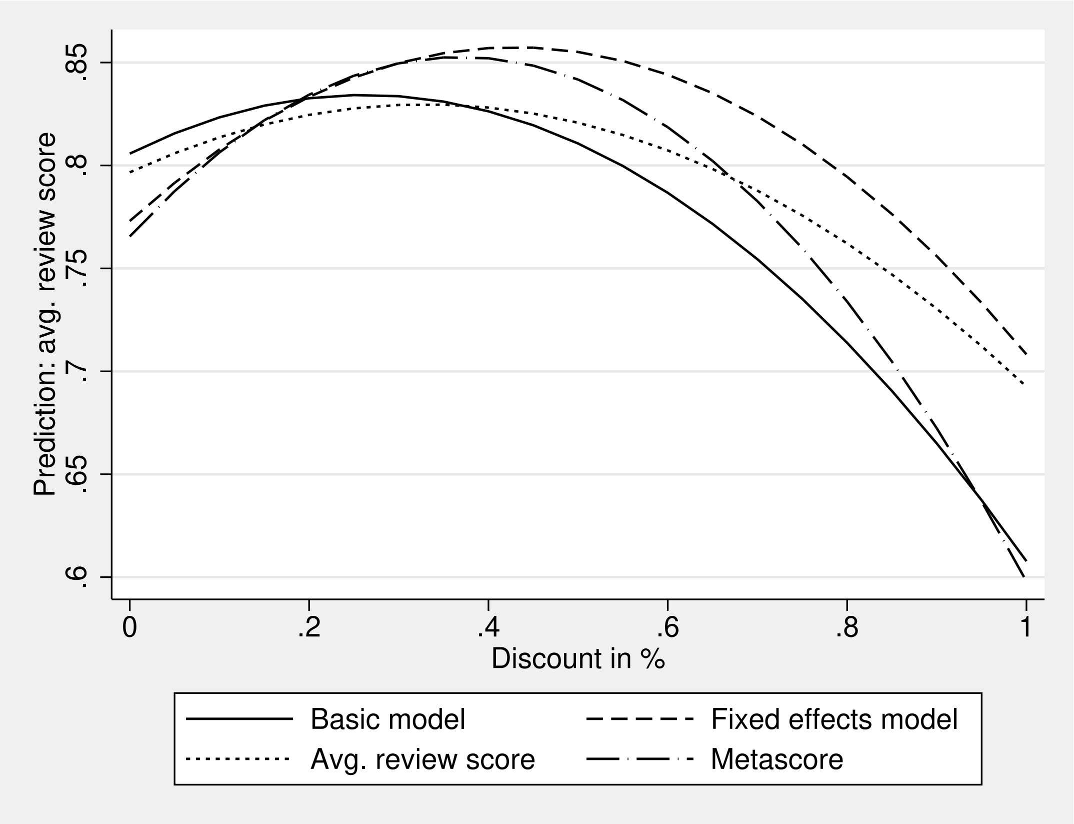
	Avg. Score	Metascore	Avg. Score Logit	Metascore Logit
Discount in %	0.200***	0.472***	1.545***	3.256***
	(0.072)	(0.153)	(0.545)	(1.099)
Discount in % × Discount in %	−0.305***	−0.640***	−2.330***	−4.453***
	(0.100)	(0.205)	(0.712)	(1.420)
Average review score	0.704***		4.574***	
	(0.040)		(0.274)	
Metascore (0-100)		0.004**		0.024**
		(0.002)		(0.011)
Initial Price in Dollar	0.001***	−0.002*	0.006**	−0.010*
	(0.000)	(0.001)	(0.003)	(0.006)
Rel. Playtime	0.011***	0.017**	0.118***	0.162**
	(0.004)	(0.007)	(0.042)	(0.071)
Days to first played	−0.002**	−0.003**	−0.013**	−0.019**
	(0.001)	(0.001)	(0.006)	(0.009)
Share purchases on Steam	−0.061***	−0.040	−0.449***	−0.342
	(0.020)	(0.036)	(0.156)	(0.264)
Avg. helpfulness rating	0.006	0.092	0.082	0.721
	(0.035)	(0.067)	(0.277)	(0.464)
Total number ratings for reviews	0.000	−0.000	0.002*	−0.000
	(0.000)	(0.000)	(0.001)	(0.002)
Avg. ratings other games (user)	0.426***	0.393***	2.838***	2.412***
	(0.024)	(0.055)	(0.176)	(0.318)
Number reviews written (user)	−0.000	0.000**	−0.001*	0.001
	(0.000)	(0.000)	(0.000)	(0.001)
Age of account (years)	−0.004***	−0.006**	−0.036***	−0.043**
	(0.001)	(0.003)	(0.011)	(0.019)
Game is self-published	0.009	0.056**	0.098	0.359**
	(0.011)	(0.026)	(0.082)	(0.179)
Available on multiple platforms	−0.001	−0.011	−0.017	−0.060
	(0.012)	(0.028)	(0.098)	(0.193)
No. of publishers	−0.056***	−0.118***	−0.359***	−0.685***
	(0.018)	(0.033)	(0.108)	(0.171)
No. of developers	0.004	0.020	0.097	0.108
	(0.013)	(0.044)	(0.104)	(0.281)
Achievement system	−0.009	0.081	−0.046	0.531*
	(0.012)	(0.052)	(0.092)	(0.321)
Days	0.003***	0.002	0.015***	0.008
	(0.001)	(0.002)	(0.004)	(0.009)
Days × Days	−0.000***	−0.000**	−0.000***	−0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
Days × Age of game	0.000	0.000*	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.016	0.260	−3.490***	−1.765*
	(0.048)	(0.162)	(0.350)	(0.994)
adj. R^2	0.18	0.14		
pseudo R^2			0.18	0.15
Observations	5516	1593	5516	1593
Number of clusters	1606	415	1606	415

Standard errors in parentheses (clustered on the game level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Using professional review scores (Metascore) as a proxy for game quality avoids this problem, as they should be based on the initial price of the product, if anything. In fact, for low to moderate discounts the estimates obtained when using this measure are quantitatively very similar to those obtained from the fixed-effects model. Nevertheless, there are two major problems with this measure, which are also visible in the results: Firstly, as I have discussed in the previous section, a considerable amount of games, especially cheaper and independently produced games, do not receive professional reviews. This does not only mean that I lose about 70% of my sample but also that the results are only applicable to a specific subset of the products, namely high-budget projects that naturally receive lots of attention. Secondly, professional review scores appear to be a somewhat imperfect measure of quality as it is perceived by consumers. In fact, when comparing the predictive power of average review scores by users and by professional reviewers, using average review scores by users as a quality proxy yields a better fit than using professional review scores.

Figure 1.13: Prediction of average review scores based on discount level



In general, average review scores seem to be a way more powerful predictor of individual review scores, to the point that they almost function as a game-specific constant in the model, while the Metascore is only able to explain some variation in individual review scores. Interestingly, when using professional review score as a quality proxy, the coefficient on initial price becomes negative, indicating that higher-priced games are rated significantly worse by consumer reviewers relative to professional reviewers.

An advantage of using quality proxies is that they allow for analyzing the impact of game-level characteristics on review scores. However, out of those only the number of publishers seems to (negatively) affect review scores. In principle, this could also be achieved by using a random effects model for estimation, which comes with the additional benefit of more efficient estimation. Using a random effects model is, however, only justifiable if the individual effect is not correlated with the covariates. Both, the fact that the correlation is estimated at around -0.1 in Table 1.6 specification (3) and a Hausman specification test confirm that this is not the case in my sample.

1.8 Conclusion

In this chapter, I have provided empirical evidence on the effect of prices on online review scores for a large online video game store using a unique dataset created by web tracking techniques that matches individual level purchasing data to consumer reviews. I have shown that there is robust evidence of a relationship between prices and review scores and that the direction of the effect may depend on the price level itself. This result is consistent with the presence of two countervailing effects. A (direct) price effect that encourages consumers to write better reviews based on a value-per-money consideration and a selection effect that captures the change in the population of buying consumers, who may show more (less) taste for the product with price increases (decreases).

Although the previous literature has extensively analyzed the relationship between review scores and demand, a potential feedback effect between prices and review scores has only recently received attention. To gain a thorough understanding of the relationship between review scores and demand it is however necessary to fully comprehend the strategic motives of sellers and the means through which they aim to increase review scores. If quality improvements are costly or difficult to implement and direct score manipulation is ineffective or impractical, lowering prices appears to be a convenient and easily-implementable strategy for sellers to induce more favorable ratings.

I build a unique tracking dataset that links individual level purchasing prices to consumer reviews on a large online platform that sells video games. Observing both prices as well as review scores allows me to directly estimate the effect of discounts on review scores. I find that for low to moderate discount levels, higher discounts increase the probability of receiving a positive review score. This indicates that the selection effect is either positive or at least outweighed by the price effect. On average, consumers rate more favorably if they pay a lower price for the product. For high discount levels, the selection effect becomes smaller until it becomes negative for very low prices. At very low prices it seems that even consumers with very low valuations “give it a try” and end up rating the product badly as they don’t enjoy it. However, most notably and contrary to free giveaways, for a very wide range of discount levels, lowering prices will result in higher average review scores.

The non-linear relationship between prices and review scores also indicates that there exists a “reputation-maximizing discount level”. More precisely, although for high discount levels the effect is still positive relative to the initial price, I estimate the discount level that maximizes the probability of obtaining a positive review score to be around 43%. When lowering prices even further, sellers face an additional trade-off between higher demand but lower review scores, which in turn may again affect demand.

These results may also help reconciling recent and seemingly contradictory findings by Zegners (2019) as well as Carnehl et al. (2021a) and Luca and Reshef (2021). While Zegners (2019) finds a positive relationship between prices and ratings, both Carnehl et al. (2021a) and Luca and Reshef (2021) find the opposite result. A notable difference is that Zegners (2019) looks at the effect of a 100% discount while both Carnehl et al. (2021a) and Luca and Reshef (2021) analyze continuous price changes at a positive price level. The dataset I analyze allows me to estimate the effect of prices on review scores for a broad range of discount levels. My results are qualitatively in line with both, seemingly contradictory findings. For small discount levels, i.e. small price changes at a positive price level, I find a negative relationship between prices and review scores which is consistent with the result in Carnehl et al. (2021a) and Luca and Reshef (2021). For large discounts, I find that the relationship between prices and review scores is positive, which is qualitatively consistent with the result obtained by Zegners (2019).

I think that this chapter takes an interesting further step towards analyzing the relationship between prices and review scores. However, there is ample scope for further work. Whereas I have focused on the demand side and argued that price changes are exogenous from the viewpoint of a consumer, sellers certainly should factor them into their profit maximization. My results indicate that a seller’s dynamic pricing problem not only involves trading off price and demand when review systems are an important information transmission or signaling device for consumers. If the second-order effect of prices on review scores feeds back to future demand, this should be taken into account by profit-maximizing sellers. Martin and Shelegia (2021) and Carnehl et al. (2021b) have taken first theoretical steps into this direction and Carnehl et al. (2021a) have provided evidence that entrants may strategically set lower prices to obtain better ratings and charge higher prices in future periods, but more empirical evidence remains yet to be explored. The necessary additional data in form of demand and pricing data is unfortunately not readily available for most platforms. Utilizing advanced tracking techniques, future research could certainly benefit from building a larger database that enables empirical analyses of supply side behavior. Finally, it seems worthwhile to continue exploring the relationship between review scores and prices in other settings. Although my findings are consistent with recent contributions, it remains to be explored if and how they translate to other sectors. Most notably, the finding that the direction of the effect may depend on the level of the price or discount is, to the best of my knowledge, novel. It remains to be seen if this finding can be confirmed with other settings and datasets.

1.9 Appendix

In this Appendix, I outline the details of the tracking algorithm that was used to collect the data

1.9.1 Selection of Users

Due to technical limitations only a subset of users could be tracked to ensure that purchases were identified in a reasonable time. The Steam web API that was used to record purchases allows 100,000 calls per day. Furthermore, it is recommended to keep a one second interval between calls to avoid being blocked by the server. For these reasons, I had to restrict tracking to a representative set of user. Moreover, in order to observe enough reviews that could be attributed to purchases within a reasonable time frame, I had to assure that the selected users were likely to write reviews. The unconditional probability of a random user ever writing a review lies around 0.1%. For this reason, I restrict attention to users that have shown enough interest in writing reviews in the past. This set of active reviewer is constructed in the following way:

1. In March 2017 all reviews written during the last year for every game were collected from the Steam website.
2. For all reviews the unique Steam ID of the reviewer is collected from the Steam user page.
3. For all Steam IDs the number of reviews written is computed.
4. For all Steam IDs with at least two reviews per year, the individual profile settings are collected using the Steam web API.
5. Only reviewers who have set their profile status to "public" and who have inidicated that they live in the U.S. are part of the group that is tracked.

1.9.2 Tracking Details

A set of 50,424 active reviewers was tracked daily from March 2017 to July 2017. The tracking script was written in Python 3 and was run daily as far as possible on a linux-based Amazon EC2 server. The data was stored in a PostgreSQL database on an Amazon RDS server. Users are identified by 64-bit Steam IDs that are unique and tied to a specific account. The detailed tracking algorithm for is described below:

1. For each user in the set of tracked users:
   (a) It is checked whether the user's profile is still publicly available.

   (b) If the profile is not public anymore, the account is marked as private and the process continues with the next Steam ID.

   (c) If the profile is public, the user's games library is downloaded.

   (d) If the profile was private during the last run, the games library is treated as an initial state and no changes are identified.

   (e) If the profile was public during the last run, the new games library is compared to the one recorded during the last run and changes are identified.

   (f) Removed games are treated as refunds and are deleted from the user's games library.

   (g) For each newly added game, the current initial price and discount of the game are automatically collected from the store page of the game.

   (h) For each game in the user's games library the total playtime of that game at the time is collected.

   (i) The changes are saved in the database and the process continues with the next Steam ID.

2. When the last user was scanned, the process starts over and continues with the first Steam ID.

Chapter 2

Excess Volatility in Belief Streams: Evidence From Sports Betting Data[*],[†]

2.1 Introduction

The efficient markets hypothesis has been one of the cornerstones of financial economics at least since Fama (1970) formalized and refined the theory.[1] Financial markets are considered to be generally efficient in the sense that they are believed to immediately and completely incorporate any publicly available information into the market price. Consequently, there is no potential source of arbitrage in these markets as the market prices perfectly reflect the underlying fundamental values of assets. In the recent years the behavioral finance literature has challenged this view. Optimistic biases, overconfidence, misperceptions and imperfect learning have proven to be important determinants of consumer and investor behavior and have been shown to distort prices in many market settings. The literature has even called into question the general efficiency of financial

[*] I am grateful to Martin Peitz, Philipp Schmidt-Dengler, Emanuele Tarantino and the seminar audiences of the IO and Theory seminar in Mannheim and the 9th Alhambra meets Barcelona Workshop for their valuable comments and suggestions.

[†] This chapter is based on a dataset that was gathered and prepared for my master thesis (Dittrich, 2013). The analysis in Sections 2.5, 2.6 and 2.7.1 is related to the analysis in Dittrich (2013). The main contribution of the chapter, namely the estimation of the "hot hand" effect from high-frequency sports betting data, following from Section 2.7.2 onwards is a completely novel contribution of this chapter.

[1] Early work on the efficiency of financial markets dates back at least to Bachelier (1900) and is closely related to the work of Hayek (1945).

markets. A fairly popular and early example of this view is the theoretical noise-trader-model by De Long et al. (1991). *Empirically* showing the inefficiency of financial markets is, however, quite challenging. Besides data limitations, the fact that fundamental values themselves often remain unknown as most financial assets do not converge to or do not have well-defined terminal values in the short-run complicates the analysis.

I argue that using high-frequency betting market data to test the efficiency of market prices can help overcome a substantial share of the inherent limitations of financial market data, while still contributing to the understanding of financial markets. In fact, modern-day live-betting markets for sports events are structurally very similar to financial markets. Historically, sports betting used to be dominated by traditional bookmakers who allowed placing bets only up until the start of the event at very infrequently adjusted prices. Since the early 2000s, the expansion of the internet has facilitated a development that has brought sports betting closer to an exchange based business model. In contrast to bookmakers acting very similarly to price-setting competitive firms, betting exchanges such as Betfair act as intermediaries, merely moderating the interaction of bettors and providing the platform infrastructure. On these platforms, consumers can act on both sides of the market by freely trading pre-defined contracts on the outcomes of (sports) events at market prices. Placing bets is possible at any point before or during the event up until the end if a trading partner can be found. The advantage of analyzing betting markets can easily be seen. In contrast to financial markets, news (e.g. major events happening during match time) are easily observable and outcomes realize contemporarily as sports events inherently terminate very cleanly with an easily observable and well-defined result. The facts make empirical evaluations both much more convenient and convincible as financial contracts (bets) always converge to their terminal value.

The purpose of this chapter is to study belief and price formation in a sports betting market environment that is very similar to a standard financial market. I use price and volume data from Betfair, the world's largest online betting exchange to analyze the evolution of betting odds during US Open 2012 men's tennis matches. I argue that tennis provides an ideal setting to study belief and price formation. It attracts public attention and at the same time provides a high number of easily observable and interpretable news events during the match. In particular, whenever a point is scored in tennis, the probability of the players winning the game, the set and the match changes significantly. While this constant change is capable of producing a lot of volatility in match-winning probabilities, the market is also fairly active during the match. In contrast to other sports betting markets, belief formation is not affected by frequent suspensions of the market as a whole (e.g. after goals, penalties or red cards in soccer). Hence, the high frequency of news and the constant activity of the market closely resemble a standard financial market.

I obtain detailed second-by-second betting data on the US Open 2012 men's tennis matches, as well as detailed point data on the same set of matches. Applying a statistical model of tennis play similar to the one described by Klaassen and Magnus (2001 and 2003), I show that point-winning probabilities in tennis matches can be thought of as random processes. At any point during the match, the probability of one player winning the game is a function of the winning-on-service probabilities of both players and the state of the match. More precisely, this function is nothing but a mathematical representation of the rules of tennis. If winning-on-service probabilities are independent, i.e. there is nothing like a hot hand or momentum-based player performance in tennis, prices on the betting exchange should form according to this statistical model. In contrast, I show that bettors on average over-infer from events during the match and tend to excessively favor underdogs if they observe them performing well. This finding can be interpreted as a conditional, dynamic version of the favorite-longshot bias, an extensively discussed phenomenon in the betting and prediction markets literature that refers to a situation where the favored player's probability of winning is collectively underestimated and the longshot's (underdog's) probability of winning is overestimated. My results suggest that on average bettors slightly overestimate the favorite's probability of winning. However, if the unfavored player sends a positive signal, e.g. by winning a point, or at crucial points during the match when a lot is at stake, the market seems to overvalue the longshot. This indicates that, although bettors act relatively rational on average, they tend to over-infer from salient events during the match which leads to excessive volatility of prices compared to the fundamentals.

Section 2.2 reviews the related literature. Section 2.3 describes the Betfair betting mechanics. The datasets used throughout the analysis are introduced in Section 2.4. Section 2.5 describes the statistical model of tennis play. In Section 2.6, I show that professional tennis can be approximated as a random process. Section 2.7 analyzes the betting market data and Section 2.8 concludes.

2.2 Literature Review

This chapter is related to several strands of the economics literature: There is a large literature on optimistic biases, overconfidence and misperceptions in a variety of economic settings. Malmendier and Tate (2005 and 2008) study CEO overconfidence, showing that managerial biases can result in substantial distortion of investment decisions. Odean (1999) shows that investors have incentives to trade too much relative to what would be the individually optimal amount. Grimes (2002) analyzes students' predictions about their exam performance, providing evidence that they tend to be overconfident concerning their own test results. For a detailed review of this literature see DellaVigna (2009).

Bettors in online sports-betting markets have always been suspected to be overly optimistic with respect to winning probabilities. Using data on NFL (National Football League, American Football) betting, Levitt (2004) shows that bookmakers are likely capable of predicting game outcomes more accurately than bettors. As a result, they can systematically exploit bettor biases, thereby increasing their own profits. Kuypers (2000) derives a theoretical model of bookmaker profit maximization and shows that firms are able to increase profits by setting inefficient betting odds. He provides evidence that bookmakers indeed distort prices, which eventually leads to inefficient market outcomes. Forrest and Simmons (2008) identify team popularity as a factor that can lead to inefficient pricing using data on soccer matches in Spain and Scotland. In contrast, Page (2009) argues that a loyalty bias towards favored teams does not affect betting odds in British soccer. However, using data in the same setting, Franck et al. (2011) find evidence that bettor sentiment may indeed affect betting odds.

Betting exchanges have also been the focus of several studies. Croxson and Reade (2011) compare price determination on Betfair and price setting by traditional licensed bookmakers in soccer betting. They provide evidence that bookmakers follow prices set on the exchange and that Betfair offers superior returns on bets. Croxson and Reade (2013) study the news effect of goal arrival in soccer and find that beliefs adjust instantly. There is not much evidence of an inefficient news drift. Easton and Uylangco (2010) compare forecasts of the Klaassen and Magnus (2003) model to self recorded Betfair data, especially focusing on responses to break points[2] in tennis. In contrast to their analysis, this chapter will focus on systematic and persistent biases using a large second-by-second high-frequency dataset and highlighting the importance of dynamic factors and market liquidity. Instead of just comparing the model set up by Klaassen and Magnus (2003) to betting market data, I extend, parameterize and estimate the model on detailed high-frequency betting market data.

[2] In tennis points are called break points whenever a player has a chance to win the opponent's service game with the next point

Another string of literature related to this analysis has focused on studying the hot hand effect using data from professional sports. The hot hand has been a controversial and highly debated topic in the literature starting with the seminal paper by Gilovich et al. (1985). Gilovich et al. (1985) first documented that a hot hand effect cannot be found on data from professional basketball even though expert as well as amateur audiences hold such beliefs. Their conclusion that a belief in the hot hand resembles a general misconception is based on an analysis of both actual basketball results from the Philadelphia 76er and the Boston Celtics NBA teams as well as results from controlled shooting experiments with collegiate players. The results by Gilovich et al. (1985) have spawned a debate on whether the hot hand can be found in other sports or using different datasets. Koehler and Conley (2003) and Avugos et al. (2013) have replicated and confirmed the findings by Gilovich et al. (1985) and found no evidence of a hot hand effect when studying results of the NBA "Three Point Contest" (Koehler and Conley (2003)) or controlled shooting experiments (Avugos et al. (2013)).

More recently, there has also been a discussion about the methodology that Gilovich et al. (1985) have employed in their test for the hot hand. Miller and Sanjurjo (2018a) have casted doubt on the results presented by Gilovich et al. (1985). Using data from a controlled shooting experiment with semi-professional basketball players and an improved statistical approach that corrects downward biases in the analysis conducted by Gilovich et al. (1985) and Koehler and Conley (2003), they find statistically significant and sizable hot hand effects.[3]

The hot hand has also been analyzed in sports other than basketball. Klaassen and Magnus (2001) study the hot hand by estimating match winning probabilities from tennis point data. They derive a dynamic random coefficients panel data model to estimate winning-on-service probabilities and find that points in tennis are neither independent nor identically distributed. Klaassen and Magnus (2003) use an i.i.d. version of the Klaassen and Magnus (2001) model to forecast match winning probabilities in tennis from any point in a tennis match. Green and Zwiebel (2018) have recently provided evidence of hot hand effects in professional baseball. Green and Zwiebel (2018) also point out that there is a relevance of taking into account endogenous responses when analyzing the hot hand. This is of a particular relevance in team sports where players on streaks may for instance defended more fiercely. Suppose a player is on a streak and

[3] See also Miller and Sanjurjo (2018b) for of a discussion of the bias in Gilovich et al. (1985) and related studies. The bias pointed out by Miller and Sanjurjo (2018b) is particularly prevalent in studies with low statistical power. These concerns do therefore not apply to the same extend to the dataset analyzed in this chapter.

the opposing team shifts their attention to defending this particular player, who will then have a harder time to score even if his current success rate would be temporarily elevated above his normal performance. If this effect is not taken into account, a hot hand effect may be underestimated. This should not be a relevant concern with the setting I use due to the fact that I study 1-on-1 tennis matches so that shifting attention to a hot player cannot occur. Aside from the sports mentioned, studies have also been conducted for bowling (e.g. Dorsey-Palmateer and Smith, 2004), golf (e.g. Clark, 2003a,b, 2005; Connolly and Rendleman Jr, 2008), volleyball (Raab, 2002) or horseshoe pitching (Smith, 2003).[4]

There are also several studies that have touched on the topic of beliefs about the hot hand. That outside spectators and experts hold hot hand beliefs has already been documented by both Gilovich et al. (1985) and Camerer (1989) in the context of basketball. Gilovich et al. (1985) conduct a survey among basketball fans and also ask players and coaches about their beliefs. They find that both amateur audiences (fans) as well as experts (players and coaches) believe in streak shooting. Camerer (1989) investigates betting market spreads and concludes that evidence for a belief in the hot hand can be found in betting market data although it is likely too small to exploit. In contrast to Gilovich et al. (1985), Miller and Sanjurjo (2018a) find that basketball players are in fact able to correctly predict when their teammates show signs of a hot hand effect. Based on this finding they conclude that the belief in the hot hand in fact exists but is not a fallacy.[5]

2.3 Betting on Betfair

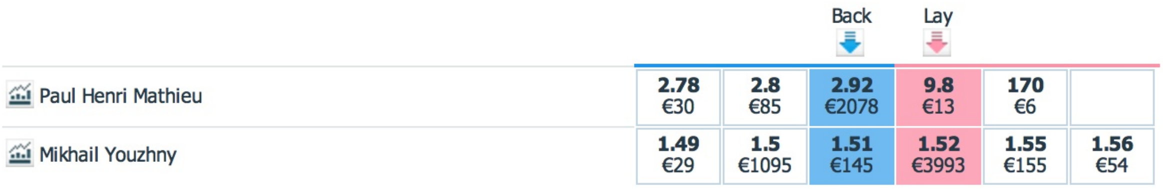
Traditionally, sports betting markets have been dominated by licensed bookmakers, accepting bets from the public at previously quoted betting odds. More recently, betting exchanges have become increasingly popular. These platforms allow individuals to bet against each other directly. Betfair was by far the leading company in this market segment in 2012 with a self-reported number of four million registered customers and more

[4] See Bar-Eli et al. (2006) for a comprehensive review of the literature.

[5] The hot hand and related gamblers fallacy has also been documented outside of sports settings. Examples include casino gambling (e.g. Croson and Sundali, 2005; Narayanan and Manchanda, 2012; Sundali and Croson, 2006) or lotteries (e.g. Guryan and Kearney, 2008; Yuan et al., 2014).

than seven million transactions per day.[6,7] Figure 2.2 shows the quickly increasing popularity of Betfair in the early 2000s by depicting an index of the frequency of worldwide Google searches for the term “betfair” from 2004 to 2019. Noticeable is especially the stark increase in searches from 2004 to 2010. The volumes traded on the betting exchange are also fairly significant. Figure 2.3 shows the total amount of money (in million GBP) matched on the two players competing in the US Open 2012 final.

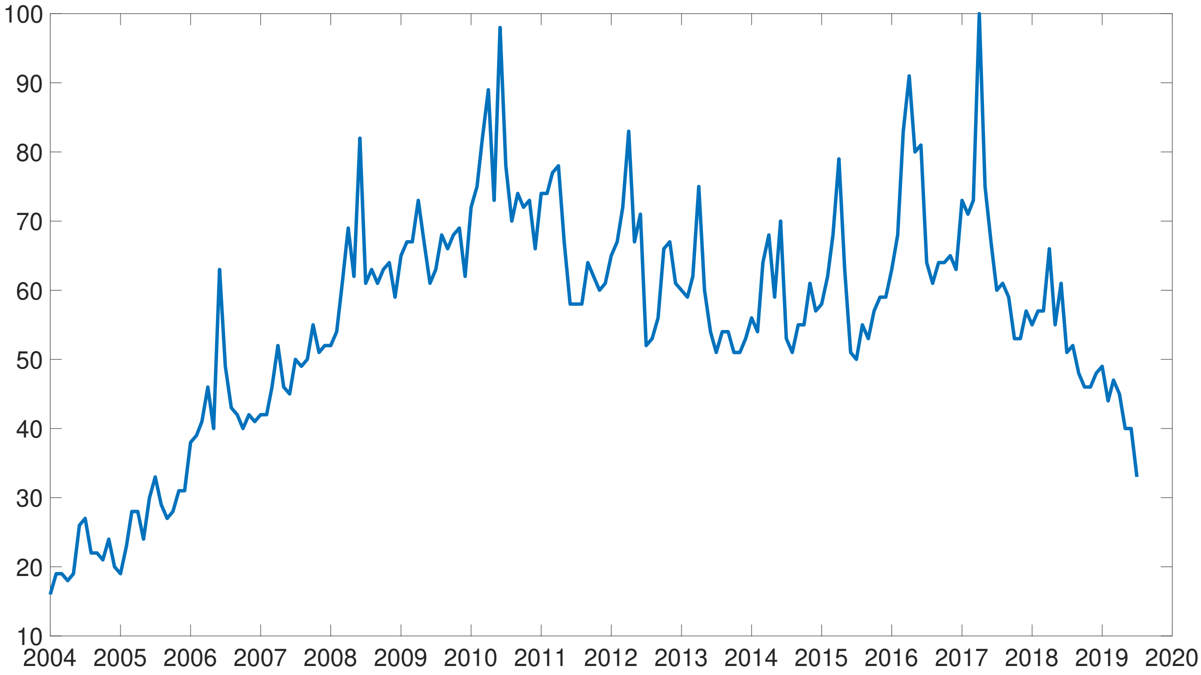
Figure 2.1: Betfair order book example: Mathieu vs. Youzhny



			Back	Lay		
Paul Henri Mathieu	2.78 €30	2.8 €85	2.92 €2078	9.8 €13	170 €6	
Mikhail Youzhny	1.49 €29	1.5 €1095	1.51 €145	1.52 €3993	1.55 €155	1.56 €54

Source: www.betfair.com (accessed on 22/07/2013)

Figure 2.2: Frequency of Google searches for the term “betfair”

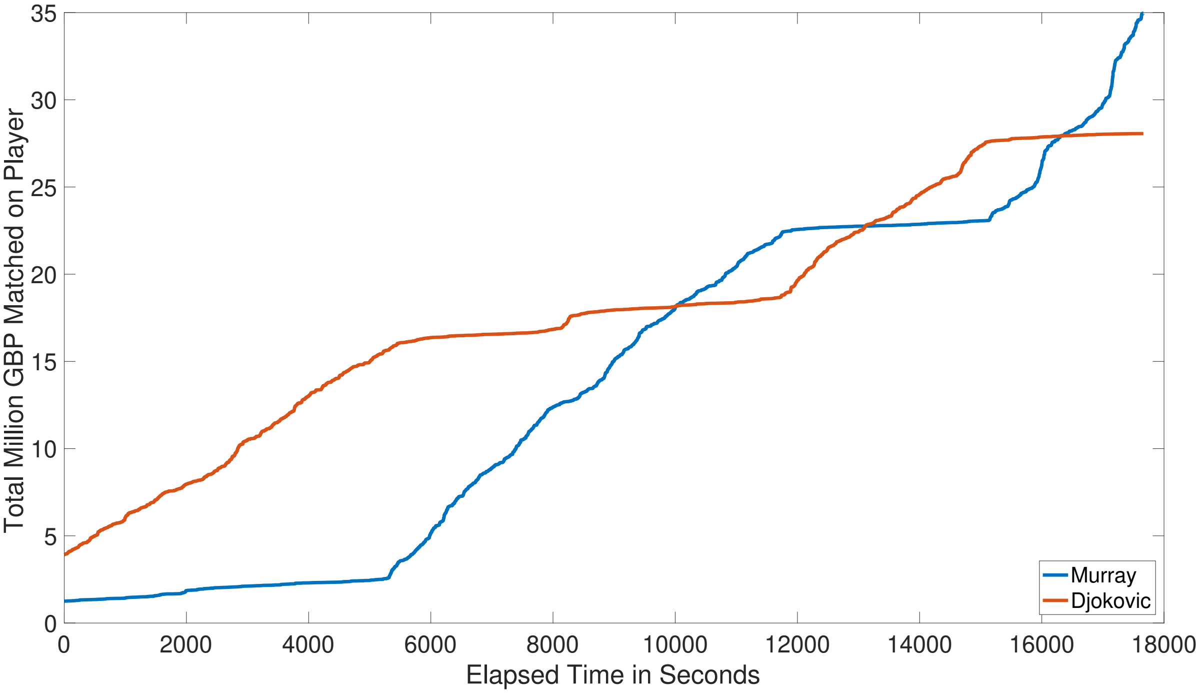


Source: Google Trends (accessed on 04/07/2019)

[6] See https://media.investis.com/B/Betfair/PDFs/Annual-Reports/Betfair_Annual_Report_2012.pdf (accessed on 28/06/2022).

[7] In February 2016 Betfair merged with Paddy Power. The merged firm is now known as Flutter Entertainment.

Figure 2.3: Money matched on players during US Open 2012 final



Source: Betfair, data recorded and supplied by Fracsoft

I continue by briefly describing the mechanics behind Betfair. Suppose agent $\mathcal{A}$ intends to bet on (back) outcome O_1 and agent $\mathcal{B}$ is willing to take the opposite position, i.e. wants to bet against (lay) the same outcome O_1. In so-called fixed-odds betting, $\mathcal{A}$ and $\mathcal{B}$ agree on an odds-ratio, $o \geq 1$, which is payed for every dollar at stake to the agent whose bet has succeeded. Typically, the position of agent $\mathcal{B}$ is taken by licensed bookmakers who offer betting at a fixed "price". Consider the following example: Suppose $\mathcal{A}$ would want to bet \$10 on Andy Murray playing Novak Djokovic in the 2012 US Open Final. $\mathcal{A}$ may buy a future contract from $\mathcal{B}$ that pays out $\$10 \cdot o$ in case $\mathcal{A}$'s bet is correct and 0 otherwise.

In the bookmaker setting, $\mathcal{B}$ acts like a price setting firm, maximizing profits conditional on his own belief about match outcomes. Betting exchanges like Betfair are organized in a different way. The main distinction is, that any participant can take the position of $\mathcal{B}$ and offer bets to the public by announcing a price (odds ratio) and the corresponding volume. Similarly, anyone is allowed to announce order prices with the corresponding volumes. Aggregating those offerings yields the order book. Figure 2.1 shows an example for an order book on the Betfair website. Betfair itself does not directly benefit from each bet but keeps a portion of the profits generated by the successful bets as provision. It is therefore not directly involved in the market transaction.[8] Because of these mechanics, Betfair provides an ideal setting to study the formation of consumer beliefs since the prices are determined by offers and orders itself but not set by a profit maximizing firm.

2.4 Data

I use high frequency data on betting odds from Betfair, the world's largest betting exchange.[9] The data is recorded second-by-second and covers the pre- as well as the in-play phase of almost every US Open 2012 tennis match. For every second in each of the matches I have information about the last price matched on the market, the three best back and lay prices from the order book and all the corresponding available volumes as well as the total amount matched on each of the options.

Match events are considered news that have the power to directly affect match outcome probabilities and thus market prices. I use in-play match data recorded and provided online by the official Grand Slam partner IBM. A tennis match consists of points, games and sets. I have detailed information about every point that is played throughout the match, including point winner, point server, double faults, aces, unforced errors and current game as well as set score. Moreover, I know the elapsed match time at any service, which allows me to match the point data to the empirical belief streams I build from the Betfair dataset. This matching procedure is described in more detail in Section 2.7.

To complement the high-frequency data, I acquire match level data on current rankings and ranking points of both players as well as further game characteristics from www.tennis-data.co.uk. Rankings and ranking points are fixed throughout the tournament and used as proxies for player quality. I can match the information contained in each of the three datasets by player names.[10]

As the analysis requires the exact matching of both panels according to their timestamps, dealing with measurement error is crucial. There are multiple sources of such

[8] Similarly, bookmakers have over-rounds (markups) on their betting odds (prices). Comparing bookmaker prices and expanses on the Betfair exchange, Croxson and Reade (2011) find that Betfair offers superior returns compared to traditional bookmakers.

[9] The data is recorded and provided by the company Fracsoft.

[10] Unfortunately, the point data does not include player names but only a unique identifier. I extract the player names from the official US Open 2012 website which features both, names and the unique identifier.

error in my tennis point data. In some matches time intervals between points are obviously incorrect. In particular, intervals between points are equally spaced. I drop those matches from the sample whenever I find three equally spaced time intervals in a row. Moreover, I drop all matches where at one or more points the winner, the server or the score are unknown and all matches where the data recording is not complete. For some matches there is no provision of point data at all, whereas for a few others recording of the data stops before the match has officially ended.[11] It is further important to split the sample by gender. Klaassen and Magnus (2001) show that men's tennis is statistically very different from women's tennis. As the focus of this chapter is not on gender differences, I will limit the analysis on men's tennis.

Table 2.1: Summary statistics for point data

Variable	Obs	Mean	Std. Dev.	Min	Max	Median
Service point won	15730	.64	.48	0	1	1
Importance of point	15730	.04	.04	0	.45	.03
Ranking point sum	15730	4.28	3.76	.93	18.56	2.84
Ranking point difference	15730	-.09	3.38	-11.55	11.55	-.09
Ace	15730	.09	.28	0	1	0
Server unforced error	15730	.19	.39	0	1	0
Returner unforced error	15730	.14	.35	0	1	0
Double fault	15730	.04	.2	0	1	0
Server net point	15730	.15	.35	0	1	0
Returner net point	15730	.06	.24	0	1	0
Set score difference	15730	-.02	1.06	-2	2	0
Time between points (across games)	2526	86.84	52.45	0	299	67
Time between points (within game)	13204	33.77	15.79	0	300	31

There are fewer limitations with the Betfair data. While the data itself is generally of good quality, I miss certain matches that were not recorded. Given the cleaned point data, this is the case for exactly three matches. Moreover, I drop seven matches, where I have serious concerns regarding incomplete recording.[12] A last, but nevertheless serious, concern are injuries which may indeed be incorporated into the belief streams obtained from the Betfair data. In contrast, the point data does not include any information on injuries unless the match ends prematurely. Consequently, matches with injuries should be removed from the dataset as the model described in the following section is not

[11] This is not only the case for presumably less popular first and second round matches. Interestingly, point data is also missing for quite a number of half and quarter finals.

[12] In particular, the duration of the matches according to the Betfair data is shorter than the duration according to the point data.

capable of incorporating injury and retirement risk.[13] Cutmore and Knottenbelt (2013) argue that it is possible to infer occurrences of injuries through the odds gap between the standard Betfair "Match Odds" market and the Betfair "Set Betting" market, which pays differently in case of retirement of a player. Using their methodology, I have not found any evidence of serious injuries during matches in my dataset. For the US Open 2012 this leaves me with 15,730 observed points in 69 matches. Table 2.1 shows summary statistics for the matched point data.

2.5 A Structural Model of Tennis Play

In this section, I describe the statistical model of tennis play that will be used in the estimation. Figure 2.4 depicts the event tree of a tennis game. It shows the transition from point to point with the corresponding transition probabilities. Here, s represents the probability of winning a point on service. This specific structure enables me to calculate the probability of winning a game at any point by simply adding up the probabilities of the histories that lead to the server being the game winner. Having obtained this probability I can compute the probability of winning the current set (given the set score) and finally the probability of winning the whole match in a similar way.

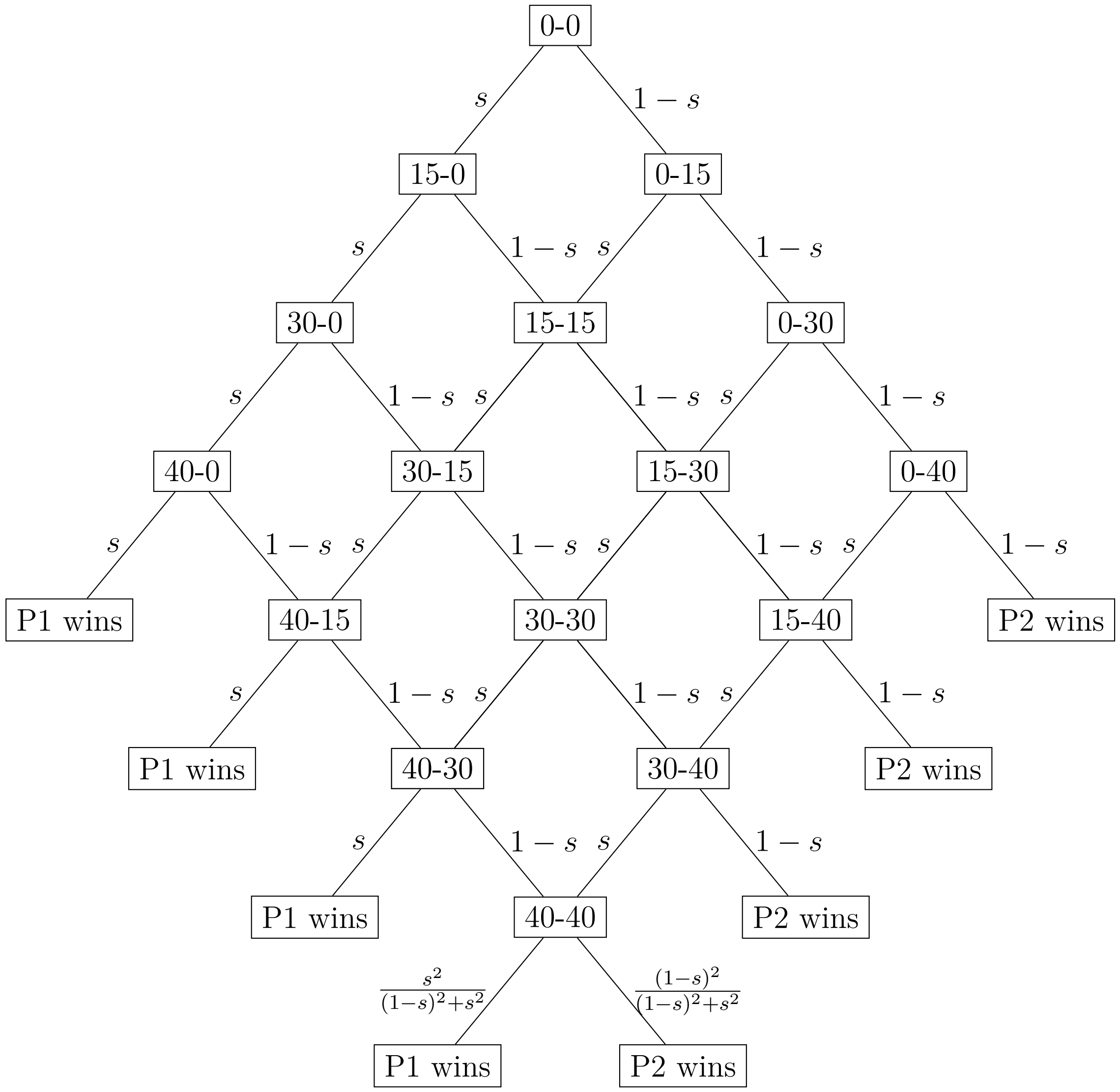
One particular issue is that tennis games do not have a fixed number of points that are played. In tennis, players have to win a game with a two point advantage. Therefore, given the players arrive at a deuce (i.e. a score of 40-40), the probability of winning the current game is the probability of winning two points in a row from a deuce after any history that has not yet produced a winner. Suppose we observe the following histories of point winners after having arrived at 40-40:

- P1 (AD-40), P2 (40-40), P2 (40-AD), P1 (40-40), P1 (AD-40), P1 (Game P1)
- P1 (AD-40), P2 (40-40), P1 (AD-40), P1 (Game P1)
- ...

The pattern's common property is that, starting at deuce, a player has to win two points in a row to win the game. Moreover, from one deuce to another there can be at most two points and each player has to win exactly one of them. Otherwise, one of the players would have won the game already. We thus observe that every possible history H consists of sub-histories of point winners H^S of the following two types:

[13] Injuries would correspond to a sudden drop in point-winning probability that cannot be matched to any of the observables.

Figure 2.4: Evolution of point score in a tennis game



Source: Depiction similar to Huang et al. (2011)

- P1 (AD-40), P2 (40-40)
- P2 (40-AD), P1 (40-40)

Therefore, the probability of winning after a deuce is given by

$$\pi_{P1} = s^2 \sum_{i=0}^{\infty} \left[2s(1-s)\right]^i .$$

Notice that this can be written as

$$\sum_{i=0}^{\infty} [2s(1-s)]^i = \frac{1}{1-2s(1-s)} = \frac{1}{(1-s)^2+s^2}$$

and therefore

$$\pi_{P1} = \frac{s^2}{(1-s)^2+s^2}, \qquad \pi_{P2} = \frac{(1-s)^2}{(1-s)^2+s^2}.$$

The question remains to what extend this model is capable of capturing the forecast of a rational agent. I think that it is indeed reasonable to assume that match-winning probabilities are tied to the relative and absolute strength of players. I further suppose that a player's probability of winning the current point should be affected by the strength of players. Winning-on-service probabilities, the current point score, the history of the match and knowledge of tennis rules are a good approximation of the information set a sophisticated bettor would use to form a belief about match-winning probabilities.

I want to emphasize that, throughout the match, winning-on-service probabilities implied by betting market prices are volatile rather than static. Hence, I need to extend the model to allow for dynamic adjustment of winning on service probabilities. I will therefore formalize the notation slightly. Let S_{it} describe the state of the match at point t from the perspective of player i. S_{it} simply represents the currently serving player and the relative score of the match. More precisely, S_{it} collects the difference in sets, games and points won from the view of player i.

$$S_{it} = \{sets_i - sets_j; games_i - games_j; points_i - points_j; service_t = i\}$$

To illustrate, consider the following hypothetical example: Suppose Andy Murray serves against Novak Djokovic. Murray has won the first set and two games in the second set, while Djokovic has won only one game in the second set. The current game's point score is 30-0, i.e. Murray has won two points, whereas Djokovic has not yet won a point. Then, from the view of Murray we have

$$S_{Murray,t} = \{1; 1; 2; service_t = Murray\}.$$

For the statistical model it is enough to know that Djokovic is returning at point t and that he is one set, one game and two points behind. To describe the state of the match, I do not require information about the particular history. The reason for this is that the state will only be used to calculate how many sets, games and points anyone of the players is missing to win the match. History dependence, however, can only possibly affect winning-on-service probabilities.

The probability of winning the match is not only a function of the state, but also a function of the point-winning probabilities. Let p_{it} be player i's probability of winning point t on service. If points were independent ($p_{it} = p_i \ \forall t$) for any possible state, the match could be fully described by the winning-on-service probabilities of both players and the rules of tennis. The literature has justified and used this assumption to forecast tennis matches, using estimated winning-on-service probabilities for both players. Notice that the model does allow for heterogeneity across players. Saying that Roger Federer is a better player than Philipp Kohlschreiber is nothing but saying that $p_{Federer} > p_{Kohlschreiber}$, i.e. Federer's probability of winning a point on service is higher than Kohlschreiber's and hence also Federer's probability of winning the match will be higher. I will later treat the winning-on-service probabilities as functions of the match's history. For now, I just assume that they are specific to players.

Obviously, if $p_{it} = p_i \ \forall t$ and points are treated as if they were independent, there is no scope to test for movement in winning-on-service probabilities. I will extend the model to allow for changing winning-on-service probabilities, while still keeping it computationally tractable. For this to work, I will assume that winning-on-service probabilities are constant "in the future". While this assumption appears rather strict at first sight and is first and foremost made to ensure tractability of the model, I argue that it is not too far-fetched. In principle, the underlying assumption is that players are hit by transitory shocks to their performance that fade out and are anticipated to fade out immediately after the current point is played. In other words, there is a common belief that players' strength will revert to a game-specific mean after the next point is played. Given the structure of tennis and the inherent uncertainty about the future game path, at any point during the match, the focus is clearly on the point that is currently played. Given the relatively fast pace of the game, I believe that it is reasonable to assume that projections about the general future performance of a player in a given match are mainly functions of his relative skill level rather than his instantaneous performance. On the contrary, I think bettors are very likely to focus on the immediate impact of a good or bad play on the subsequently played point. In the end, I think that there is very little reason to believe that a player's general relative skill level is re-evaluated within a given service game as points are played out very quickly. Notice that I will explicitly allow winning-on-service probabilities to adjust across games, i.e. whenever there is a slightly longer break and the serving player changes.

Specifically, let $M(p_{it}, p_{jt}, S_{it})$ be a function that maps the winning-on-service probabilities of both players and the state of the match into match-winning probabilities. As I have outlined above, $M(\cdot)$ will be nothing but the rules of tennis. Let $S^{\oplus}_{i,t+1}$ ($S^{\ominus}_{i,t+1}$) be the state of the match, given that player i has won (lost) the last point. Further, assume that $p_{i\tau} = \bar{p}_{ig}$ $\forall \tau > t$ if $G(\tau) = G(t)$, where $G(\cdot)$ is the number of the game g and t is the number of the point, i.e. having won the last point affects the probability of winning the current point, but not the probability of winning points on services in the future within a given game. As I have discussed above, this assumption is somewhat strict but not unreasonable given the pace of the game and simplifies the computations a lot. Observe that there are only two transition possibilities from the current state: Either the serving player wins the point or he does not. The first case happens with probability p_{it}, whereas the second one happens with probability $(1 - p_{it})$. We can thus split up the match-winning probability at any point t as follows:

$$M(p_{it}, p_{ig}, p_{jg}, S_{it}) = p_{it} M(p_{ig}, p_{jg}, S^{\oplus}_{i,t+1}) + (1 - p_{it}) M(p_{ig}, p_{jg}, S^{\ominus}_{i,t+1})$$

In Section 2.7, I will show how this model can be applied to the betting market data.

2.6 Point Data Analysis

In this section, I show that treating winning-on-service probabilities in professional tennis as independent is indeed a good benchmark. Although this has already been shown by Klaassen and Magnus (2001) using data on Wimbledon matches, I will repeat the analysis for my dataset and use more robust estimation methods.

In contrast to Klaassen and Magnus (2001), I use a dynamic Probit model with random effects instead of a simple linear probability model. Even though the predicted probabilities are oftentimes far away from the boundaries using the methodology of Klaassen and Magnus (2001), a dynamic random effects Probit model (as outlined by e.g. Jeffrey Marc Wooldridge (2005)) fits the data better and is still computationally feasible. Assuming that the probability of winning a point on service depends on numerous unobservable factors, a random effects model seems to be an appropriate choice, as it accounts for unobservable heterogeneity. Notice that estimating a fixed effects model is not appropriate, since it prevents estimation of player-match-specific time-invariant coefficients. More precisely, relative strength of the players (measured by the difference in ranking points) and overall quality of the match (measured by the sum of the ranking points) cannot be separately identified from the fixed effects.

Let y_{ijt} be 1 if player i wins his t-th service point against player j and 0 otherwise. The random effects Probit model is defined as follows:

$$P(y_{ijt} = 1|Q_{ij}, H_{ijt}) = \Phi\left(Q_{ij}\beta + H_{ijt}\delta + \eta_{ij}\right)$$

Here, Q_{ij} is the quality of player i when playing against player j, H_{ijt} is the set of regressors reflecting the history of the match at point t and also includes the lagged dependent variable $y_{ij,t-1}$ that measures the hot hand effect,[14] η_{ij} is the match-specific random effect and $\Phi(\cdot)$ is the cumulative standard normal distribution function. Notice that $\Phi(\cdot)$ could be replaced by other appropriate distribution functions. I will additionally provide results using a linear probability function. In this case the corresponding linear probability model would be

$$y_{ijt} = Q_{ij}\beta + H_{ijt}\delta + \eta_{ij} + \varepsilon_{ijt}$$

where ε_{ijt} is the respective error term. For simplicity, I subsequently omit the index of the respective opponent and write

$$P(y_{it} = 1|Q_i, H_{it}) = \Phi\left(Q_i\beta + H_{it}\delta + \eta_i\right)$$

or

$$y_{it} = Q_i\beta + H_{it}\delta + \eta_i + \varepsilon_{it}$$

respectively. Assuming that the η_i are i.i.d. $N\left(0, \sigma_\eta^2\right)$, the Probit model can be estimated via standard maximum likelihood methods. As the random effect cannot be integrated out analytically, the log-likelihood is approximated by adaptive Gauss-Hermite quadrature. The linear probability model can be estimated via OLS.

As I am interested if there is a hot hand in professional tennis, I need to incorporate dynamic factors into the estimation.[15] To keep the estimation feasible, I restrict histories to be relevant during service games only. Including more potentially relevant information

[14] The hot hand literature has frequently also considered longer histories to determine the hot state. Gilovich et al. (1985) has investigated the hot hand effect using information on the previous 1 to 3 shots. Green and Zwiebel (2018) has defined the hot hand as the average winning rate over the previous 10, 25 or 40 actions. As games in tennis are rather short, I restrict attention to the outcome of the previous point.

[15] Such an effect could occur through multiple psychological channels. One may suspect that a player who has won a point is more confident or motivated in the following play. Similarly, the opponent's strength may be affected by losing the previous point.

about the match history (e.g. score differences or winning streaks) is left for future work. During tiebreaks the service pattern differs substantially, because the server changes every two points. To ensure that the different serving pattern during tiebreaks does not interfere with the analysis, I drop all tiebreak-observations from my sample.

Similar to Klaassen and Magnus (2001), I want to test whether the points of the match are independent and identically distributed. To test the former, I add information about previous match outcomes. To account for the latter, I measure the statistical importance of the current point. I do take into account the relevance of the history for the current point. The last point of a game may, however, not affect the probability of winning the first point of the next game. First of all, there is a significant time break between two games and secondly there is a change of service, implying a structural change in the point-winning probability. Following Klaassen and Magnus (2001), I therefore restrict the history to be relevant only within games. This is done by defining a dummy variable, d_{it}, that is 1 whenever the previous point was in the same game as the point currently played i.e.

$$d_{it} = \begin{cases} 0 & \text{if } G(t) = G(t-1) \\ 1 & \text{else} \end{cases},$$

where $G(\cdot)$ is the number of the game and t is the number of the current point. As I want to focus on histories within games, I have to zero out the first observation in every game. Thus, the lagged dependent variable is given by

$$\tilde{y}_{i,t-1} = \begin{cases} y_{i,t-1} & \text{if } G(t) = G(t-1) \\ 0 & \text{else} \end{cases}.$$

Regarding the importance of a point, I_t, I follow Morris (1977) and define

$$I_t = P(y_{iT} = 1 | y_t = 1) - P(y_{iT} = 1 | y_t = 0), \tag{2.1}$$

where T is the very last point of the game. Equation (2.1) is the difference in match-winning probabilities conditional on whether the current point is won or lost. A set point, for example, is considered to be more important than the first point of the game as the former gets the set-winning player much closer to victory than the latter. Both the serving and the returning player may therefore put in more effort when their benefit from winning the point is larger, which in turn may affect their probability of winning a point.

Furthermore, I will test if the degree of dependence as well as the effect of the importance of a point on the winning-on-service probability depend on the absolute and relative quality of both players. This accounts for the possibility that players' behavior and "professionalism" depends on the circumstances of the match. In particular, lower ranking, less professionally experienced players could be more susceptible to (negative) momentum – especially when playing against higher ranked opponents. Accordingly, a hot hand effect could occur if the lower ranked player gives up on a return game after the higher ranked player scores a point on own service. Similarly, the lower ranked player might gain confidence and trust in his abilities after scoring a point on own service. Both effects may not be present if two high rank players are playing against each other. Interacting the dynamic regressors with the match quality measures discussed above should account for that possibility.

Table 2.2: Point data estimation results

	Probit model		Linear prob. model	
Last point won	0.049	(0.036)	0.020	(0.013)
× ranking point sum	−0.001	(0.006)	−0.001	(0.002)
× ranking point diff.	0.002	(0.006)	0.001	(0.002)
Importance of point	−0.568	(0.423)	−0.228	(0.155)
× ranking point sum	0.073	(0.080)	0.027	(0.030)
× ranking point diff.	0.075	(0.108)	0.018	(0.039)
Ranking point sum	−0.005	(0.006)	−0.002	(0.002)
Ranking point difference	0.026^{***}	(0.007)	0.009^{***}	(0.002)
Last point ace	0.063	(0.046)	0.030^{*}	(0.017)
Last point double fault	−0.041	(0.059)	−0.012	(0.022)
L. p. server unforced error	−0.026	(0.034)	−0.014	(0.013)
L. p. returner unforced error	-0.072^{**}	(0.036)	-0.029^{**}	(0.013)
Last point server net point	−0.032	(0.033)	−0.013	(0.012)
Last point returner net point	0.023	(0.047)	0.006	(0.017)
Set score difference	0.020	(0.018)	0.013^{**}	(0.006)
× ranking point sum	0.001	(0.003)	0.001	(0.001)
× ranking point diff.	−0.002	(0.003)	−0.001	(0.001)
Constant	0.380^{***}	(0.035)	0.646^{***}	(0.012)
Observations	15730		15730	

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.2 shows the results of the estimation procedure described above. The most relevant estimates are the coefficients on the "last point won" and the corresponding interaction terms. I do not observe that any of these parameters is estimated to be statistically significantly different from zero. This implies that, independent of the absolute and relative strength of players, there is no evidence pointing towards a hot hand or "winning mood". Thus, treating points in professional tennis as independent appears to be the correct benchmark across matches. In fact, winning-on service probabilities mainly appear to be a function of the quality difference between players.

This result confirms previous findings by Klaassen and Magnus (2001) and is perfectly well explained by the fact that, at least in the short run, professional players should not be susceptible to the history of a game. Because winning a tennis game (and ultimately also a tennis match) requires scoring a final decisive point, it is always optimal for players to put in their best effort to win a given point. This is all the more true as the tennis tournament analyzed here is played according to the single-elimination system, which does not allow a player to move on after a loss. Professional players seem to understand that putting in the optimal effort to win a given point is the best strategy independent of who scored the last point.[16]

Nevertheless, it should be highlighted that there is some evidence of match events affecting players' winning-on-service probability. In particular, it seems that scoring aces (direct point wins with the services) as well as unforced errors and the set score difference have an effect on players' winning-on-service probability. Aces and return player unforced errors are relatively rare events which according to the estimation results may indeed affect players' short run performance. Interestingly, while the effect of aces corresponds to the intuition that scoring an ace leads to momentum and thus increases the probability of scoring the next service point, the effect of the return player's unforced error has the opposite sign. This implies that the return player would actually perform especially well after making an unforced error, hence "making up" for the earlier mistake.

The positive and significant effect of the set score difference (at least in the linear probability model) can be taken as evidence of a long-term momentum effect. The

[16] This may partially differ in other tournament modes or leagues systems if the contest is not effort inducing in every match, e.g. when players or teams cannot improve or worsen their final ranking independent of the match outcome. There is for instance an everlasting discussion about team motivation on the final match days in soccer leagues if teams have "nothing to play for" (see e.g. an analysis for the English Premier League by Jonathan Liew in "The Daily Telegraph", www.telegraph.co.uk/sport/football/competitions/premier-league/10810560/Do-Premier-League-teams-ease-up-when-they-have-nothing-to-play-for.html, accessed on 28/06/2022).

further the serving player is ahead (behind) in terms of the set scores, the higher (lower) is his probability of winning a point on service. It should be noted though, that this coefficient may partly measure other match trends and effects like players' form on the particular day. All else given, the better performing player in a particular match has a higher chance of winning the first set and therefore is also favored in the following sets. Hence, the coefficient on the set score difference may partially reflect that the better performing (not necessarily better ranked) player in a particular match will tend to have set score advantages towards the later stages of the match. This is not perfectly controlled by the other match quality measures as these do not reflect players' form of the day. A different explanation for the coefficient on set score could be that it partly measures unobserved trends during a particular match. For instance, if a player is (slightly) injured and still continues playing, his performance may decrease over the course of the match and he may fall behind in terms of sets at the same time. An effect like this would also be picked up by the set score coefficient.

Finally, it seems that the importance of a particular point, in terms of the point's contribution to the match winning probability of a player, does not affect the winning on service probability. While it should be expected that players try very hard to win important points, it is important to note that this does not only hold for the point server but also for the point returner. In fact, because tennis matches cannot end in a tie in the tournament setting analyzed here, points are always equally important to both players as an increase in the match winning probability of one player corresponds to an equally-sized decrease in the match winning probability of the other player. Hence, if both players would try harder to win an important point, an effect on the server's point-winning probability may not occur.

2.7 Betfair Data Analysis

In this section, I will proceed by bringing the model derived in Section 2.5 to the betting market data. Before discussing the empirical model, I show how probabilities are obtained from the price data and how the betting market data is matched to the point data.

2.7.1 Extracting Empirical Belief Streams from Betfair Data

Whereas the model derived in Section 2.5 operates with probabilities, the Betfair data is provided in prices. To test the structural model on the Betfair data, I need to map the observed prices into the probability space. Suppose o_{it} is the equilibrium market odds ratio[17] for a bet on player i at time t. Similarly, let o_{jt} be the corresponding odds ratio on the opponent, j, at the same time. As the events market participants bet on will actually realize after the match has ended, market prices are directly linked to beliefs. More precisely, inverting and normalizing the odds ratios will yield the corresponding match-winning probabilities. The higher the odds ratio, i.e. the price that is paid in case the bet was successful, the lower the probability of the corresponding event.

Normalizing the inverse odds ratios is necessary to obtain the correct probabilities. For the non-normalized odds-ratio to imply the correct probabilities, I would require that they some up to one, i.e.

$$\frac{1}{o_i} + \frac{1}{o_j} = 1 \quad \Rightarrow \quad \mathcal{O} = 0 = 1 - \frac{1}{o_i} - \frac{1}{o_j}. \tag{2.2}$$

Equation (2.2) is called the "over-round", $\mathcal{O}$, of the odds and usually estimated to be around ten percent for traditional bookmaker betting markets. I observe that there is almost no over-round in my betting market data. I estimate the mean over-round to be 1.0015 with a standard deviation of 0.0127. Figure 2.5 shows the estimated density of the over-round using all available observations in the dataset. Even during the matches, the over-round implied by the odds changes constantly. Figure 2.6 depicts the over-round during the US Open 2012 Men's Final between Andy Murray and Novak Djokovic. Despite a number of spikes and short-run deviations, the over-round does not systematically differ from 1 both within and across matches.

Even though there may not be a systematic deviation in the over-round, at any point t, I normalize the inverse odds ratios of both players by the sum of the inverse odds ratios at the same point. This ensures the comparability of the calculated probabilities over time as well as across matches.

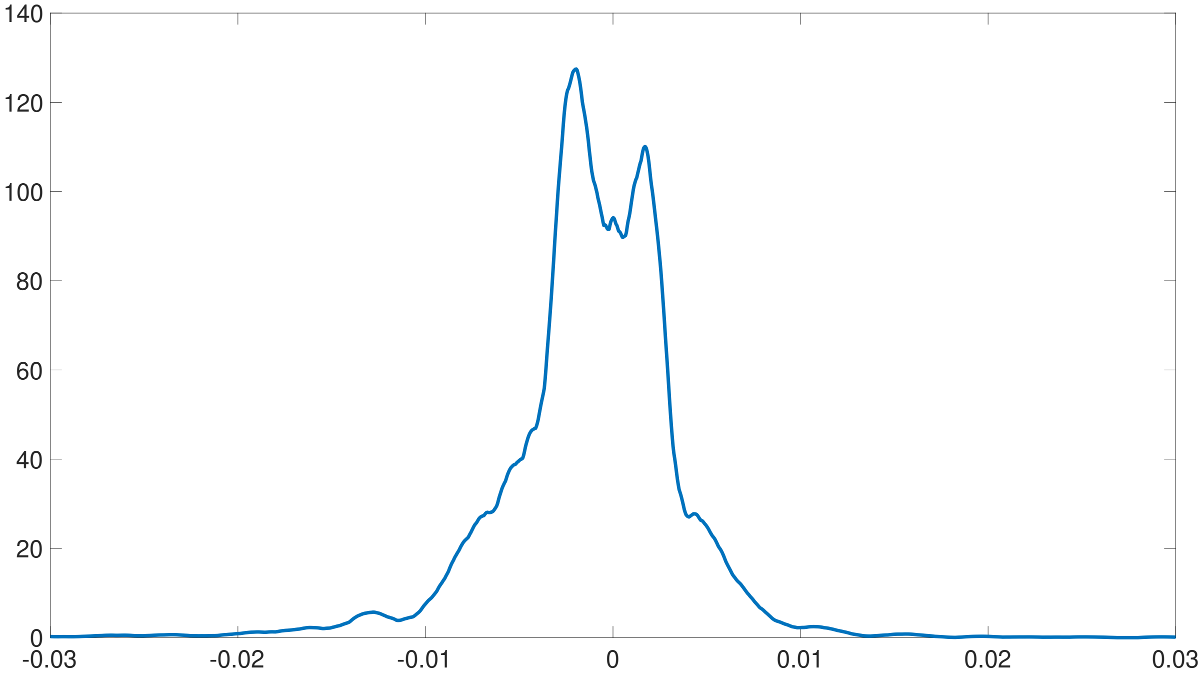
Suppose π_{it} is player i's match-winning probability at time t and π_{jt} is the respective probability put on player j. Then,

$$\pi_{it} = \frac{\frac{1}{o_i}}{\frac{1}{o_i} + \frac{1}{o_j}} \qquad \text{and} \qquad \pi_{jt} = \frac{\frac{1}{o_j}}{\frac{1}{o_i} + \frac{1}{o_j}}$$

are player i's and player j's normalized match-winning probabilities at time t. By construction, π_{it} and π_{jt} add up to one at any time t.

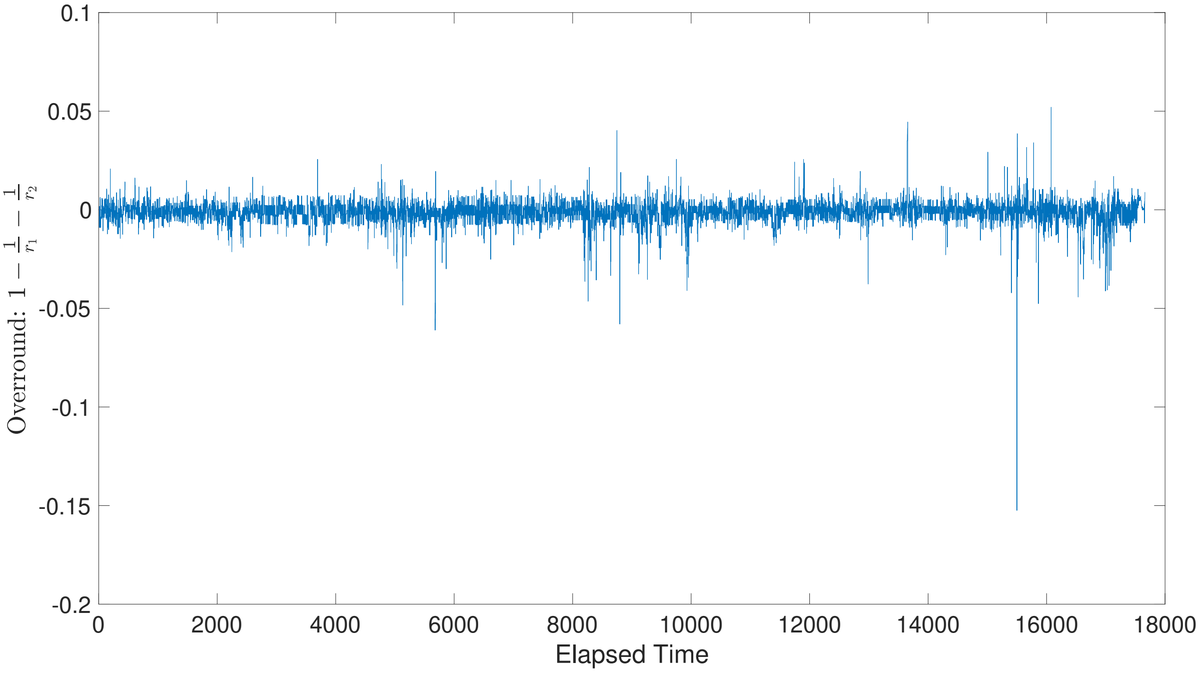
[17] I will discuss different equilibrium price concepts later on.

Figure 2.5: Estimated density of over-round in Betfair markets



Note: Epanechnikov kernel, Bandwidth: 0.0006, Observations: 1,079,667

Figure 2.6: Over-round over time: US Open 2012 final



2.7.2 Matching the Time Series

To study how the belief streams calculated from the price data react to events during the match, it is necessary to perfectly match both time series. If both time series are not perfectly aligned, I cannot relate price movements to news events that occur throughout the match. Comparing a very simple model forecast done similarly to Klaassen and Magnus (2003) and the Betfair data, it is evident that both panels are indeed not perfectly aligned. This is due to the fact that Betfair sometimes sets the market active several minutes before the first service.[18]

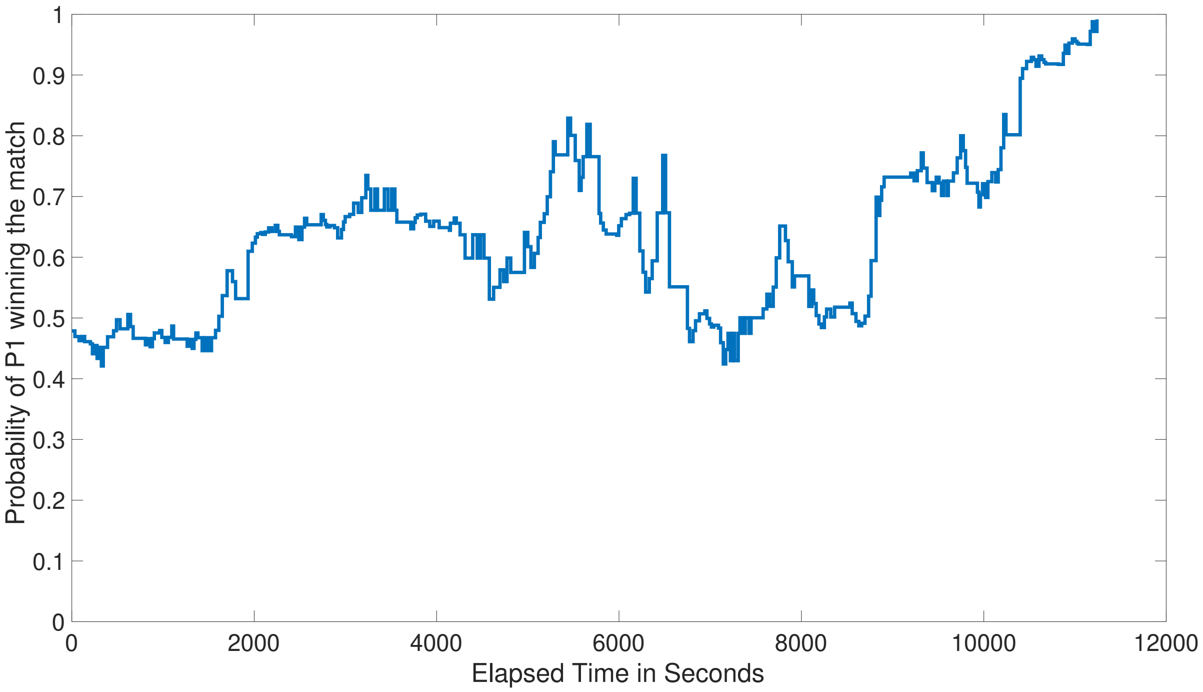
Even though this is not always the case[19] it is necessary to correct imprecise starting marks. Figure 2.7 shows the model forecast for a match from the dataset. Based on the discussion in Section 2.5 and the results in Section 2.6, I use the winning-on-service probabilities predicted by the point data regression model to construct a forecast of the match winning probability at any point of the match. For the same example match, Figure 2.8 shows the aggregated empirical beliefs computed form the betting market data. It is readily seen that both time series do not start at the same time. In the example, for about 25 minutes the betting market prices remain approximately stable at the pre-match implied probability of player 1 winning the match. Similarly, after the match has officially ended, the market remains active with no betting activity taking place. In contrast to the betting market data, the point data does not contain the exact timestamps. The only information available in the point data is the elapsed time between two points. Therefore, matching by timestamps is not possible.

Nevertheless, information about the exact time interval between two serves enables me to match the data streams. Let $\boldsymbol{M}$ be the $\boldsymbol{(n \times 1)}$ vector of probabilities predicted by the forecasting model. I construct the $\boldsymbol{(n \times 1)}$ vector $\boldsymbol{D(t)}$ with probabilities calculated from the betting market data in the following way: Starting at time t, I pick probabilities calculated from the betting market price data using the time intervals observed from the

[18] This constitutes an issue as I only observe the time between points but not the exact time of the day at which a point was played, whereas the betting data includes standardized timestamps. Consequently, I need the first service at match time “00:00:00” to coincide with the correct price. By similar arguments matching from the end is also not possible as firstly, the market oftentimes technically remains in an active state even after the match has ended and secondly, betting activity oftentimes comes to a standstill before the match has officially ended. As both datasets contain data that is exact to the second, matching based on the time intervals between points remains to be the technically sound option.

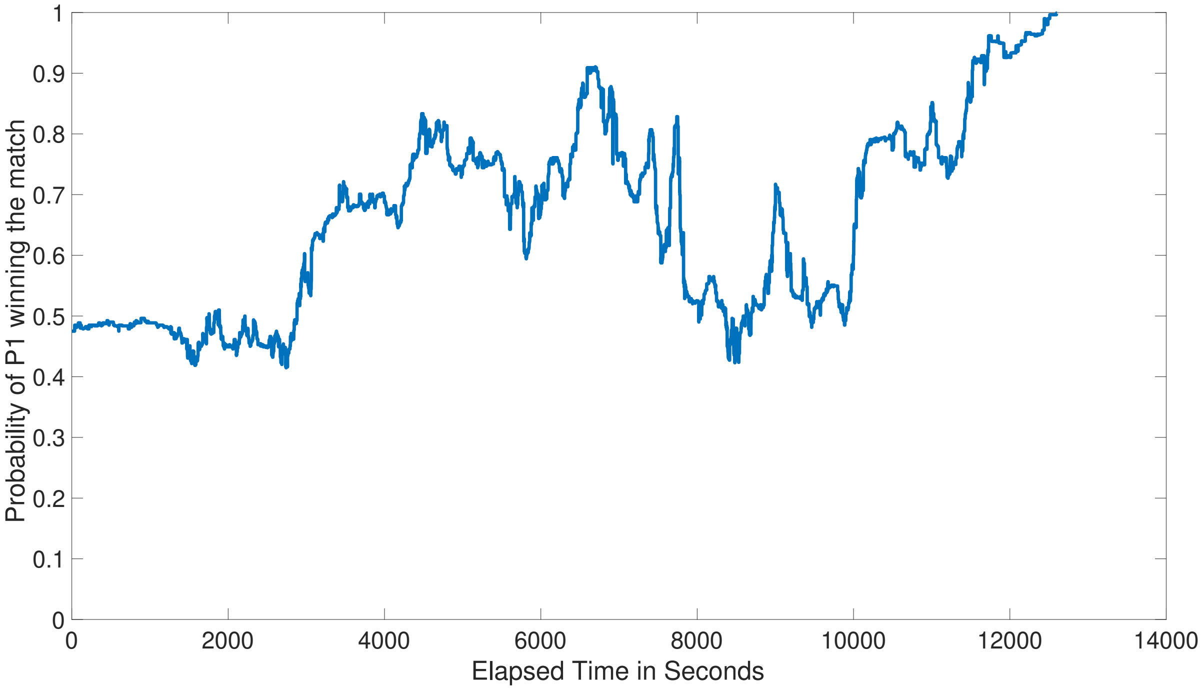
[19] Especially markets for more important matches like finals, half finals or quarterfinals are very accurate. However, those markets may not necessarily be representative samples as they are likely to attract a less sophisticated audience compared to an average match.

Figure 2.7: Example of basic model forecast



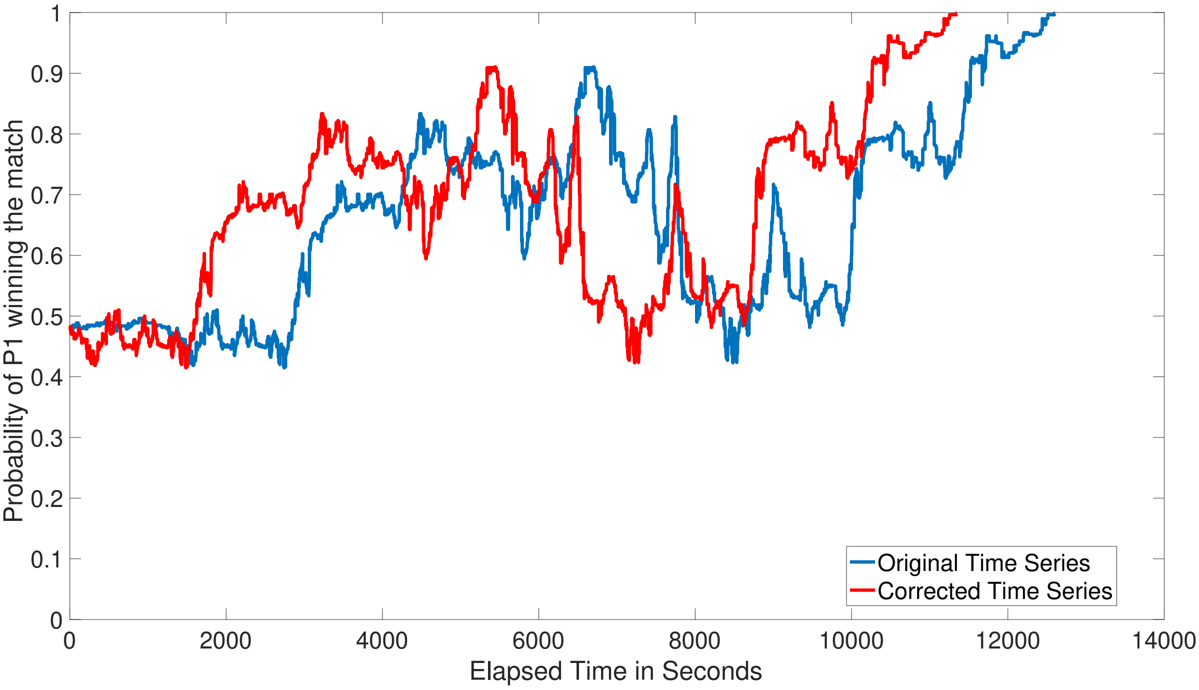
Note: US Open 2012 match between N. Almagro and R. Stepanek

Figure 2.8: Empirical belief stream: unmatched time-series



Note: US Open 2012 match between N. Almagro and R. Stepanek

Figure 2.9: Empirical belief stream: gap between matched and unmatched time-series



Note: US Open 2012 match between N. Almagro and R. Stepanek

point data. I define the optimal starting point t^* as the solution to the following problem. The requirements of the matching criterion are summarized in the box below.

$$t^* = \underset{t}{\operatorname{argmin}} \sqrt{\sum_{i=1}^{n} (M_i - D_i(t))^2}$$

Matching Criterion

The time-series are matched such that the Euclidean metric between the probabilities predicted by the model and the ones calculated from the betting market data are minimized.

To work well, this criterion relies on two major assumptions:

1. At every point during the match, the change in probabilities predicted by the model and the change in probabilities observed in the betting market time-series go in the same direction.
2. By the time of the service, the market has converged to a new equilibrium, i.e. there is no inefficient drift by the time of the service.

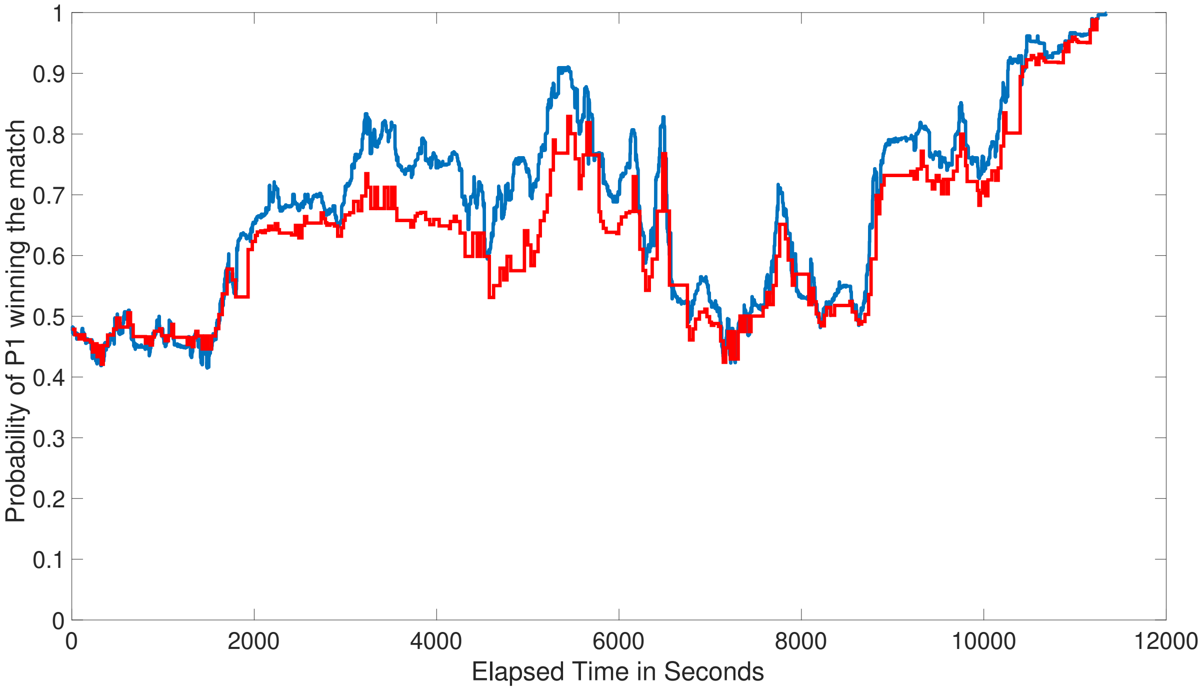
Given that the market is sufficiently liquid at any point of the match, both assumption seem fairly weak. For Assumption 1 to be satisfied, it is sufficient if odds ratios for a player are falling (i.e. the implied probability that he is winning the match is increasing) if he is winning a point. Assumption 2 is slightly more critical. Croxson and Reade (2013) show empirically that Betfair markets do not always adjust immediately. Although I have no means of testing the speed of adjustment with my data, the fact that the break between the scoring of a point and the subsequent service is fairly large on average and the market is sufficiently active during the matches should ensure a timely adjustment of prices. Even if adjustment was slow, the match quality would still be good as long as there is sufficient movement in the prices. This is due to the fact that the "spikes" in match-winning probabilities which are caused by a player winning or losing an important point have a relatively high weight in the matching algorithm and can therefore be captured relatively easily. Notice that persistent deviations in terms of levels between the forecast and the observed time-series are not problematic as these are very unlikely to affect the quality of the matching. Towards the end of a tennis match the probabilities of players winning the match naturally converge to either 0 or 1 such that even for a match with a heavily favored player, there is nothing like a "flat" time series that could severely complicate the matching.[20] Figure 2.10 shows the matching outcome for the example in Figures 2.7 to 2.9.

It is worthwhile to note that, by applying the outlined matching procedure to the data, I avoid another potential problem. When calculating match-winning probabilities from the data, I use the last price matched on the exchange (LPM). The LPM is the price of the last betting agreement. The advantage of using this price measure is that it is closest to the definition of a market equilibrium. However, even though Betfair matches offers on both market sides instantly, the LPM may potentially be a lagged variable. Especially in markets that hit the zero volume boundary (e.g. matches with a low public interest) this could potentially lead to issues caused by slow price adjustment in some matches.[21] As the matching procedure used here removes the price adjustment period, this is not as much of a concern anymore.

[20] As robustness checks, I have also deployed matching only on the second half of the match where changes in match-winning probabilities are larger, used the importance of a point as a weight in the matching function or used higher powers in the matching function, thereby weighting outliers higher. All approaches have led to very similar results.

[21] There are other, faster updating measures like average back or lay prices or averages between back and lay prices. The drawback of those measures is that they either reflect the beliefs of only one market side or impose strong assumptions on supply and demand. Using the average between back and lay prices instead of the LPM yields very similar results.

Figure 2.10: Matching: Model forecast and empirical belief stream



Note: US Open 2012 match between N. Almagro and R. Stepanek

Figure 2.9 displays the corrected betting market time-series for the data depicted in Figure 2.8. It is evident that, in the particular example, substantial adjustments to the starting point of the match need to be made as the market is switched to the active state almost 2,000 seconds before the first point is played.[22]

2.7.3 Empirical Model

Having matched the point data to the betting market data, I proceed by analyzing how Betfair prices react to events during the match. For each point that is played, I use the statistical model derived in Section 2.5 and a functional form for the winning-on-service probabilities to calculate a predicted match-winning probability. This probability is then compared to the one calculated from the betting market data. Using a nonlinear least squares approach, I search for the parameters that are most likely to generate the belief data.

[22] The example here is chosen to show a case that requires a substantial adjustment of the starting point. There are also matches in my dataset that only require slight adjustments or even no adjustments at all.

As I have shown in Section 2.5, the statistical model of tennis play can be rewritten as

$$M(p_{it}, p_{ig}, p_{jg}, S_{it}) = p_{it} M(p_{ig}, p_{jg}, S^{\oplus}_{i,t+1}) + (1 - p_{it}) M(p_{ig}, p_{jg}, S^{\ominus}_{i,t+1})$$

where p_{it} is the probability of player i winning point t, p_{ig} and p_{jg} are the long-term point-winning probabilities of players i and j at the start of game g, $S^{\oplus}_{i,t+1}$ is the state of the match if player i would win point t and $S^{\ominus}_{i,t+1}$ is the state of the match if player i would lose point t. Because the match winning probability $M(\cdot)$ depends on p_{it}, p_{ig} and p_{jg}, it is not invertible. In other words, there is an infinite number of combinations p_{ig}, p_{jg} that would generate the same match-winning probability. I solve this problem by assuming a particular functional form for p_{it}, p_{ig} and p_{jg} that is the same within and across matches. The parametrization is chosen to reflect the point data model estimated in Section 2.6:[23]

$$p_{it} = \Phi\left(\eta + Q_i \beta + H_{it} \delta\right).$$

Alternatively, I could also assume that p_{it} follows a linear probability specification, which allows for an easier interpretation of parameters. In that case, p_{it} would be given by

$$p_{it} = \eta + Q_i \beta + H_{it} \delta.$$

In both cases, I assume that the future expected winning on service probabilities reduce to

$$p_{ig} = \Phi\left(\eta + Q_i \beta + \tilde{H}_{ig} \delta\right) \quad \text{and} \quad p_{jg} = \Phi\left(\eta + Q_j \beta + \tilde{H}_{jg} \delta\right)$$

in case of the Probit formulation and

$$p_{ig} = \eta + Q_i \beta + \tilde{H}_{ig} \delta \quad \text{and} \quad p_{jg} = \eta + Q_j \beta + \tilde{H}_{jg} \delta$$

in case of the linear probability formulation. Here, $\tilde{H}_{ig}$ and $\tilde{H}_{jg}$ represent the subset of regressors in H_{it} and H_{jt} that only contain variables which are updated between games and not points (e.g. the set score difference). Specifically, $\tilde{H}_{ig}$ and $\tilde{H}_{jg}$ are identical to the values of H_{it} and H_{jt} at the first point played in a game where by assumption no information about the previous points is contained in H_{it} and H_{jt}. This is equivalent to

[23] For computational feasibility, the match-level random effect η_i is replaced by a general level effect η that is assumed to be constant across players.

assuming that bettors' expectations about future realizations of the players' winning-on-service probabilities are not a function of the immediate match history up to the current point but may still adjust between service games and sets. Under these assumptions, a point win would allow bettors to adjust their beliefs about the next service in the same game but not on a player's long-term point-winning probability. This is in line with the typical definition of the hot hand being a temporary elevation above a players long-term or average performance while still permitting an update of long-term point-winning probabilities throughout the match, e.g. if one player is ahead or behind in sets. Using the parametric assumptions on winning-on-service probabilities and the statistical model of tennis play, the predicted match-winning probability of player i is given by $\hat{M}(\eta, \beta, \delta, Q_i, H_{it}, \tilde{H}_{ig}, \tilde{H}_{jg}, S_{it})$. To fit the model to the data, I minimize the sum of squared deviations between the model predictions $\hat{M}_{ngt}(\cdot)$ and match-winning probabilities calculated from betting odds M_{ngt} over all matches, $n \in N$, games in the match, $g \in G_n$ and points played during a game $t \in T_g$.

$$(\hat{\eta}, \hat{\beta}, \hat{\delta}) = \arg\min \sum_{n \in N} \sum_{g \in G_n} \sum_{t \in T_g} \left[\hat{M}(\eta, \beta, \delta, Q_p, H_{ngt}, \tilde{H}_{ng}, \tilde{H}_{ng}, S_{ngt}) - M_{ngt}\right]^2$$

Given the parametrization of winning-on-service probabilities, variation in point characteristics and match-winning probabilities calculated from prices identifies the parameters.

2.7.4 Results

Table 2.3 shows estimation results for both model specifications. For a direct comparison to the point data benchmark, Table 2.4 shows both the estimation results from the point data and the betting market model for the linear probability specification.[24] As I have already shown in the preceding section, winning-on-service probabilities can actually be treated as independent and are mainly a function of players' relative strength. As already discussed in Section 2.6, the results confirm previous findings by Klaassen and Magnus (2001) and are perfectly well explained by the fact that, at least in the short run, professional players should not be susceptible to the history of a game. Indeed, the only evidence of momentum that can be found in the point data estimates is a slightly higher chance of winning a point for the player that is ahead in terms of sets and after scoring a point win with the service directly (ace).

[24] A comparison of the Probit specifications yields similar results. The linear probability specifications are discussed here for the benefit of an easier interpretation of the coefficients as direct marginal effects.

Based on the point data estimation results, one would expect that, beyond the change in match winning probabilities directly caused by a point win, prices on the betting market do not show reactions to events like previous point wins, which supposedly create momentum. It is immediately evident from Table 2.3 that, judging based on the point data benchmark, betting market prices show much more movement than expected. Betting market prices react stronger than expected to almost all traceable events during the match, thereby showing a lot more volatility than what would be expected according to the benchmark.

Table 2.3: Betting data estimation results

	Probit link		Linear prob. link	
Last point won	0.140***	(0.019)	0.049***	(0.007)
× ranking point sum	-0.010^{***}	(0.003)	-0.003^{***}	(0.001)
× ranking point diff.	-0.015^{**}	(0.006)	-0.005^{**}	(0.002)
Importance of point	-1.497^{***}	(0.218)	-0.976^{***}	(0.074)
× ranking point sum	0.020	(0.060)	0.112***	(0.016)
× ranking point diff.	-0.217^{***}	(0.033)	-0.067^{***}	(0.011)
Ranking point sum	0.087***	(0.007)	0.004**	(0.002)
Ranking point difference	0.041***	(0.001)	0.013***	(0.000)
Last point ace	-0.324^{***}	(0.031)	-0.115^{***}	(0.011)
Last point double fault	-0.226^{***}	(0.061)	-0.079^{***}	(0.023)
L. p. server unforced error	0.016	(0.016)	0.005	(0.006)
L. p. returner unforced error	-0.013	(0.025)	-0.009	(0.009)
Last point server net point	-0.017	(0.017)	-0.009	(0.006)
Last point returner net point	-0.016	(0.050)	-0.013	(0.017)
Set score difference	-0.013^{***}	(0.002)	-0.008^{***}	(0.000)
× ranking point sum	0.002***	(0.000)	0.002***	(0.000)
× ranking point diff.	0.040***	(0.006)	-0.001	(0.002)
Constant	0.201***	(0.035)	0.682***	(0.012)
Observations	15730		15730	

Standard errors (based on a numerical approximation of the Hessian in the betting market model) in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The results suggest that the motivating example of this analysis, an alleged belief in the hot hand indeed appears to be present in the data. All else given, a player is expected to have a 4.9 percentage points higher chance of scoring the next point on service if has won the last service point. Notably, the estimated coefficients of the interaction terms indicate that the degree of dependence depends on the specific circumstances of the match. If the sum of both players' ranking points is high, i.e. if the match is of higher quality, the hot hand plays a lesser role. In fact, given the estimates, betting market prices in the US Open 2012 final between Murray and Djokovic do not show an effect of the hot hand. The effect of the hot hand also seems to vary significantly with the relative strength of both players. In a given match, the lower ranked player is considered to be more susceptible to the hot hand. It should be noted that there are two potential effects at play which are not easily separable. It is possible that bettors in general tend to overestimate the degree to which the lower ranked player is affected by momentum or that matches with a high rank difference attract a different bettor population. These effects cannot be separated without additional information on bettor characteristics, even though it seems unlikely that matches with a high rank difference, i.e. matches with low public interest, systematically attract less sophisticated bettors.

Table 2.4: Point data estimation vs. betting data estimation

	Point Data Model		Betting Market Model	
Last point won	0.020	(0.013)	0.049***	(0.007)
× ranking point sum	−0.001	(0.002)	−0.003***	(0.001)
× ranking point diff.	0.001	(0.002)	−0.005**	(0.002)
Importance of point	−0.228	(0.155)	−0.976***	(0.074)
× ranking point sum	0.027	(0.030)	0.112***	(0.016)
× ranking point diff.	0.018	(0.039)	−0.067***	(0.011)
Ranking point sum	−0.002	(0.002)	0.004**	(0.002)
Ranking point difference	0.009***	(0.002)	0.013***	(0.000)
Last point ace	0.030*	(0.017)	−0.115***	(0.011)
Last point double fault	−0.012	(0.022)	−0.079***	(0.023)
L. p. server unforced error	−0.014	(0.013)	0.005	(0.006)
L. p. returner unforced error	−0.029**	(0.013)	−0.009	(0.009)
Last point server net point	−0.013	(0.012)	−0.009	(0.006)
Last point returner net point	0.006	(0.017)	−0.013	(0.017)
Set score difference	0.013**	(0.006)	−0.008***	(0.000)
× ranking point sum	0.001	(0.001)	0.002***	(0.000)
× ranking point diff.	−0.001	(0.001)	−0.001	(0.002)
Constant	0.646***	(0.012)	0.682***	(0.012)
Observations	15730		15730	

Standard errors (based on a numerical approximation of the Hessian in the betting market model) in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Interestingly, betting market prices do not only show higher volatility, but also seem to overvalue the serving player compared to the point data benchmark. The unconditional winning-on-service probability of a player is about 3.6 percentage points higher when estimated from the betting market rather than the point data, which indicates that bettors seem to generally overestimate the point scoring probability of players. Indeed, bettors seem to believe that the winning-on-service probability of a player is even higher for good players (relative to their opponent) and in higher quality matches, i.e. matches where both players are very skilled relative to the rest of the field. Although the former is in principle supported by the point data estimates, the effect is larger in the betting market model compared to the point data model. This can be interpreted as evidence against the favorite-longshot bias as bettors seem to systematically overestimate the winning-on-service probability of better players. The favorite-longshot bias is a phenomenon that is sometimes observed in fixed-odds betting markets. It refers to a situation where betting odds are systematically biased against favorites and towards “longshots”, i.e. teams or players with a relatively low ex-ante probability of winning. While betting odds are too high for favorites, they are too low for longshots, implying that bettors systematically underestimate the winning probability of favorites while overestimating the winning probability of underdogs. The favorite-longshot bias has been documented extensively in the literature, although it’s presence seems to somewhat depend on the particular markets and even the characteristics of the match.[25]

There is also some significant evidence of betting against well performing players. In particular, bettors seem to underestimate the probability of winning on service for the player that is ahead in terms of sets, although this player should be more likely to score a point according to the point data estimates. Thus, although there is no evidence of a favorite-longshot bias across matches, within a given match and across sets bettors seem underestimate the leading player’s chance of winning. As discussed above, on a very short time horizon, the market again tends to favor the player who is scoring points on service. However, while bettors do not seem to believe that relatively strong players benefit from scoring points on service, there is significant evidence that they seem to believe that relatively weak players are building up momentum by scoring points. Consequently, this can be interpreted as evidence of a conditional short-run version of the favorite-longshot bias, where the longshot’s probability of winning is only overestimated if he send a “good” signal by scoring a point.

[25] Thaler and Ziemba (1988), Sauer et al. (1998) or Snowberg and Wolfers (2007) provide more extensive surveys of the literature on the favorite-longshot bias and Vaughan Williams and Paton (1998) discusses why the favorite-longshot bias might be reversed in some markets. The favorite-longshot bias has been found very consistently in horse race betting and other sports, for instance by Snowberg and Wolfers (2010) for horse race betting, by Cain et al. (2003) for a number of other sports betting markets and by Forrest and McHale (2007) and Lahvicka (2014) for tennis specifically. Lahvicka (2014) concludes that the favorite-longshot bias appears more prevalent in tennis matches between lower ranked players and in later round matches. It should be noted that these studies usually analyze fixed-odds betting markets where prices are set by bookmakers.

The preceding discussion highlights the importance of the time-horizon with respect to belief formation and betting behavior. In fact, my results show that well-established behavioral phenomena like the favorite-longshot bias frequently observed in betting markets may have multiple dimensions. In my setting, bettors seem to generally overestimate a favorite's ability to score points on service but start to believe in the longshot (underdog) whenever he performs well, i.e. scores a point. A potential explanation for these two, at first sight contrary, observations might be the time horizon of information gathering and decision making. While there is enough time to evaluate player's chances of winning between matches, the time between two points is rather short and hence bettors may tend to stick to rules of thumb or their intuition. Again, it is worth noting that empirically almost no evidence of momentum-based performance can be found in the point data. Even for relatively low skilled professional players[26] there is no sign of bad plays affecting their future performance.

The observation that bettors may over-infer also translates to other events during a match. Most notably, bettors seem to believe that players struggle to score at crucial points during the match when a lot is at stake.[27] Bettors seem to believe that, the more important a point is for the match outcome, e.g. if it is a game, set or even a match point, the less likely is the current server to score that point. Again, this is less pronounced in high quality matches. Strikingly, bettors also seem to believe that relatively lower skilled players tend to do better on important points than relatively higher skilled players. Similar to the belief in a hot hand, at important points during the match, the market seems to put too much weight on the longshot while underestimating the favorite's probability of winning.

2.8 Conclusion

Using high-frequency betting market data as well as detailed point data, I have provided evidence that bettors on an online betting exchange over-infer from certain events during tennis matches. In line with previous research (e.g. Klaassen and Magnus, 2001, 2003), I have shown that modeling tennis matches as random processes is indeed a good approximation. During a match, the future performance of professional tennis players

[26] It is also worth noting that the US Open is a high stakes tournament. Hence, the average skill level of players is certainly very high. Consequently, the results may not apply to a low stakes professional or even amateur environment.

[27] An alternative and somewhat complementary explanation would be that opponents shine at defending these points.

is, independent from their skill level, not affected by their current performance. There is almost no empirical evidence that points towards the existence of “momentum” that players can “build up” in those matches. The probability of winning a point on service is almost constant throughout a match.

In stark contrast, winning-on-service probabilities implied by prices on the betting market do not seem to follow a random process but specific patterns. Overall, the betting market model provides clear evidence that prices on Betfair are not only driven by the belief in a hot hand but more generally show more volatility than warranted based on the fundamentals. The point data estimates show that treating winning-on-service probabilities as independent is indeed a good approximation. Empirically, the probability of winning a point is mainly determined by the relative strength of players, but is not changing significantly during the match. In stark contrast, the betting data estimates show that prices are moving a lot throughout the match. Bettors seem to infer a lot more from certain events during the match than is justifiable based on the player’s performance. Interestingly, while they generally seem to slightly overestimate the favorite’s probability of winning, at important points during the match or after seeing the underdog score a point on service, they start overvaluing the ex-ante unfavored player’s chances. This can be interpreted as a conditional and dynamic version of the favorite-longshot bias, which has been discussed extensively in the betting market literature. While bettors do not seem to generally overestimate the longshot’s probability of winning, they do so only if the longshot sends a positive signal by scoring a point on service or when a lot is at stake during the match. This casts doubt on the wide-spread assumption that betting markets, much like financial markets, are generally efficient if they are sufficiently liquid.

Although I believe that this chapter takes an important step towards a better understanding of high frequency betting markets, which are structurally very similar to financial markets, there is still a lot to be done. I have not investigated the role of market liquidity and its relationship with betting market prices. Market liquidity tends to vary enormously, both across and during matches mostly based on the players’ relative and absolute strength. One could expect less liquid markets to show larger or smaller deviations from the point data benchmark, depending on whether the bettor population they attract is more or less biased compared to the average match. Another drawback of the model is that it is computationally very demanding. This limits the number of covariates and interaction terms I am able to control for and prevents the use of more advanced estimation mechanisms. Optimizing the estimation routine would not only allow to improve the robustness of the estimation, but also to increase the sample size.

Finally, it seems worthwhile to validate my findings in different settings. Prediction markets do not only cover sports events but also political campaigns or even scientific studies[28]. The general consensus is that probabilities implied by prices on those markets contain all publicly available information and therefore prices are efficient. My results cast some doubt on this view and it would be interesting to see whether they translate to other, potentially simpler markets. One point that should not be underestimated is that tennis has a relatively complex structure. The marginal value of scoring a point is not constant throughout a match and depends heavily on the state a the game. Thus, evaluating the impact of a scored point on the overall match-winning probability is a non-trivial task – especially in a highly active and dynamic market environment. Making accurate predictions in these situations requires either a lot of experience or sophisticated computational methods, both of which the majority of bettors may not have access to. In contrast, evaluating the probability of a football team winning the game after seeing a goal or updating the probability of a candidate winning an election after seeing a debate is structurally a much simpler task. On the other hand, albeit hard to evaluate the impact of scoring a point on the probability of winning the match is technically very straightforward. Tennis matches are very close to random processes as the updating step after seeing one player score a point is mainly a function of the state of the match and players' relative strength. Analyzing the impact of different degrees and dimensions of complexity on the potential bias of betting market prices certainly appears to be a very interesting direction for future research.

[28] See e.g. the Experimental Economics Replication Project (https://experimentaleconreplications.com, accessed on 28/06/2022).

Chapter 3

Does the “Hot Hand” Vanish With Higher Skill and Experience? Evidence From Winning Streaks in Online Games[*]

3.1 Introduction

Since the seminal paper by Gilovich et al. (1985) the “hot hand fallacy” has been a well documented phenomenon, which has been used as a candidate explanation for puzzles and observed behaviors in the areas of economics and finance. The hot hand describes a state in which an agent (most often a player) temporarily achieves a higher performance due to past successes. This pattern, which is also commonly referred to as “streakiness” or “momentum”, has been extensively studied in different contexts, predominantly related to various types of sports.[1] Apart from a more general interest in playing behavior and delivery of performance, a natural question is if there is a discrepancy between the degree to which such a hot hand effect can be observed in actual player performance and the degree to which a belief exists that such a hot hand effect is present.

[*] I am grateful to Sebastian Panthöfer and Katharina Momsen for their valuable comments and suggestions.

[1] The seminal paper by Gilovich et al. (1985) and much of the literature spawned by it focusses on basketball but examples of other sports include Klaassen and Magnus (2001) for Tennis matches or Green and Zwiebel (2018) for baseball matches.

In this context, the terminology "hot hand fallacy" refers to the alleged cognitive misperception that such a hot hand state is driving player performance even if it is in fact not. Such beliefs in a hot hand or streakiness are most often encountered in sports betting and gambling markets but are also frequently discussed in relation to financial markets, for instance in the context of traders who are presumed to over-infer from recent information to predict the future performance of stocks or market trends.[2] In Chapter 2 I have already provided evidence that prices on an online betting exchange exhibit signs of a belief in the hot hand even though the performance data from respective professional tennis matches does not show any signs of such a pattern.

The presence or lack of such a hot hand effect has been documented by researchers in a variety of settings.[3] Although Bar-Eli et al. (2006) note that "the scientific support for the hot hand is controversial and fairly limited", they conclude that the "question whether the hot hand phenomenon does or does not exist remains for the meantime unresolved". In fact, a hot hand effect may be detected in some settings or under specific circumstances while it remains absent in others. In accordance with this, Koehler and Conley (2003) conclude that "no single study can be the last word on this topic". Indeed, what is largely missing from the literature is a discussion of potential heterogeneity in the hot hand effect and the factors driving it. In line with this Bar-Eli et al. (2006) note that the existence of a hot hand effects may very well hinge on factors like the "nature of the task performed, the level of expertise, or some psychological (or emotional) variable". Indeed, most of the evidence on the hot hand comes from professional sports, where players are typically very familiar with the setting they are playing in. In such an environment one may not expect that player performance exhibits extensive streakiness or at least one would not expect much heterogeneity in the hot hand if players' experience and skill levels are in a relatively narrow range relative to the general population (which would be expected in competitive sports).

However, this does not imply that streakiness is a delusion. The results presented in this chapter indicate that heterogeneity may be a key factor to studying hot hand effects and that skill and experience are key drivers of the hot hand effect and different

[2] See e.g. Rabin (2002), Rabin and Vayanos (2010) or the discussion in DellaVigna (2009). Rabin and Vayanos (2010) also establishes a link between the gambler's fallacy, the (false) belief in systematic reversals in random sequences. The hot hand fallacy, which describes the belief in systematic persistence rather than systematic reversal, can be a consequence of the gambler's fallacy when the underlying data generating process is not known. For an independently and identically distributed (i.i.d.) data generating process (e.g. the return to a stock), agents may then form beliefs that the underlying process is in fact not i.i.d. and exhibits streakiness.

[3] See Bar-Eli et al. (2006) for a comprehensive summary of the literature.

degrees of the hot hand may exist in parallel – even in the same setting. To study the effects of skill and experience on the hot hand effect, I use a novel panel dataset on match outcomes of the online video game Dota 2. This dataset allows me to draw comprehensive inference on the distribution of hot hand effects across a wide population of varying skill and experience level. Importantly, the range of skill and experience levels I observe in the data is much wider compared to studies that focus on professional sports. Beyond that, Dota 2 also inhibits many desirable features of a controlled setting that simplify the identification of a hot hand effect:

1. Particularly in sports settings, which the literature has been very focused on, studies of the hot hand are frequently plagued by the problem of endogenous response. Endogenous responses refer to situations where not only a player's own performance changes in response to recent successes but also the behavior of other players interacting with him. As an example take the basketball setting studies by Gilovich et al. (1985). When a shooter starts scoring frequently, it is likely that the opposing team will shift focus and defending resources from other players to him.[4] This is not an issue in my setting as players are randomly matched up with other players such that there is only a very low probability that someone else in the current match has observed a player's previous performance and alters her playing behavior towards him.

2. In Dota 2 the probability of success is a function of the player's performance but also of the performance of multiple other players. Therefore, from the perspective of the player, the outcome of the match is inherently random. This should be known to players such that they should expect that their successes are the result of an i.i.d process if their performance is consistent.

3. The expected long-term success rate of players around 50% such that wins and losses are equally likely and thus winning and losing streaks are equally likely to occur.

Observing the full match history of all players as well as the skill rank on which they are playing, I can analyze the presence of a hot hand effect depending on their current skill level and match experience. I then show that the hot hand effect differs substantially across the population of players. I find that there is an over 5 percentage point higher winning chance following a previous win rather than a loss for unexperienced players

[4] For an extensive discussion about the problem of endogenous responses in Basketball and other sports see e.g. Green and Zwiebel (2018).

playing on medium-low to medium-high skill level. For more experienced players this hot hand effect decreases to below 1 percentage point. This demonstrates the substantial impact of the environment on the estimation of a hot hand effect. More experienced and higher skilled players exhibit only a weak hot hand effect, while the performance of less experienced players is much more prone to streakiness. These findings are robust to changes in definition of the hot hand state, a distinction between the hot and cold hand effect, considering individual player statistics as outcome variables or only allowing for short breaks between matches. While many of these robustness checks lead to an estimation of higher baseline hot hand effects, the pattern that is robust among them is the stark reduction of the hot hand with higher game experience and a strong dependency on a player's skill level.

The remainder of the chapter is organized as follows. Section 3.2 discusses the related literature on the hot hand effect. An overview of the dataset and a short explanation of the game mechanics is provided in Section 3.3.[5] The empirical strategy used to identify the hot hand effect and its interaction with skill and experience level are discussed in Section 3.4. Descriptive results are presented in Section 3.5 while Section 3.6 discusses the regression results and the robustness checks considered. Section 3.7 concludes.

3.2 Literature

More than 30 years after the seminal paper by Gilovich et al. (1985) and despite a multitude of empirical studies, the hot hand remains a highly debated and controversial topic.[6] In their seminal paper, Gilovich et al. (1985) first documented that a hot hand effect, contrary to the popular belief by experts as well as amateur audiences, is indeed not detectable in professional basketball. Gilovich et al. (1985)[7] have investigated both actual basketball results from the Philadelphia 76er and the Boston Celtics NBA teams as well as results from controlled shooting experiments with collegiate players and concluded that the belief in the hot hand and the detection of streaks in random sequences is attributed to a general misconception of chance according to which even short random sequences are thought to be highly representative of their generating process.. While they find that coaches and players as well as audiences believe in a hot hand, the data does not support this presumption.

[5] A more detailed description of the game mechanics is relegated to the appendix.

[6] See Bar-Eli et al. (2006) for a comprehensive review of the early literature on the hot hand effect.

[7] Summaries and discussions of this research have also been published in Tversky and Gilovich (1989b) and Tversky and Gilovich (1989a).

The results of Gilovich et al. (1985) have been replicated, challenged and confirmed numerous times in the same or other settings. Koehler and Conley (2003) and Avugos et al. (2013) confirm the findings by Gilovich et al. (1985) and find no evidence of a hot hand effect analyzing outcomes of the NBA "Three Point Contest" (Koehler and Conley (2003)) or controlled shooting experiments (Avugos et al. (2013)). More recently, Miller and Sanjurjo (2018a) have casted doubt on these results by showing that there is a strong hot hand effect using data from a controlled shooting experiment with semi-professional basketball players and applying an improved statistical approach that corrects downward biases in the analysis conducted by Gilovich et al. (1985) and Koehler and Conley (2003).[8] Miller and Sanjurjo (2021) extend these findings to the NBA "Three Point Contest" studied by Koehler and Conley (2003).

While the focus of Miller and Sanjurjo (2018a) and Miller and Sanjurjo (2021) is in principle a different one, they point out a strong degree of heterogeneity in the individual hot hand effects but do not further investigate if these individual estimates of the hot hand effect correlate with information about the shooters. Their results do however illustrate the importance of accounting for heterogeneity in the hot hand effect rather than assuming that an average hot hand effect applies to the entire population. In this chapter, I show that heterogeneity in the hot hand effect is at least partially attributable to differences in player characteristics.[9]

The hot hand has also been investigated in a number of settings other than basketball. In Chapter 2 of this dissertation I have shown that, in line with earlier results by Klaassen and Magnus (2001), there is no robust evidence for a hot hand effect in professional tennis using data from the US Open 2012 but have also shown that there is evidence of a belief in the hot hand in sports betting data.[10] Green and Zwiebel (2018) have found robust evidence of hot hand effects in professional baseball when accounting for endogenous responses (e.g. if players on streaks are defended more fiercely).[11] Further studies have

[8] See also Miller and Sanjurjo (2018b) for of a discussion of the bias in Gilovich et al. (1985) and subsequent studies. The bias they point out is particularly prevalent in studies with low statistical power. These concerns do not apply to the same extend to the dataset analyzed in this chapter.

[9] This is not to say that there is no remaining unexplained heterogeneity in hot hand effects across players.

[10] That outside spectators and experts hold hot hand beliefs has already been documented by Camerer (1989) in the context of basketball. See Chapter 2 for a discussion of the relevant literature.

[11] In particular, they point out a concerns that is of relevance for many hot hand studies in the context of team sports. If a player is on a streak, then the opposing team may shift their focus to defend this particular player more fiercely. If this hampers the hot player's ability to score, the hot hand effect will be underestimated if the endogenous response is not taken into account. With the data and setting I am using this effect is not a concern. In Dota 2, players are matched randomly with

also been conducted for bowling (e.g. Dorsey-Palmateer and Smith, 2004), golf (e.g. Clark, 2003a,b, 2005; Connolly and Rendleman Jr, 2008), volleyball (Raab, 2002) or horseshoe pitching (Smith, 2003).[12]

To the best of my knowledge, no studies exist that formally evaluate the existence of a hot hand in online gaming. Kou et al. (2018) collects a series of quotes and player reactions related to streaks in League of Legends, a game that is similar to Dota 2. They list various quotes of very diverse anecdotal evidence related to what players thoughts about reasons of streakiness, advice to break streaks, relation to game mechanics and other factors. Some of quotes also touch on the topic of the interconnectedness between successes and performance.[13]

3.3 Data

Dota 2 is a so called multiplayer online battle arena game ("MOBA") in which two teams (called "Radiant" or "Dire" depending on the side of the map on which they are playing) are fighting each other to take control over a map. Dota 2 is an ideal setting to study the effect of skill and experience on the hot hand because players performance histories and outcomes are easily tractable and allow to precisely calculate both the skill level on which the player is playing as well as her experience with the game measured by the matches she has played. A more detailed description of the game mechanics can be found in Appendix 3.8.

The data on Dota 2 matches was collected via Valve's official Steam Web API service.[14] For each match, information on the players participating in the match, the winning team, start time and duration of the match, game mode, map status at the end of the match, the hero played and items used by each player as well as different

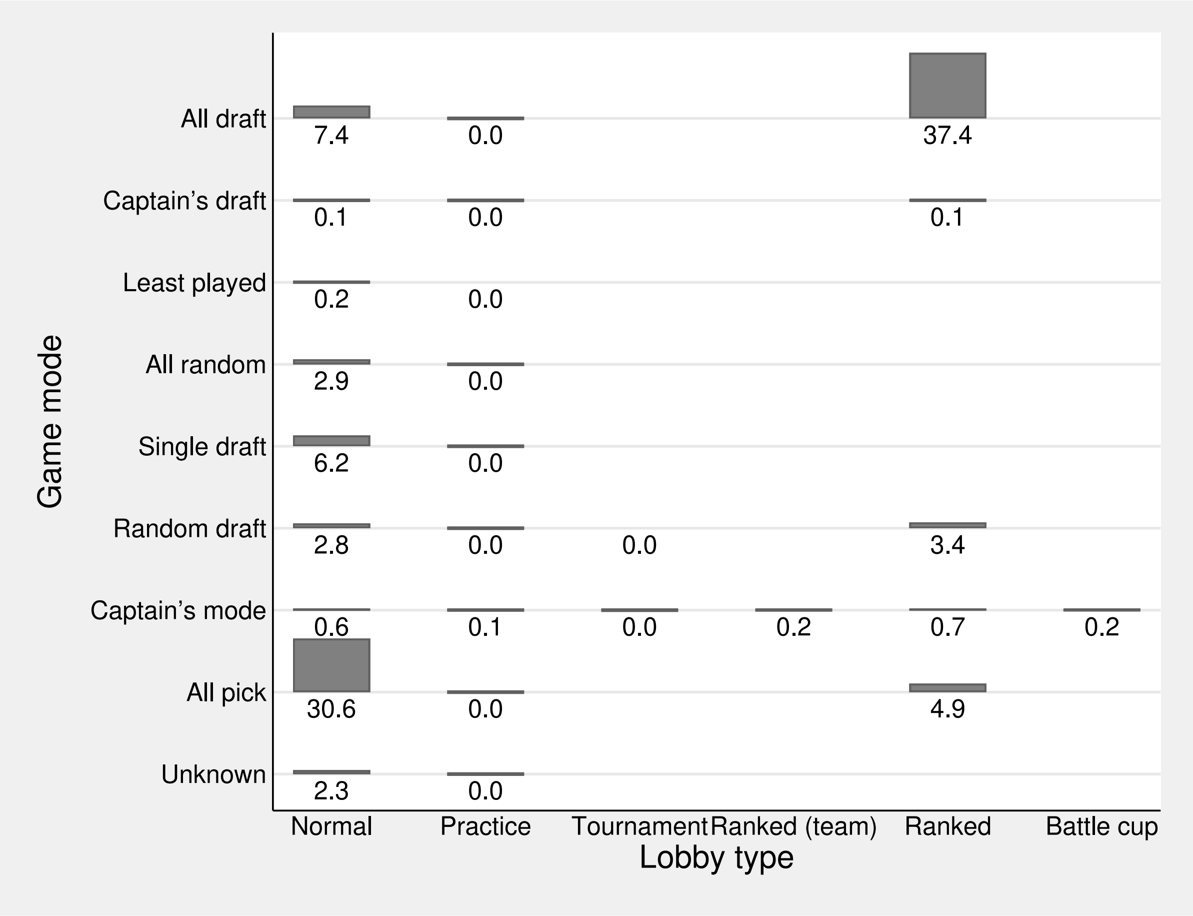
players of the opposing team such that there is only a very low likelihood of repeated interaction and thus it is unlikely that players on the opposing team would have observed recent performances of the hot player.

[12] See Bar-Eli et al. (2006), Table 1 for a comprehensive overview.

[13] Examples include: "You win a game you feel happy... You play well one game you're likely to play well the next. You're directly making your games correlated which makes these streaks much more likely."; "I was P3 at 87 LP, dropped to G2 with 13 consecutive losses. I know I am playing angry so I am not playing as well as usual..."; "It's quite random but it is entirely possible to lose subsequent games that you play in a short amount of time but mostly because your attitude changes and carries over from game to game." or "This is called tilt. It is a disease that no one knows how to cure. Every professional player gets it as well. :P just comes and goes".

[14] The endpoint used to collect the data was api.steampowered.com/IDOTA2Match_570.

Figure 3.1: Percentages of game mode and lobby type combinations



performance statistics[15] for each player at the end of the match was collected. For the following analysis a random sample of 41089 users was selected from the data gathered.

This data was supplemented by information on players' match histories from the OpenDota API service.[16] A match history contains each Dota 2 match ever played by a specific player, including key match information like game mode, match outcome, hero played as well as match performance indicators. In addition, OpenDota also collects data on the skill bracket in which the match was played. The skill bracket is assigned on the match level by the game developer Valve and differentiates between *normal*, *high*

[15] For each player this included the typical Dota 2 performance statistics: kills, deaths, assists, denies and last hits during the match as well as gold and hero experience per minute and the hero level achieved at the end of the match.

[16] OpenDota (https://api.opendota.com/api/) is a third-party service which collects, processes and stores large amounts of data from the official Steam Web API.

and *very high* skill. This indicator therefore allows an assessment of the average skill level of each player at each point in time.[17]

Dota 2 features various different modes in which the game can be played. Generally, the ruleset of matches can be varied in the following two ways. The so called *game mode* determines how the game is played. In most modes, this alters how heroes are picked or which heroes are available but may also further affect the ruleset in certain modes. The *lobby type* on the other hand determines the set of other players a player is matched up against.

Both game mode as well as lobby type may affect playing behavior. For instance, players may use the *practice* or *normal* lobbies to try to improve their playing performance with certain heroes or practice new strategies, which in turn may lead to a suboptimal performance and lower than usual winning chances. Similarly, the restrictions on hero choices in some game modes may result in players getting assigned heroes they are not comfortable with. For the analysis below, the data was therefore restricted to only cover matches played in the game mode *all draft* with lobby type *ranked.* As evident from Figure 3.1, ranked/all draft was the most prevalent combination of game mode and lobby type in the data extracted. Besides ranked/all draft being the most popular way to play the game, it also resembles a competitive environment and therefore is most likely to incentivize optimal play. Winning ranked matches brings players further up the leaderboards, which is also visible to others in their player profiles. The all draft mode enables players to both pick from their pool of preferred heroes and also prevent heroes being chosen against which they feel uncomfortable (so-called “banning”).

3.4 Empirical Strategy

In this section, I will discuss the empirical strategy to estimate the hot hand effect and identify the interactions with player skill and player experience.

3.4.1 Defining a Match Series

The hot hand or “streakiness” is typically defined as a state in which players perform above their normal level. Gilovich et al. (1985) note that “(t)hese phrases express a belief that the performance of a player during a particular period is significantly better than expected on the basis of the player’s overall record.”. According to this notion of a

[17] The skill bracket is determined by the average matchmaking ranking (MMR) of the players in the match.

temporary elevation of performance, I will only consider sequences of matches that are not interrupted by long breaks.

The selection of an appropriate break time is not straightforward and also subject to a trade-off. With over 40 minutes on average in the dataset used to produce the results for this chapter, Dota 2 matches are relatively long. Very long streaks with very short breaks are therefore rare in the data – most likely as they are just too time consuming. While this is not necessarily an issue when conditioning only on the previous match, it makes more difficult to understand the impact of longer streaks which may exhibit much stronger hot hand effects as discussed in Section 3.6.4. To enable such an analysis while still preserving enough statistical power, I will consider breaks between matches of up to one day sufficiently short to not interrupt a series of matches.

It should be noted that this assumption will lead to a conservative estimate of the hot hand effect if players are able to "cool down" during longer breaks and leave the "hot" state. To get a sense of the degree to which the hot hand effect varies with break length, I present results using much stricter assumptions on maximum break length in Section 3.6.5. A match series, S is therefore defined as a sequence of matches with breaks up to 15 minutes.

Match series

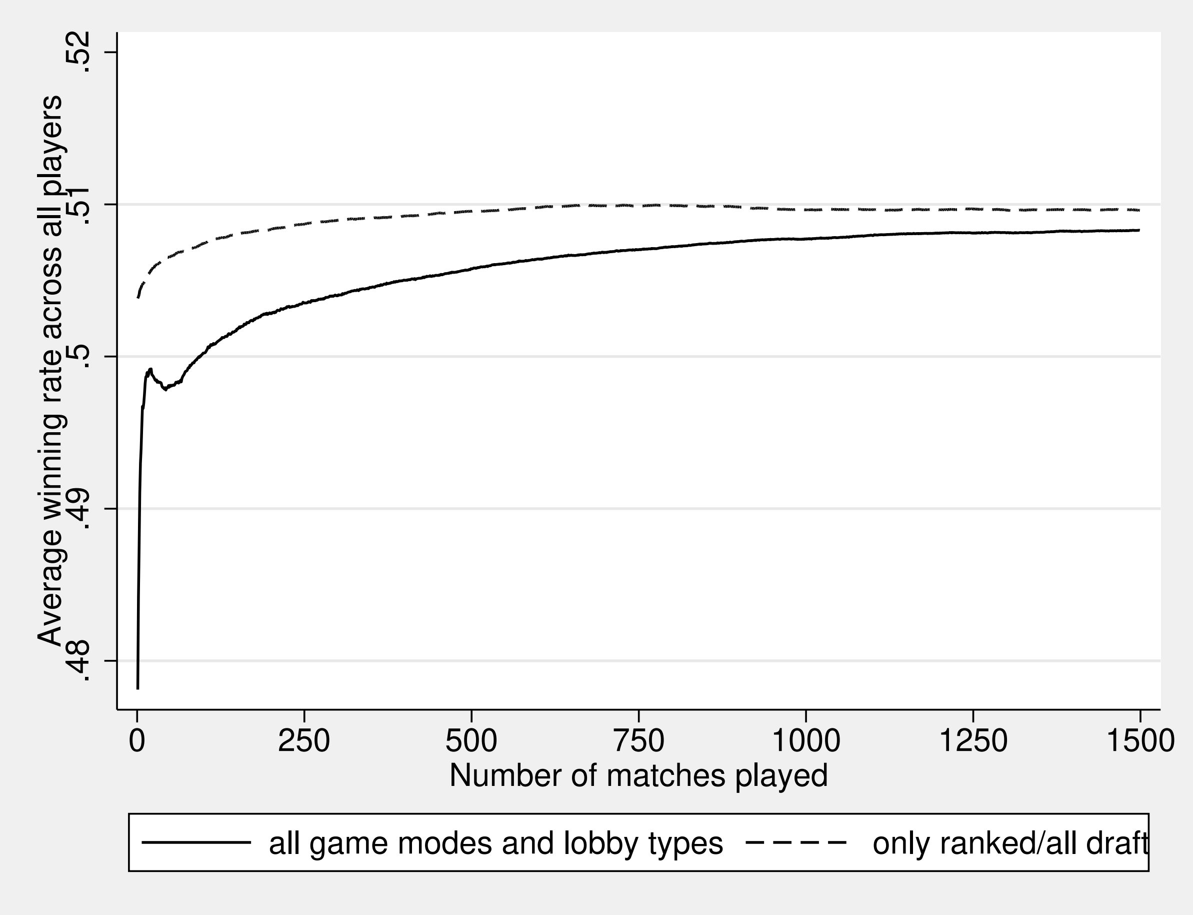
Matches belong to the same match series if breaks between all of the matches are not too long. The break between two matches is considered to be not too long if it is less or equal to a day.

3.4.2 Match Experience

I measure experience with the number of matches a player has played up to a certain point. As players play more matches, they should get to know the game better and familiarize themselves better with the environment, which should ultimately translate to a better performance and higher winning rates as they take better decisions. Similarly, learning about the game may not only translate to higher average winning rates but also more consistent performance, i.e. a lower hot hand effect.

To produce an unbiased estimate of the hot hand effect, one may still want to disregard the first matches played by a player - even when directly controlling for experience. With very little prior experience with the game, additionally to players not having reached their long-term performance level, the matchmaking algorithm may also need to learn about their true skill level to appropriately match players to equally skilled opponents. Moreover, learning may be very heterogenous over the first matches as some players may be more familiar with the concepts of the game if they have played similar games before or may simply learn certain aspects of the game faster. This could also affect the consistency of their performance, i.e. the measured hot hand effect. That learning is indeed taking place is shown in Figure 3.2 below. Figure 3.2 displays average winning rates of players over the course of their entire match history both for all matches as well as matches only in ranked/all draft mode. When all game modes and lobby types are considered, players' average winning rates are gradually increasing over time as can be seen from the solid line in Figure 3.2. When only ranked/all draft matches are considered (dashed line in Figure 3.2), such a stark increase in winning rate is only visible for the very first matches played.

Figure 3.2: Change in average winning rate with number of matches played



There may be different explanations for these observation. Firstly, ranked/all draft is certainly a much more competitive environment than some other game modes which may incentive a much faster learning as well as less experimentation with different playing styles to optimize performance. Secondly, it should be stated that players will have to play at least 100 hours in other game modes before having access to the ranked game mode such that they will always have a acquired a certain amount of experience with the game before starting in ranked mode. For the latter reason it should also not be necessary to trim the sample used in the chapter, which focusses on the ranked/all draft game mode, as players will always have acquired a certain amount of prior experience before entering the game mode.

3.4.3 Controlling for Player Skill

As discussed above, players playing on high skill levels could be less susceptible to a hot hand effect, potentially as they may show a more professional attitude or are more trusting in their on ability and performance. For the internal DOTA matchmaking algorithm, player skill level is measured by the MMR, which directly determines the skill level of a player. However, the MMR but is not readily available in the extracted dataset such that it cannot be used to measure player skill directly in the estimation.

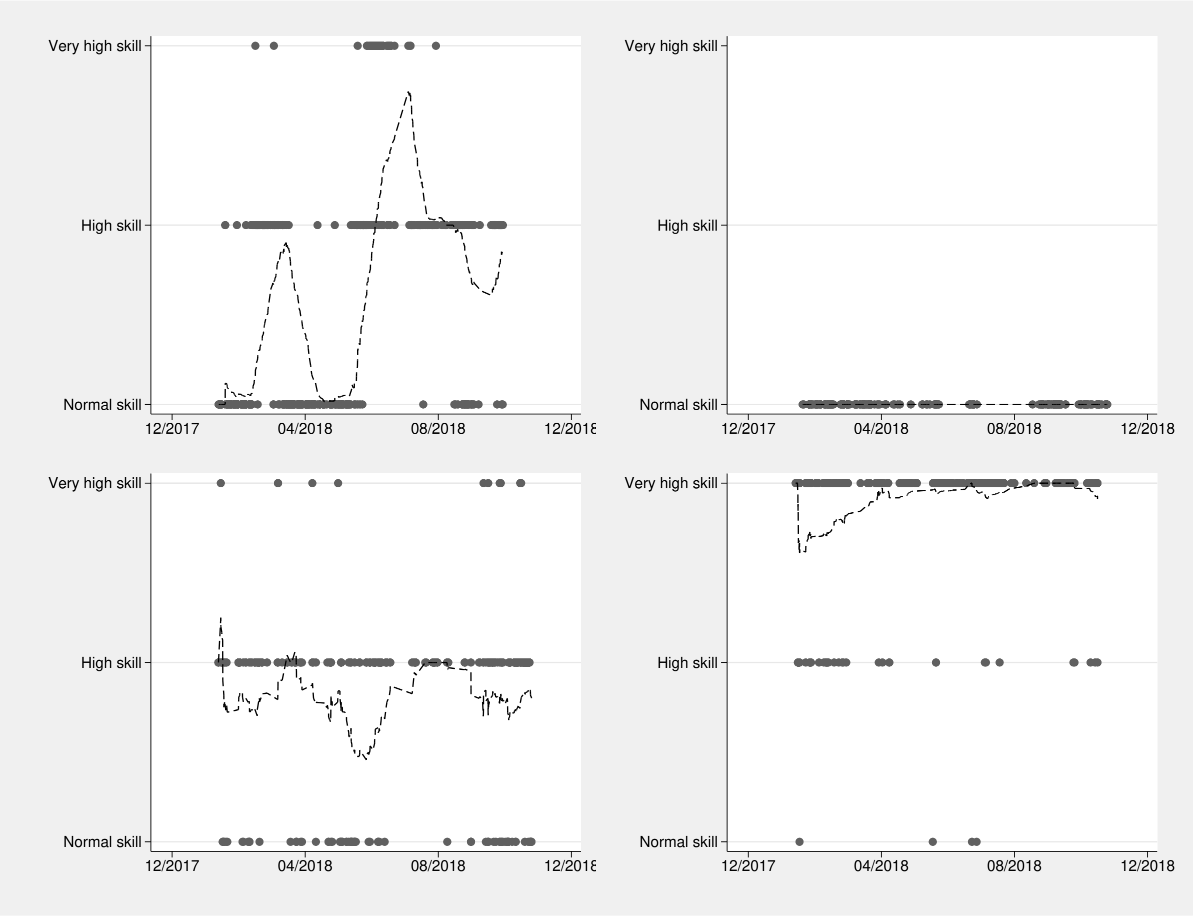
An alternative approximation to player skill is the skill level on which a match is played. This metric separates matches into the bins “normal skill”, “high skill” and “very high skill” and is available for approximately 60% of matches played by a player.[18] As this is a match-level and not a player-level statistic, I construct a statistic for players' approximate skill level as a rolling average of skill levels of matches played by them over the 30 days before and after the current match. Let M_{it} be the set of matches played by player i at day t, let $m_{jt} \in M_{it}$ be match number j played at day t and let $l_{jt} \in \{1, 2, 3\}$ the corresponding skill level of match m_{jt}, where 1 corresponds to “normal skill”, 2 corresponds to “high skill” and 3 corresponds to “very high skill”. The average skill level of player i at day t, s_{it}, is then calculated as

$$s_{it} = \frac{\sum_{\tau=T_0}^{T_1} \sum_{m_{j,t+\tau} \in M_{i,t+\tau}} s_{j,t+\tau}}{(T_1 - T_0) \sum_{\tau=T_0}^{T_1} n(M_{i,t+\tau})} \tag{3.1}$$

where $T_0 = -30$ and $T_1 = 30$ days is chosen for the analysis below.

[18] This information is not available for all matches as it can only be extracted for a very limited time period after the match has been finished. Accordingly, the extraction and recording scripts run by OpenDota, from which this data is sourced, may not be able to record this information for some matches. As is recording does not discriminate by other match parameters, whether or not skill level for a match recorded is inherently random and therefore no selection effect is expected.

Figure 3.3: Examples of evolution of skill levels



Constructing a statistic for player skill level as outlined above has several advantages. Firstly, it allows to construct a relatively noise-free measure of player skill. As the categorization of matches into skill brackets is based on the skill levels of all players, the match skill level may not perfectly reflect the skill level of each individual player. Averaging over a series of matches should result in a more representative measure of skill level as players are not systematically matched with better or worse players. Secondly, it allows to also preserve match observations where data on the match skill level is missing. As discussed above, the match skill level is sometimes not recorded for non-systematic technical reasons. It can therefore be presumed that the average skill level of matches played in close temporal proximity will be reflective of the average skill level when match skill data is missing.[19] Thirdly, using a moving averages allows the skill level of a player

[19] It should be noted that in some cases player skill level may still not be calculated in this way when all match skill information is missing for a player's matches in the interval $t \in [T_0, T_1]$

to still adjust over time when her individual skill level changes - e.g. with more game experience.

Figure 3.3 presents both the skill level measure calculated in (3.1) as well as the match skill levels for four different players during the year 2018. The top right panel of Figure 3.3 marks an example of a player only playing on the lowest (“normal”) skill level, whereas the bottom right panel shows a player almost always playing on the highest skill level and the bottom left panel shows a player mostly on the medium (“high”) skill level. Finally, the top left panel gives an example of a player changing average skill level over time, starting out predominantly at the lowest skill level and then switching between matches on both “normal”, “high” and “very high” skill levels.

3.4.4 Controlling for Team Composition

Besides player and team performance, the outcome of a Dota match also depends on the players’ choice of characters (so-called “heroes”). In Dota every hero has a unique set of abilities and a corresponding playing style. Due to the stark differences between heroes, their relative strength is not always equal.[20] This is evident from comparing average match winning rates across heroes. Figure 3.4 displays violin plots, showing the distributions of winning rates and frequency of selection of heroes across all matches in the sample. The distributions indicate that both the winning rate as well as the probability with which a hero is chosen starkly differs.

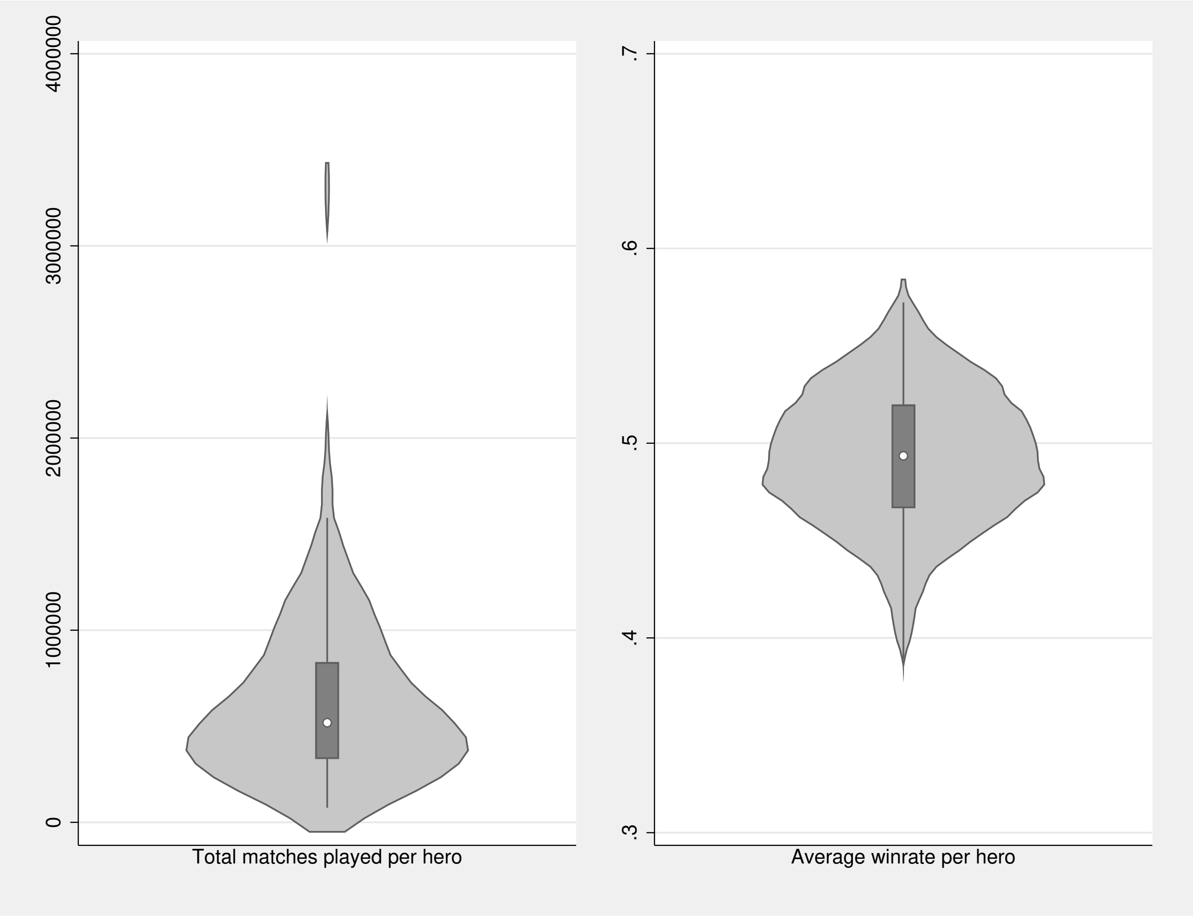
A priori, the fact that certain heroes achieve a higher winning rate on average, should not affect whether or not players exhibit a hot hand effect. However, a player’s choice of heroes could also be affected her previous performance. To flexibly capture the effect of the presence of certain heroes in the match on winning rates, I include fixed effects in the regression model indicating whether a certain hero was present and if it was present on the player’s team or the opposing team.[21]

Besides the strength of a given hero, it also matters how heroes match up against each other in a given match. Due to their differences in abilities and playing style a hero may perform better against some heroes and worse against others. Similarly, a hero may have synergies with other heroes in the same team that complement his particular playing

[20] As heroes are changed with updates to the game, their relative strength may also change over time. The analyses in this chapter are based on a relatively narrow timeframe such that this should not substantially alter results.

[21] The latter is important as a strong hero increases a player’s chance to win when being selected on the player’s team, while it decreases the player’s winning chances when chosen by the opposing team.

Figure 3.4: Matches played and average winrate for different heros



style. In the hero selection stage before the start of the match (the so-called “draft”) heroes may therefore be chosen strategically to counter or support other heroes. Thus, besides the first order effect of hero choice on winning chances, there may exist a second order effect of combinations of heroes on winning chances. Although in principle, one would therefore want to control for all combinations of hero choices, this is technically not feasible.[22] I therefore choose a more flexible way to control for the composition of the teams competing in a match. Again, an effect on the hot hand should only be found if a player’s strategic choice or line-up is affected by her previous performance.

In Dota, each hero has both a primary attribute as well as an attack type. The primary attribute concerns the damage a hero can cause and typically also determines their role. There are only three primary attributes in Dota: “Strength”, “Intelligence”

[22] With in total 115 heroes being playable in Dota this would result in far too many possible combinations.

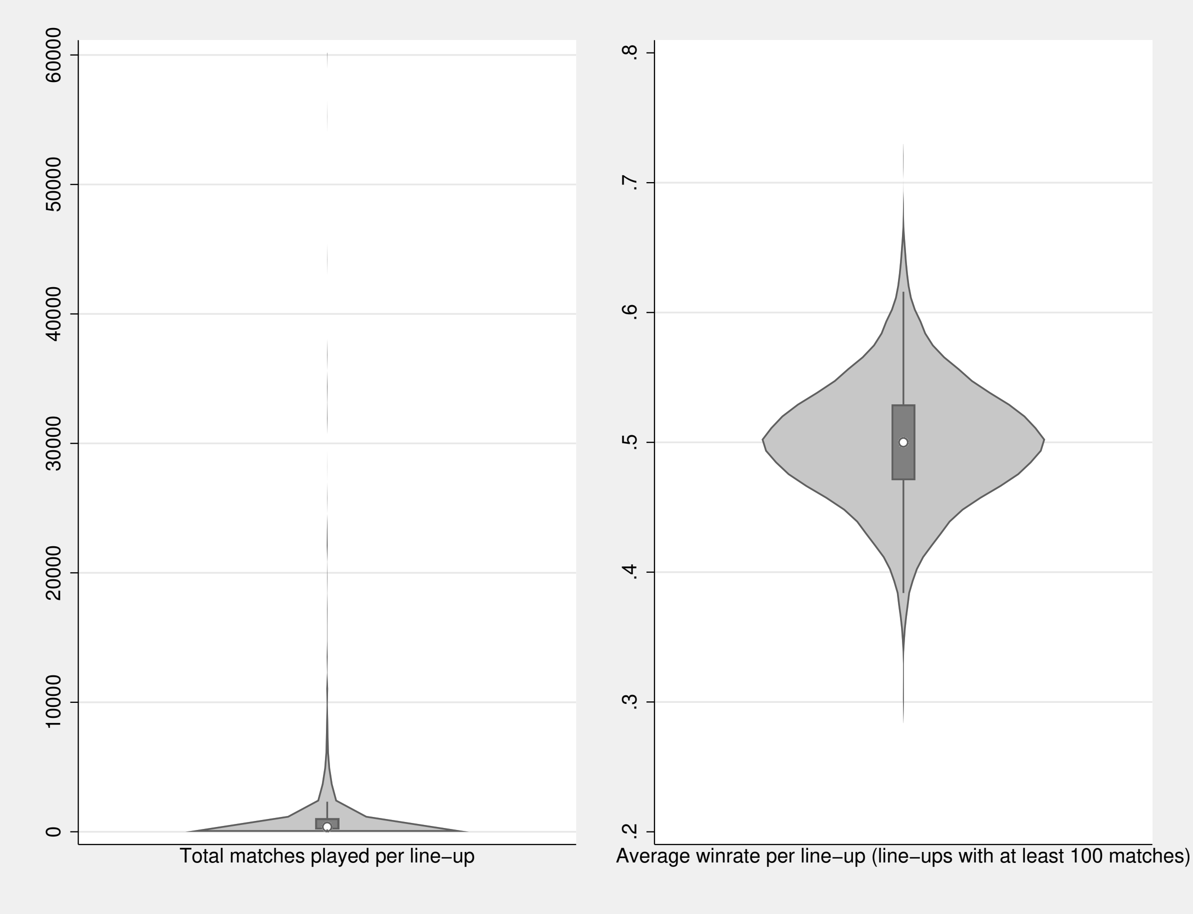
and “Agility”, which broadly define how a hero functions in fights.[23] The attack type of heroes is either “Ranged” or “Melee” and defines if a hero attacks from a safer distant or a more risky close position. I will define each hero in the game as a combination of its primary attribute and its attack type. The set of hero types will therefore be $H = \{MA, MI, MS, RA, RI, RS\}$, with MA representing attack type “Melee” with primary attribute “Agility” and so forth. The player’s team in match m, T_m, is then defined as a combination of five heroes from set H. The opponent team in match m, O_m is defined similarly. The line-up of match m, L_m, is then defined as the combination of sets T_m and O_m and therefore represents the primary attribute and attack type combinations of the player’s team and the respective combinations of the opposing team.[24] For each of these line-up combinations, I will include a fixed effect in the regression model. This reduces the number of combinations to a feasible amount but preserves a high degree of flexibility.

Figure 3.5 displays the distributions of the average winning rate of all team compositions (right panel) and choice of line-ups (left panel) in the data. As can be seen from the graph, the median line-ups (represented by the white dot) has a winning rate at almost exactly 50%. The interquartile range of winning rates (represented by the dark rectangle) is fairly narrow around the median. However, there is a wide range of line-ups with winning chances substantially above and below 50%. The left panel of Figure 3.5 shows that the majority of line-ups has a fairly low number of matches played. The majority of matches is played with a fairly limited number of line-ups.

[23] The website https://dota2.gamepedia.com/Attributes (accessed on 28/06/2022) notes that “Strength is a measure of a hero’s toughness and endurance. Strength determines a hero’s maximum health and health regeneration. Heroes with strength as their primary attribute can be hard to kill, so they often take initiator and tank roles, initiating fights and taking most of the damage from enemy attacks. [...] Agility is a measure of a hero’s swiftness and dexterity. Agility determines a hero’s armor and attack speed. Heroes with agility as their primary attribute tend to be more dependent on their physical attacks and items, and are usually capable of falling back on their abilities in a pinch. Agility heroes often take so-called carry and gank roles. [...] Intelligence is a measure of a hero’s wit and wisdom. Intelligence determines a hero’s maximum mana and mana regeneration. Heroes with intelligence as their primary attribute tend to rely on their abilities to deal damage or help others. Intelligence heroes often take support, gank, and pusher roles.”

[24] An example of a possible line-up would therefore be $MS, MA, RA, RI, RI - MS, MS, MA, RA, RS$.

Figure 3.5: Matches played and average winrate for different team compositions (from the perspective of the player's team)



3.4.5 Player Performance

A natural determinant of match success is a player's performance in a match. There are multiple individual performance indicators in Dota, which could be seen as candidate control variables to explain match success. I will consider the following indicators to measure player performance in a match:[25]

[25] There are further potential performance indicators available which are not considered here:

1. So-called "last hits" occur when a player deals the killing blow to a unit, building or hero from the opposing team. Players are awarded gold and experience for "last hits". These indicators are therefore preferred as more direct performance indicators.
2. Player level indicates which level the player's hero has reached at the end of the match. Higher level heroes are generally stronger. As player level is awarded based on experience gathered, this is preferred as a more direct performance indicator.

1. **The hero experience per minute.** Hero experience is gained by killing units or other heroes on the map or being present when they are killed. With more experience gathered heroes grow stronger during a given match. The hero experience per minute is therefore an indicator for the activity of a player.

2. **The gold gathered per minute.** Gold is the currency awarded in a given match and is mainly used to improve a hero's power through the purchase of certain items. Gold is gathered from several sources but most prominently from killing enemy units, heroes or building. The gold gathered per minute is therefore a good indicator of a player's success during the match.

3. **The Kill/Death/Assist (KDA) ratio.** The KDA ratio measures how often the player kills or assists in killing enemy heroes relative to being killed herself and is therefore an indicator how well a player does in fighting heroes of the opposing team. Formally, to be defined in instances where a player achieves zero deaths, I will calculate the KDA ratio as $KDA = \frac{Kills+Assists}{1+Deaths}$

Although these performance indicators will likely be highly correlated with the match outcomes, they should not be included in the model directly as they are so-called "bad controls" due to them not only being predictors of the match outcome but rather being a potential outcome of the treatment (i.e. whether or not the last game was lost). Including these variables could therefore lead to significant biases in coefficient estimates.[26] The

[26] See e.g. Angrist and Pischke (2008) Section 3.2.3. In fact, the problem here is similar the the second example of inappropriate proxy controls in Angrist and Pischke (2008). Suppose match success, $w_{it} = \{0, 1\}$, is a function of two player performance indicators: a transitory systematic shock to performance due to the momentum of a win in the last match, $w_{i,t-1}$, and a random shock to performance e.g. due to better or worse team synergies, k_{it} such that

$$w_{it} = \alpha + \beta w_{i,t-1} + \delta k_{it} + \varepsilon_{it}. \tag{3.2}$$

Now suppose that k_{it} is to be approximated by some match-level performance statistic, $\hat{k}_{it}$, similar to the ones discussed above but that these performance statistics not only reflect the random shock to performance but also the transitory shock to performance $w_{i,t-1}$ such that

$$\hat{k}_{it} = \theta + \gamma w_{i,t-1} + \pi k_{it} + \mu_{it}. \tag{3.3}$$

If this is now substituted for k_{it} in equation 3.2, we have

$$\begin{aligned} w_{it} &= \alpha + \beta w_{i,t-1} + \delta \left[\frac{\hat{k}_{it} - \theta - \gamma w_{i,t-1} - \mu_{it}}{\pi} \right] + \varepsilon_{it} \\ &= \left[\alpha - \delta \frac{\theta}{\pi} \right] + \left[\beta - \delta \frac{\gamma}{\pi} \right] w_{i,t-1} + \frac{\delta}{\pi} \hat{k}_{it} + \left[\varepsilon_{it} - \frac{\delta}{\pi} \mu_{it} \right]. \end{aligned} \tag{3.4}$$

If we assume that $\delta > 0$, i.e. better performance increases the chances of winning, $\gamma > 0$, i.e. there is also a potential hot hand effect on player performance in the current match and $\pi > 0$, i.e. $\hat{k}_{it}$ is a intuitive reasoning is as follows: if a player's performance is negatively affected by a loss in the last match (or vice versa positively affected by a win), she may perform worse in the current match and therefore have both worse performance statistics as well as a lower chance of winning the match. The performance statistic will therefore "soak up" part of the hot hand effect. The essential issue here is that player performance in the current match is not predetermined relative to the variable of interest (the hot hand effect) and therefore performance statistics are likely "bad controls". I will therefore not include player performance statistics of the current match in the regression.[27]

Player performance statistics are nevertheless useful to analyze, as they may in fact be used as alternative outcome variables. If a hot hand effect can be found with respect to winning rates, it is likely that a corresponding effect will be found with respect to individual player performance.

3.4.6 Regression Model

In this section, I will discuss the model that is used to estimate the hot hand effect taking into account the points discussed above. Let $w_{i,m,s}$ be a binary variable indicating whether player i has won (w_{ims}=1) or lost (w_{ims}=0) match m in playing session s. Let $w_{i,m-1,s}$ be the outcome of the match previously played by player i in the same playing session. As discussed in Section 3.4.1, if the last match was played a while ago, it is unlikely that it still substantially affects play in the current match. I will therefore only consider matches where $w_{i,m-1,s}$ is the actual observable outcome of the last match in the current playing session.[28] To simplify notation, I will henceforth omit the index for the playing session. The basic estimation equation can be written as

$$w_{im} = \alpha + \delta w_{i,m-1} + \beta X_{im} + T_m + O_m + L_m + \varepsilon_{im} \tag{3.5}$$

valid proxy for k_{it}, the effect on the hot hand, $\left[\beta - \delta\frac{\gamma}{\pi}\right]$ will be underestimated. In the best case, one may assume that $\gamma = 0$, i.e. there is no hot hand effect in the observable performance measure. This assumption is extremely restrictive and would essentially require that the performance proxy only measures natural fluctuations in performance that are independent of a potential hot hand effect.

[27] By the same argument including performance statistics of team members or opponents may lead to a similar bias. If a player performs better in a given match, this may also open up opportunities for team members and limit possibilities of opponents. The resulting correlation between player performance statistic may therefore induce similar, although dampened "bad control" effects.

[28] This effectively removes the first observation in each playing session. An alternative approach would be to preserve the first observation and assume that there is no hot hand effect in the first match of a session. The first observation in each session could then still be used to identify other variables in the model. As the dataset used here is large enough, there is no drawback of disregarding the first observation of each playing session.

where δ is the parameter of interest indicating the hot hand effect, X_{im} is a set of controls on the player-by-match level, T_m and O_m are sets of match-level fixed effects indicating the presence of heros in the player's team (T_m) as well as the opposing team (O_m) and L_m is a match-level fixed effect indicating the lineup of the match as discussed in Section 3.4.4.[29]

As the focus is chapter is to investigate if the hot hand effect changes with player skill and experience, I will extend the model in (3.5) to include interactions with the measures for skill and player experience. As discussed above, I will define skill and experience brackets based on deciles to allow heterogenous effects for different skill and experience levels. Let $skill_{im}$ be the skill level of player i in match m and let $P_n\left(skill_{im}\right)$ an indicator variable for $skill_{im}$ being below the nth decile and above the $(n-1)th$ decile across all observations in the data.

$$\begin{aligned} w_{im} = \alpha + \delta w_{i,m-1} + w_{i,m-1} \sum_{n=1}^{10} [\gamma_{1n} P_n\left(skill_{im}\right) + \gamma_{2n} P_n\left(ranked\_exp_{im}\right) \\ + \gamma_{3n} P_n\left(match\_exp_{im}\right)] + \beta X_{im} + T_m + O_m + L_m + \varepsilon_{im} \end{aligned} \qquad (3.6)$$

3.5 Descriptive Evidence

Before turning to the regression analysis, I will discuss in this section if the data exhibits any descriptive evidence of a hot hand effect. Before turning to this, Table 3.1 provides an overview of the dataset used for the analysis, i.e restricted to the ranked/all draft mode analyzed below.

The dataset contains 7,282,263 observations in total, of which 4,450,638 contain information the skill level on which the match was played. As discussed above, I use a moving average over 30 days before and after the current match played to calculate the skill level of a player. This allows to calculate player skill level for a total of 5,943,301 matches. With regard to player experience, measured by matches played, the data further allows

[29] A potential concern with the functional form in specification (3.5) is the binary nature of the dependent variable, for which a Logit or Probit model would in principle be more suitable. However, for the data used in this chapter, the drawbacks of using a linear probability model are not very marked, while the linear probability model has clear advantages in interpretability of the coefficients as marginal effects. In particular, a linear-probability model should fit the data sufficiently well as the predicted winning probabilities will likely be very close to 50% such that the problem of model predictions outside the interval $[0, 1]$ is unlikely to occur.

Table 3.1: Summary statistics (only ranked/all draft matches)

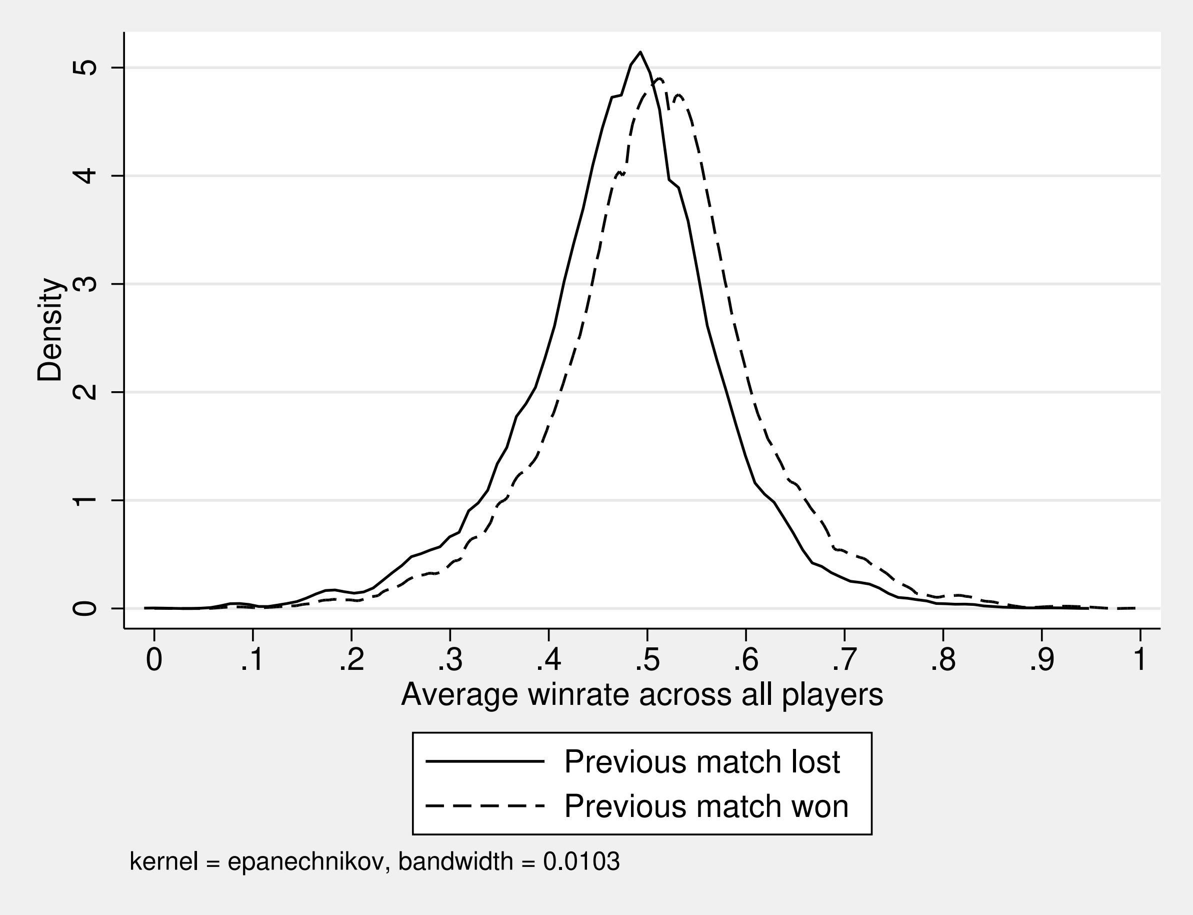
Variable	Obs	Mean	Std. Dev.	Min	Max	P50
Match win	7282263	.5	.5	0	1	0
Total matches played	7282263	2989.92	1995.84	1	14071	2716
Total matches played (ranked, all draft)	7282263	239.16	245.37	1	3040	165
Match skill level	4450638	1.88	.87	1	3	2
Approx. player skill level	5943301	1.71	.76	1	3	1.41
Player kills	7282263	7.62	5.46	0	65	6
Player assists	7282263	14.09	7.48	0	79	13
Player deaths	7282263	8.12	4.05	0	92	8
KDA ratio	7282263	3.12	2.94	0	68	2.36
Player denies	7282263	8.28	8.94	0	211	5
Player gold per min.	7282263	422.17	133.76	90	1726	405
Player last hits per min.	7282263	154.86	118.42	0	3136	129
Player level at end of match	7282263	21.01	4.22	1	25	22
Player exp. level per min.	7282263	521.73	148.28	0	1463	523
Time between matches in min.	7282201	1244.56	12001.92	0	2882845	35.28
Match duration	7282263	2493.13	659.88	361	19522	2443

to differentiate between total matches played in all game modes and matches only played in the ranked/all draft modes analyzed here. Even though average match experience in the ranked/all draft mode on average amounts to only less than 10% relative to average match experience across all game modes, it could still have a very high relevance for players' performance and susceptibility to a hot hand effect, if learning in this particular game mode is more valuable or more relevant than learning in other game modes. For this reason, I will consider both dimensions of player experience in the analysis below.

A first indication that a hot hand effect can be detected in the data may be found from a comparison of the conditional distributions of average winning rates across players. Figure 3.6 displays the estimated kernel densities of players' average winning rates conditional on whether or not the previous match was lost. As discussed in above, I restrict my analysis to consecutive matches with short breaks and matches only played in the ranked/all draft game mode.

The distribution of winning rates is evidently shifted to the right when the previous match played was won. This is clearly indicative of a hot hand effect in the data. Interestingly, the distribution appears to be shifted across all winning rates, which indicates that both high as well as low winning rate players are equally affected by the hot hand. As the aim of this chapter is to discuss how the hot hand varies with player skill and experience, Figures 3.7 and 3.8 compared these distributions both for matches played on low and high skill levels (Figure 3.7) as well as matches played on low and high experience levels (Figure 3.8).

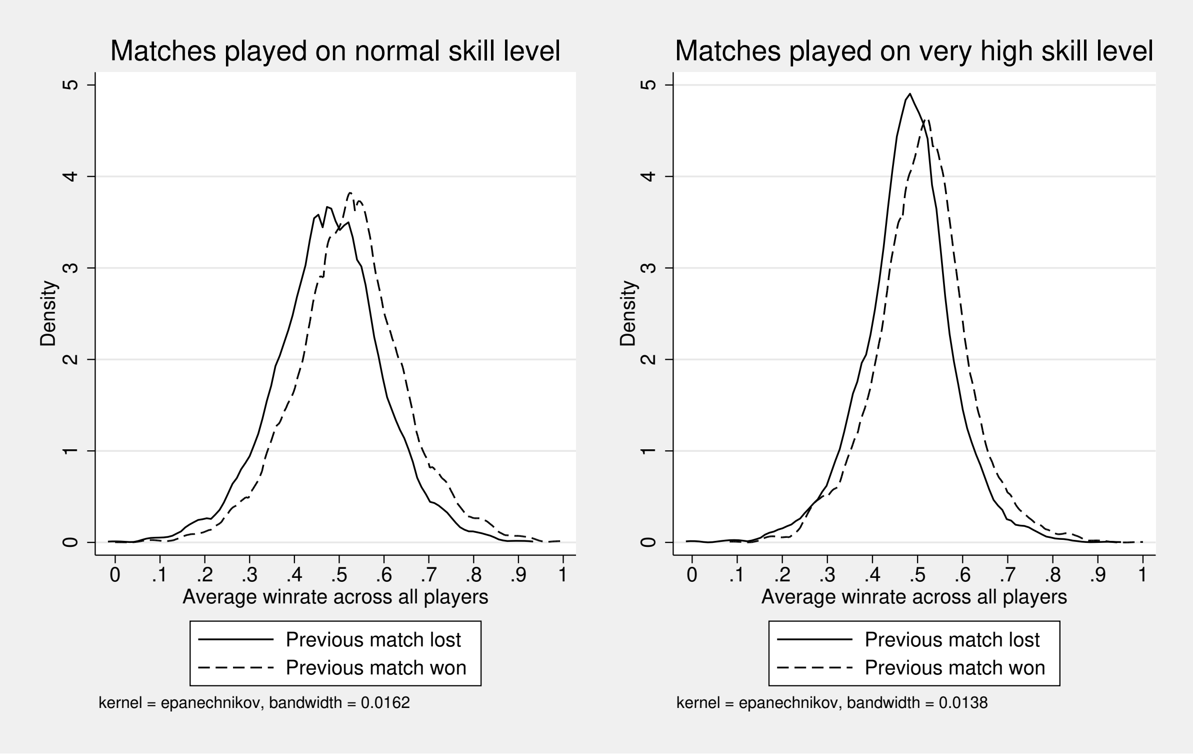
Figure 3.6: Kernel density estimate of average winning rates of players when last match was lost or won



Clearly, both figures show that a hot hand effect also tends to be present on higher skill and experience levels. Both figures also indicate that this effect might be lower on high skill and experience levels. Notably, the distribution of winning also clearly differs between low and high skilled and experienced players, with the latter seemingly delivering a more consistent performance on average as shown by the higher density of winning rates close to 50%.

Interestingly, both figures show that the distribution of winning rates as well as its change is similar for high and low skill and game experience. This may be indicative of skill level and game (mode) experience having similar effects. Both high experience and high skill level yield distributions of winning rates with lower range and higher densities around 50% as well as a smaller shift to the right when the previous match was won rather than lost, i.e. a smaller hot hand effect. However, it is a priori not clear whether this implies that both skill and experience affect the hot hand or whether this is mainly

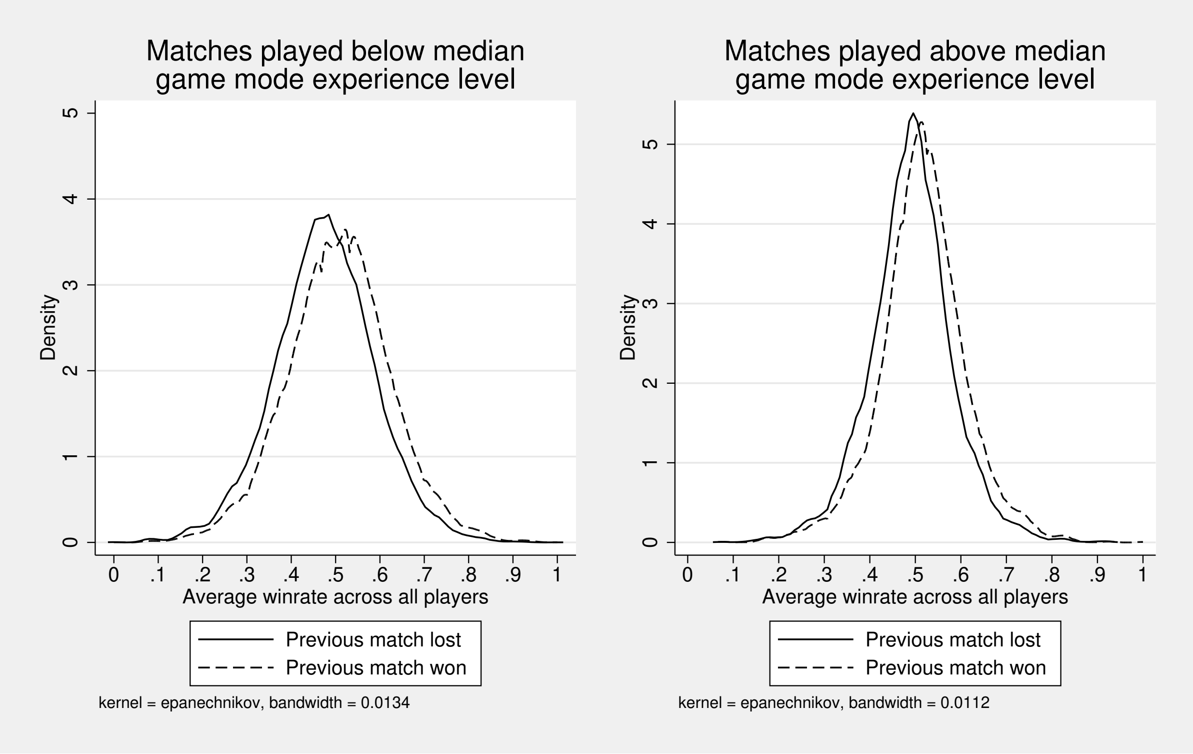
Figure 3.7: Kernel density estimate of average winning rates of players when last match was lost or won



indicative of a strong correlation between skill level and experience. If players with a higher experience level also show a higher skill level (e.g. due to learning) the observed shift in winning rate distributions may simply be caused by the same underlying factor for both low as well as high skill and experience levels.

Whether the change in winning rates is attributable more to changes in experience or skill is therefore a question that should be answered when looking at variations in both dimensions simultaneously. Moreover, the figures presented above only distinguish between low and high skill and experience levels, while the effect on the hot hand could of course also be non-linear in the sense that the hot hand effect vanishes more quickly with gaining skill and experience on low levels relative to high levels. The section below, which presents the results of the regression models described in Section 3.4, will shed further light on these questions.

Figure 3.8: Kernel density estimate of average winning rates of players when last match was lost or won



3.6 Results

In this section, I present the results of the model described in Section 3.4 and discuss results of two extensions shedding further light on nuances of the hot hand effect.

3.6.1 Average Hot Hand Effect

Table 3.2 presents results of the regression model described in Section 3.4. Specification (1) of Table 3.2 represents the most basic model of the hot hand as a regression of the indicator variable for winning a match on the outcome of a player's last match. I find that a player's winning chance increases by 2.2 percentage points if she has also won the match immediately before. Hence, the results suggest that on average a hot hand effect can indeed be found in the data.

Specifications (2) to (4) investigate if this result is robust to adding further match-level controls. Indeed, adding dummy variables for the side on which the player's team is playing (specification (2)), the heroes chosen for the match (specification (3)) and the line-up (specification (4)) does impact the estimate of the hot hand effect. None of these robustness checks affects the key result: the hot hand effect is consistently estimated to be around 2 percentage points. This shows that neither the realization of the side on which the player is playing nor the choice of heroes influences whether a hot hand effect occurs in the data. This is despite the fact that the added controls have a relevant effect on the outcome. Playing on the Radiant match side for instance has a substantial effect on the probability of winning the match. However, the realization of the player's side is random in itself. A correlation with streakiness could therefore only be expected if the realization of the match side would affect the player's performance differently depending on the outcome of the last match. The results indicate that this is not the case. It would also be surprising as players are aware that the realization of the match side is random and hence their performance should not be impacted.

Table 3.2: Regression results: average hot hand effect

	(1)		(2)		(3)		(4)	
Last match won	0.022***	(0.000)	0.022***	(0.000)	0.021***	(0.000)	0.021***	(0.000)
Playing on Radiant side			0.061***	(0.000)	0.061***	(0.000)	0.061***	(0.000)
Constant	0.489***	(0.000)	0.459***	(0.000)	0.445***	(0.010)	0.304***	(0.007)
Hero FE	No		No		Yes		Yes	
Line-up FE	No		No		No		Yes	
adj. R^2	0.00		0.00		0.05		0.05	
Observations	5832631		5832631		5832631		5832631	

Standard errors in parentheses (clustered on the player level)
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Similarly, adding fixed effects for heroes and the line-up does not change the results – even though adding fixed effects substantially increases the explanatory power of the model as can be seen from the increased R^2 in specifications (3) and (4) in Table 3.2. Hence, players' average performance is also not indirectly affected through more (less) optimal hero and line-up choices following a win (loss). As the available data allows for a relatively granular fixed effects setup, which can also be seen by the fact that the precision of the hot hand estimate does not change when adding more fixed effects, I will include both hero as well as line-up fixed effects in all subsequent analyses.

3.6.2 Does the Hot Hand Effect Depend on Skill and Experience?

I will now turn to the main research question in this chapter and discuss if and how the hot hand effect differs for players playing on a higher skill level or having more experience. As explained above, I will use decile brackets for the skill and experience levels to allow for an easier interpretation of results. Furthermore, this approach will reflect potential non-linearities in the relationship between skill and experience level and the hot hand effect.[30]

Table 3.3 shows the results of these regressions. In specification (1) interactions with the average skill level of players are introduced.[31] The results show that the average hot hand effect for players playing on the lowest skill level is even slightly below the average hot hand effect across the population that was estimated in Table 3.2. Surprisingly, for players with a medium-low or medium-high skill level, the hot hand effect is even stronger while it is not different from the lowest level for the medium skill bracket. The highest skill bracket exhibits a significantly lower hot hand effect compared to the lowest skill bracket.

Overall, the pattern of interaction terms is in line with the expectations. The hot hand effect is stronger for the lower skill levels and becomes weaker for the high skill levels. The notable exceptions are the lowest skill bracket as well as the medium-low skill bracket from skill level 1.83 to 2.19 and medium-high skill bracket from skill level 2.66 to 2.95. A possible explanation for these results might be that players on the medium-low and medium-high skill levels might be "trying harder" to win and advance out of their current position to higher ranks and may therefore be more susceptible to winning and losing streaks affecting their performance, while players in the medium skill bracket may feel more comfortable with their current ranking position.

Specifications (2) and (3) investigate how the hot hand effect varies with player experience measured as experience in matches played in all game modes (specification (2)) and matches played in the ranked/all draft game mode which is analyzed here (specification (3)). In both specifications the pattern is very clear: there is a relatively strong hot hand effect for the unexperienced players that reduces with an increase in the number of matches played. Comparing the relevance of experience across all matches and experience only in the ranked/all draft game mode, it is clear that match experience in the ranked/all draft game mode is the more relevant factor for reducing the hot hand effect – especially for the higher experience brackets. This is further confirmed in specification

[30] The results are, however, robust to both modeling with linear-quadratic terms instead of decile brackets as well as a more granular choice of brackets.

[31] It should be noted that only six interaction effects are shown here as there is a substantial proportion of players playing at the lowest skill level. The lowest bracket therefore represents more than 10% of players. Also note that the number of observations is lower in specifications (1) and (4) of Table 3.3. This is due to the fact that skill level is not recorded for all matches and therefore cannot be calculated for all players at all points in time. Whether the skill level for a match is recorded in my dataset is random, such that any sort of selection bias is not expected.

Table 3.3: Regression results: interactions with skill and experience

	(1)		(2)		(3)		(4)	
Last match won	0.019***	(0.001)	0.029***	(0.001)	0.035***	(0.001)	0.046***	(0.003)
Interaction skill level								
× 1.08 to 1.41	0.007***	(0.002)					0.008***	(0.002)
× 1.41 to 1.83	0.006***	(0.002)					0.008***	(0.002)
× 1.83 to 2.19	−0.001	(0.002)					0.001	(0.002)
× 2.19 to 2.66	0.004**	(0.002)					0.007***	(0.002)
× 2.66 to 2.95	0.001	(0.002)					0.005***	(0.002)
× 2.95 and above	−0.003**	(0.002)					0.001	(0.002)
Interaction exp all matches								
× 614 to 1168			−0.004**	(0.002)			−0.000	(0.002)
× 1168 to 1725			−0.006***	(0.002)			−0.004	(0.002)
× 1725 to 2233			−0.006***	(0.002)			−0.003	(0.002)
× 2233 to 2716			−0.010***	(0.002)			−0.007***	(0.002)
× 2716 to 3224			−0.010***	(0.002)			−0.007***	(0.002)
× 3224 to 3809			−0.010***	(0.002)			−0.006**	(0.002)
× 3809 to 4556			−0.010***	(0.002)			−0.005**	(0.002)
× 4556 to 5632			−0.010***	(0.002)			−0.004**	(0.002)
× 5632 and above			−0.012***	(0.002)			−0.004*	(0.002)
Interaction exp gamemode matches								
× 24 to 51					−0.006***	(0.002)	−0.012***	(0.003)
× 51 to 83					−0.010***	(0.002)	−0.020***	(0.003)
× 83 to 121					−0.010***	(0.002)	−0.021***	(0.003)
× 121 to 165					−0.017***	(0.002)	−0.028***	(0.003)
× 165 to 219					−0.016***	(0.002)	−0.027***	(0.003)
× 219 to 287					−0.017***	(0.002)	−0.028***	(0.003)
× 287 to 384					−0.020***	(0.002)	−0.031***	(0.003)
× 384 to 549					−0.018***	(0.002)	−0.029***	(0.003)
× 549 and above					−0.021***	(0.002)	−0.032***	(0.003)
Skill level								
1.08 to 1.41	0.006***	(0.001)					0.005***	(0.001)
1.41 to 1.83	0.008***	(0.001)					0.007***	(0.001)
1.83 to 2.19	0.009***	(0.001)					0.009***	(0.001)
2.19 to 2.66	0.013***	(0.001)					0.011***	(0.001)
2.66 to 2.95	0.016***	(0.001)					0.015***	(0.001)
2.95 and above	0.021***	(0.001)					0.021***	(0.001)
Experience over all matches								
614 to 1168			−0.001	(0.002)			−0.012***	(0.002)
1168 to 1725			−0.005***	(0.002)			−0.013***	(0.002)
1725 to 2233			−0.008***	(0.002)			−0.017***	(0.002)
2233 to 2716			−0.009***	(0.002)			−0.019***	(0.002)
2716 to 3224			−0.007***	(0.002)			−0.020***	(0.002)
3224 to 3809			−0.008***	(0.002)			−0.023***	(0.002)
3809 to 4556			−0.008***	(0.002)			−0.026***	(0.002)
4556 to 5632			−0.004***	(0.002)			−0.027***	(0.002)
5632 and above			−0.001	(0.002)			−0.029***	(0.002)
Experience over gamemode matches								
24 to 51					0.013***	(0.001)	0.025***	(0.002)
51 to 83					0.020***	(0.001)	0.040***	(0.002)
83 to 121					0.022***	(0.001)	0.046***	(0.002)
121 to 165					0.031***	(0.001)	0.057***	(0.002)
165 to 219					0.033***	(0.001)	0.061***	(0.002)
219 to 287					0.036***	(0.001)	0.064***	(0.002)
287 to 384					0.040***	(0.001)	0.069***	(0.002)
384 to 549					0.041***	(0.001)	0.071***	(0.002)
549 and above					0.042***	(0.001)	0.073***	(0.002)
Playing on Radiant side	0.062***	(0.000)	0.061***	(0.000)	0.061***	(0.000)	0.062***	(0.000)
Constant	0.395***	(0.007)	0.320***	(0.007)	0.272***	(0.007)	0.285***	(0.008)
Hero FE	Yes		Yes		Yes		Yes	
Line-up FE	Yes		Yes		Yes		Yes	
adj. R^2	0.05		0.05		0.05		0.05	
Observations	4777224		5832631		5832631		4777224	

Standard errors in parentheses (clustered on the player level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

(4), which looks at the effects of skill level, overall match experience and ranked/all draft match experience together.

Specification (4) shows an even higher baseline hot hand effect than the other specifications. The effect of skill level on the hot hand is very similar to the results in specification (1), while the effects of match experience change slightly compared to specifications (2) and (3). Even though growing overall match experience still reduces the hot hand effect, the order of magnitude of this reduction through more experience in ranked/all draft is substantially higher. In contrast to the reduction of the hot hand effect through more experience in ranked/all draft, which is growing very consistently with more matches played, the effect of more overall match experience is predominantly relevant for the very unexperienced players.[32]

Overall, the results point towards match experience being a main determinant of the hot hand. Despite the clear relevance of skill level, the number of matches played has a significant effect on the hot hand. The results also clearly indicate that experience in the specific game mode is more relevant than overall playing experience. A player in the highest bracket of game mode experience has a 3.2 percentage point lower hot hand effect than a player in the respective lowest decile bracket. What seems to matter is therefore experience in the specific setting the player is in. The ranked/all draft game mode is clearly a competitive setting. More match experience in this setting seemingly helps players to better understand the inherently random nature of the game and deliver a more consistent performance than less experienced players.

The results also show that there is a substantial heterogeneity in the estimated hot hand effect. Even though the evidence points towards the existence of a hot hand effect even for the most experienced players, the difference in the estimated hot hand effect between e.g. a high-skilled and very experienced player and an unexperienced player with medium-low or medium-high skill is substantial. This shows that at least in this setting estimating a single average hot hand effect across the entire population does not adequately capture the observed heterogeneity. Indeed, the hot hand effect varies substantially with skill level and playing experience. This highlights the importance of

[32] Interestingly, for the direct effect of experience on the probability of winning a match, a higher overall match experience may even reduce the chance of winning a match, while a higher ranked/all draft match experience very consistently also increases winning chance. This could be driven by effects of relative match experience. For a given number of ranked/all draft matches played, players with a lot of matches played overall spend relatively less time in the ranked/all draft game mode and relatively more time in other game modes. They may therefore be less focused on current playing strategies in this game mode and may play worse even though they have played a higher number of matches overall.

the setting and the characteristics of the analyzed sample for the evaluation of the hot hand effect. When estimating the hot hand effect on a sample of highly skilled and experienced Dota 2 players (e.g. professionals), I may at most find evidence for very low effects. Conversely, when analyzing a sample of rather inexperienced (casual) players, I may conclude that there is strong evidence for the hot hand in Dota 2. However, neither the conclusion that Dota 2 players show strong hot hand effects nor the opposite finding will correctly reflect which type of player is susceptible to the hot hand.

In the following sections, I will demonstrate that these findings are robust to changes of the outcome variable, alternative definitions of the hot hand and when limiting the maximum allowed break time between streaks.

3.6.3 Is the Hot Hand Effect Found in Player Performance Statistics?

In the section above, I have shown that a hot hand effect can be found in the data. Moreover, this hot hand effect varies with player skill level and match experience. Although this is already indicative of player performance being affected by "streakiness", in principle one could ask if this dependency in winning probability is in fact attributable to players' actual match performance. As I have explained above, performance indicators are not suitable controls in this setting as they themselves are outcome variables. If a player suffered a loss and this loss affects her performance in the next match, this should not only materialize in a lower winning probability but also in lower player performance as measured by the performance statistics discussed in Section 3.4.5.

Table 3.4 replicates specification (4) of Table 3.3 but using as independent variables the player performance statistics described in Tables 3.3.[33] Broadly, I expect similar results as in specification (4) of Table 3.3 if the hot hand effect identified can be traced back to actual measured performance of players.

This is indeed the case. All of the individual player performance statistics considered strongly and statistically significantly increase when the previous match was won. Similarly, the effect of skill level and experience broadly resembles the results presented above

[33] The regression models presented in Table 3.4 only use fixed effects for the hero selected by the player. Heroes vary in the degree to which they are specialized in achieving high experience levels, gold levels or the kill/death/assist ration (KDA) as a measure of fighting performance. The selection of the hero by the respective player is therefore of particular relevance. I still consider line-up fixed affects to account for the dependence of the line-up selected by the player's and the opposing team on winning rate.

Table 3.4: Regression results: individual performance statistics

	XP per min		Gold per min		KDA	
Last match won	17.179***	(0.883)	16.243***	(0.758)	0.240***	(0.019)
Interaction skill level						
× 1.08 to 1.41	0.835*	(0.473)	0.814**	(0.407)	0.038***	(0.011)
× 1.41 to 1.83	0.458	(0.461)	0.585	(0.399)	0.036***	(0.010)
× 1.83 to 2.19	−0.435	(0.454)	−0.279	(0.385)	0.011	(0.010)
× 2.19 to 2.66	0.170	(0.454)	0.706*	(0.382)	0.028***	(0.010)
× 2.66 to 2.95	0.825*	(0.488)	0.823*	(0.422)	0.014	(0.010)
× 2.95 and above	−0.111	(0.502)	−0.012	(0.426)	0.025**	(0.011)
Interaction exp all matches						
× 614 to 1168	−1.729**	(0.692)	−1.477**	(0.616)	−0.013	(0.015)
× 1168 to 1725	−2.918***	(0.689)	−2.626***	(0.605)	−0.027*	(0.015)
× 1725 to 2233	−3.075***	(0.686)	−2.696***	(0.612)	−0.029*	(0.015)
× 2233 to 2716	−3.327***	(0.682)	−3.306***	(0.600)	−0.040***	(0.014)
× 2716 to 3224	−3.132***	(0.672)	−3.320***	(0.590)	−0.041***	(0.015)
× 3224 to 3809	−3.159***	(0.676)	−3.207***	(0.594)	−0.038***	(0.015)
× 3809 to 4556	−3.298***	(0.695)	−3.303***	(0.608)	−0.042***	(0.015)
× 4556 to 5632	−2.916***	(0.678)	−3.081***	(0.599)	−0.028*	(0.015)
× 5632 and above	−3.423***	(0.723)	−3.473***	(0.625)	−0.028*	(0.015)
Interaction exp gamemode matches						
× 24 to 51	−4.614***	(0.883)	−3.938***	(0.719)	−0.050***	(0.019)
× 51 to 83	−7.490***	(0.858)	−6.637***	(0.712)	−0.112***	(0.018)
× 83 to 121	−7.895***	(0.834)	−7.211***	(0.697)	−0.110***	(0.018)
× 121 to 165	−9.237***	(0.839)	−8.480***	(0.701)	−0.140***	(0.018)
× 165 to 219	−10.236***	(0.828)	−9.070***	(0.690)	−0.162***	(0.018)
× 219 to 287	−10.321***	(0.827)	−9.826***	(0.688)	−0.154***	(0.018)
× 287 to 384	−10.417***	(0.828)	−9.545***	(0.692)	−0.160***	(0.018)
× 384 to 549	−10.865***	(0.829)	−9.980***	(0.691)	−0.164***	(0.018)
× 549 and above	−10.407***	(0.844)	−9.936***	(0.704)	−0.164***	(0.018)
Skill level						
1.08 to 1.41	4.372***	(0.715)	8.905***	(0.656)	0.145***	(0.014)
1.41 to 1.83	3.564***	(0.757)	11.153***	(0.695)	0.194***	(0.015)
1.83 to 2.19	2.948***	(0.754)	12.951***	(0.708)	0.215***	(0.015)
2.19 to 2.66	5.359***	(0.786)	17.680***	(0.782)	0.293***	(0.016)
2.66 to 2.95	3.385***	(0.873)	19.832***	(0.838)	0.310***	(0.017)
2.95 and above	−0.498	(1.060)	22.035***	(1.045)	0.377***	(0.020)
Experience over all matches						
614 to 1168	−6.340***	(0.940)	−4.119***	(0.859)	−0.035**	(0.018)
1168 to 1725	−6.910***	(1.025)	−4.205***	(0.953)	−0.020	(0.020)
1725 to 2233	−9.159***	(1.017)	−5.998***	(0.949)	−0.029	(0.019)
2233 to 2716	−9.138***	(1.004)	−4.971***	(0.937)	−0.035*	(0.020)
2716 to 3224	−9.900***	(1.017)	−5.030***	(0.949)	−0.043**	(0.020)
3224 to 3809	−11.867***	(1.047)	−6.583***	(0.971)	−0.082***	(0.020)
3809 to 4556	−10.418***	(1.094)	−4.851***	(1.028)	−0.062***	(0.021)
4556 to 5632	−13.648***	(1.165)	−6.295***	(1.098)	−0.069***	(0.023)
5632 and above	−13.815***	(1.377)	−3.897***	(1.299)	−0.050*	(0.026)
Experience over gamemode matches						
24 to 51	15.089***	(0.655)	12.456***	(0.529)	0.137***	(0.013)
51 to 83	23.060***	(0.678)	19.184***	(0.556)	0.216***	(0.013)
83 to 121	29.057***	(0.681)	24.490***	(0.567)	0.255***	(0.013)
121 to 165	34.015***	(0.696)	28.780***	(0.582)	0.301***	(0.013)
165 to 219	38.649***	(0.712)	32.335***	(0.598)	0.331***	(0.013)
219 to 287	41.743***	(0.733)	34.782***	(0.621)	0.319***	(0.014)
287 to 384	45.187***	(0.765)	36.392***	(0.654)	0.316***	(0.014)
384 to 549	48.629***	(0.826)	38.507***	(0.721)	0.286***	(0.015)
549 and above	53.255***	(1.064)	40.804***	(0.987)	0.251***	(0.019)
Playing on Radiant side	2.227***	(0.122)	3.772***	(0.100)	0.123***	(0.003)
Constant	620.264***	(1.368)	567.793***	(1.408)	3.068***	(0.028)
Hero FE	Yes		Yes		Yes	
Line-up FE	Yes		Yes		Yes	
adj. R^2	0.25		0.40		0.05	
Observations	4777224		4777224		4777224	

Standard errors in parentheses (clustered on the player level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

when the hot hand effect is measured by individual player performance statistics. This demonstrates that the hot hand effect not also materializes in players' average winning rates but also in the individual performance they deliver following a previous win or loss. Overall, these results confirm the findings presented above. Well-experienced and skilled players deliver a more consistent performance and are less affected by the hot hand in their play.

3.6.4 Long Streaks and Hot vs. Cold States

So far I have assumed that players enter a hot state once they have won a match. While this is a reasonable starting point for estimating a hot hand effect, it disregards two potential mechanisms that max have a significant impact on the estimated hot hand effect. Firstly, players may not enter a hot state immediately after winning one match but it may take several wins in a row for them to create momentum. Therefore only conditioning on the last match may not identify a hot state sufficiently well and consequently underestimate the true hot hand effect. Since Gilovich et al. (1985) many authors have therefore used streaks of two or three successes in a row to identify a hot state. Secondly, the conditional winning probabilities may differ between streaks of wins and streaks of losses. In the literature this has commonly been referred to as the difference between the hot and the cold hand (see e.g. Gilovich et al., 1985).

In this section, I analyze if these effects are also detectable in my dataset. I follow Green and Zwiebel (2018) and model the degree of the hot state as the average winning rate over a given number of previous matches. Substituting $w_{i,m-1}$ in (3.6) for the average winning rate over the previous k matches (excluding the current match), $\bar{w}(k)_{im} = \frac{1}{k}\sum_{j=m-k}^{m-1} w_{i,m-j}$, I estimate the following model:

$$\begin{aligned} w_{im} = \alpha + \delta\bar{w}(k)_{im} + \bar{w}(k)_{im} \sum_{n=1}^{10} [\gamma_{1n} P_n\left(skill_{im}\right) + \gamma_{2n} P_n\left(ranked\_exp_{im}\right) \\ + \gamma_{3n} P_n\left(match\_exp_{im}\right)] + \beta X_{im} + T_m + O_m + L_m + \varepsilon_{im} \qquad (3.7) \end{aligned}$$

This model is a generalization of (3.6) and essentially assumes that the hot state is not defined by the outcome of the last match but the success over recent matches. For $k = 1$ the model is equivalent to (3.6). For the purpose of this analysis I estimate the model in (3.7) for $k = 5$ and $k = 8$.[34]

[34] Green and Zwiebel (2018) use much longer histories of 10, 25 and 40 previous outcomes. Im my setting, I intend to avoid using histories that are too long as the matchmaking algorithms used in

While the model in (3.7) takes into account a longer playing history for estimation of the hot hand effect, it defines the state as a continuous variable and can therefore not differentiate between the potentially different effects of a very low and very high winning rate. To investigate this potential asymmetry, I additionally estimate a model that explicitly distinguishes between hot and cold states. Here, I define hot states as a number of successive wins in a row and cold states as a number of successive losses in a row and use dummy variables to reflect that a player is in the respective state. Formally, I define the hot state of player i in match m over the k previous matches, $H(k)_{im}$, as

$$H(k)_{im} = \prod_{j=m-k}^{m-1} w_{i,m-j}$$

and equivalently define the cold state, $C(k)_{im}$, as

$$C(k)_{im} = \prod_{j=m-k}^{m-1} \left(1 - w_{i,m-j}\right).$$

I then estimate the model in (3.6) using these two dummy variables and the respective interactions with the skill and experience brackets instead of the outcome of the last match, $w_{i,m-1}$. Note that again for $k = 1$ this model exactly reduces to the model in (3.6).[35] For the purpose of this analysis, I chose $k = 2$ (two consecutive wins/losses) and $k = 3$ (three consecutive wins/losses) to define the hot and cold states.

The results of the estimation models described above are presented in Table 3.5. I omit the coefficients of the interaction effects with the skill and experience brackets to preserve readability of the table. The effects of skill and experience level on the hot/cold hand effect are instead presented in Figures 3.9 and 3.10. These figures present the "net" hot hand effect for the respective skill and experience brackets (assuming that the player is in the first bracket for the other categories).

In specifications (1) and (2) Table 3.5 displays the model in (3.7) using the last 5 matches (specification (1)) and the last 8 matches (specification (2)) to determine the recent average winning rate that represents the hot state. Compared to the results in Table 3.3, the hot hand effects are substantially larger and in the range of 13.9 to 19.7

Dota 2 will lead to average winning rates mechanically reverting to 50% as players with a lot of (very few) wins are matched with better (worse) players in following matches.

[35] However, for $k = 1$ it is obviously not possible to separately identify hot and cold hand effects, as these must be estimated relative to a neutral state (i.e. any other sequence than consecutive wins or losses), which does not exist in case of $k = 1$.

Table 3.5: Regression results: average winning rate and hot and cold states

	(1)		(2)		(3)		(4)	
avg. win rate last 5 matches	0.139***	(0.012)						
avg. win rate last 8 matches			0.197***	(0.021)				
Last two matches won					0.042***	(0.005)		
Last two matches lost					-0.040^{***}	(0.004)		
Last three matches won							0.066***	(0.007)
Last three matches lost							-0.044^{***}	(0.006)
Playing on Radiant side	0.062***	(0.001)	0.062***	(0.001)	0.062***	(0.000)	0.062***	(0.001)
Interaction skill level	Yes		Yes		Yes		Yes	
Interaction exp all matches	Yes		Yes		Yes		Yes	
Interaction exp gamemode matches	Yes		Yes		Yes		Yes	
Hero FE	Yes		Yes		Yes		Yes	
Line-up FE	Yes		Yes		Yes		Yes	
adj. R^2	0.05		0.05		0.05		0.05	
Observations	2635375		1906925		3990313		3421770	

Standard errors in parentheses (clustered on the player level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

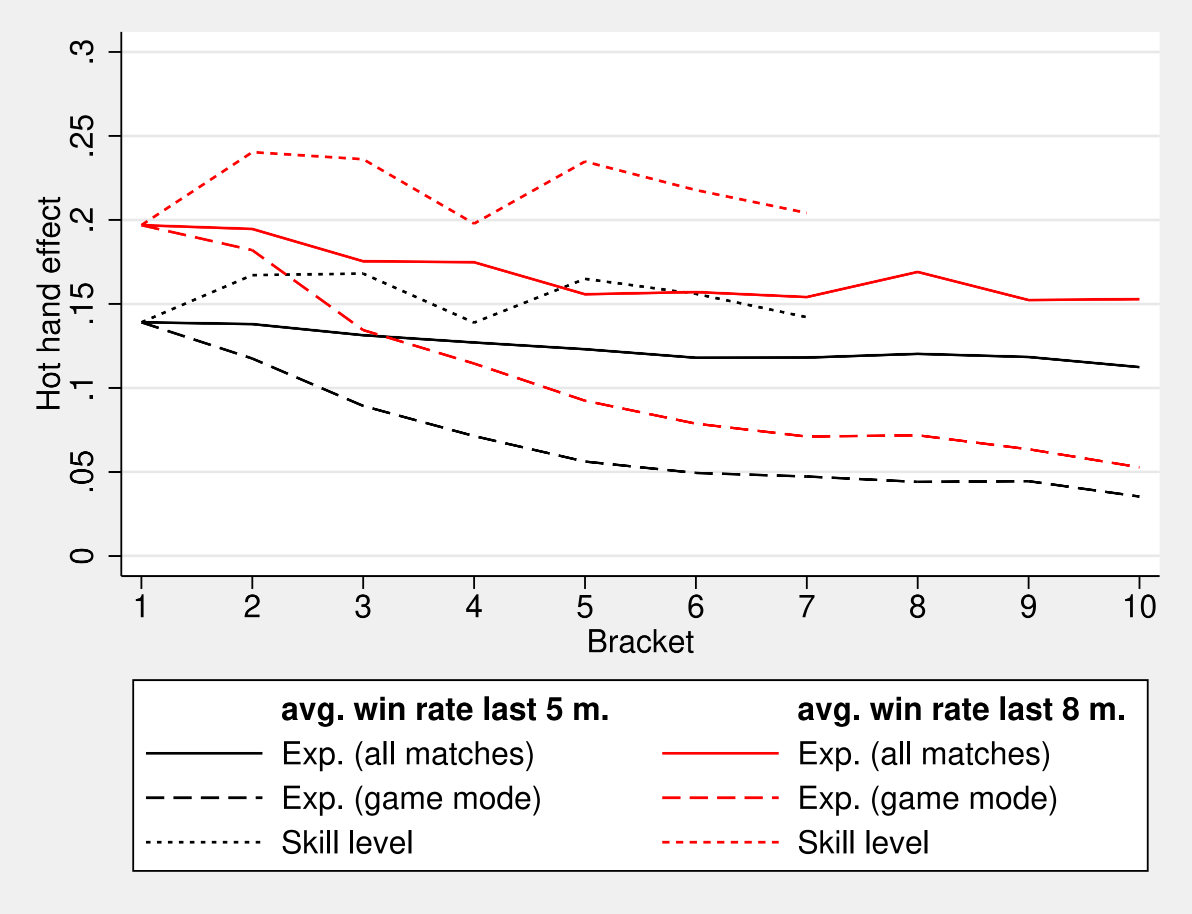
percentage points. This stark difference in effect sizes should, however, be interpreted with caution. Firstly, there might be a substantial selection effect. Calculating the recent average winning rate requires the observation of uninterrupted streaks that are long enough to include a sufficient number of previous matches (5 or 8 respectively in this case). As streak length is essentially a choice of players, the subsamples used in specification (1) and (2) (and equally also (3) and (4)) may not be directly comparable to the sample in Table 3.3. Secondly, the effects in specification (1) and (2) are estimated on average winning rates rather than a binary independent variable. Much of the variation will therefore be closer to the average winning rate of 50% and hence the size of the coefficient has to be interpreted with regards to the distribution of average winning rates.[36]

The interactions between the recent average winning rate and indicator variables for the skill and experience brackets are displayed in Figure 3.9. In the figure, the black lines correspond to specification (1) in Table 3.5 and the red lines correspond to specification (2) in Table 3.5. Each line shows the aggregate hot hand effect for a different skill or experience bracket, assuming that the player is in the lowest bracket for the respective other categories.[37] The patterns observed here are very similar to the results shown in specification (4) of Table 3.3. In particular, players playing on the lowest, medium and highest skill level show lower hot hand effects than players playing on the medium-low and medium-high skill levels. More match experience drastically lowers the hot hand

[36] Saying it differently, the entire 13.9 percentage points would only be applicable when comparing a player that has lost the last 5 (8) matches to a player that has won the last 5 (8) matches.

[37] This is of course an abstract representation as e.g. players in higher game mode experience brackets will automatically also move to higher overall experience brackets. An alternative interpretation would be that the figure displays the change in the hot hand effect when moving to higher brackets but adding the baseline hot hand effect.

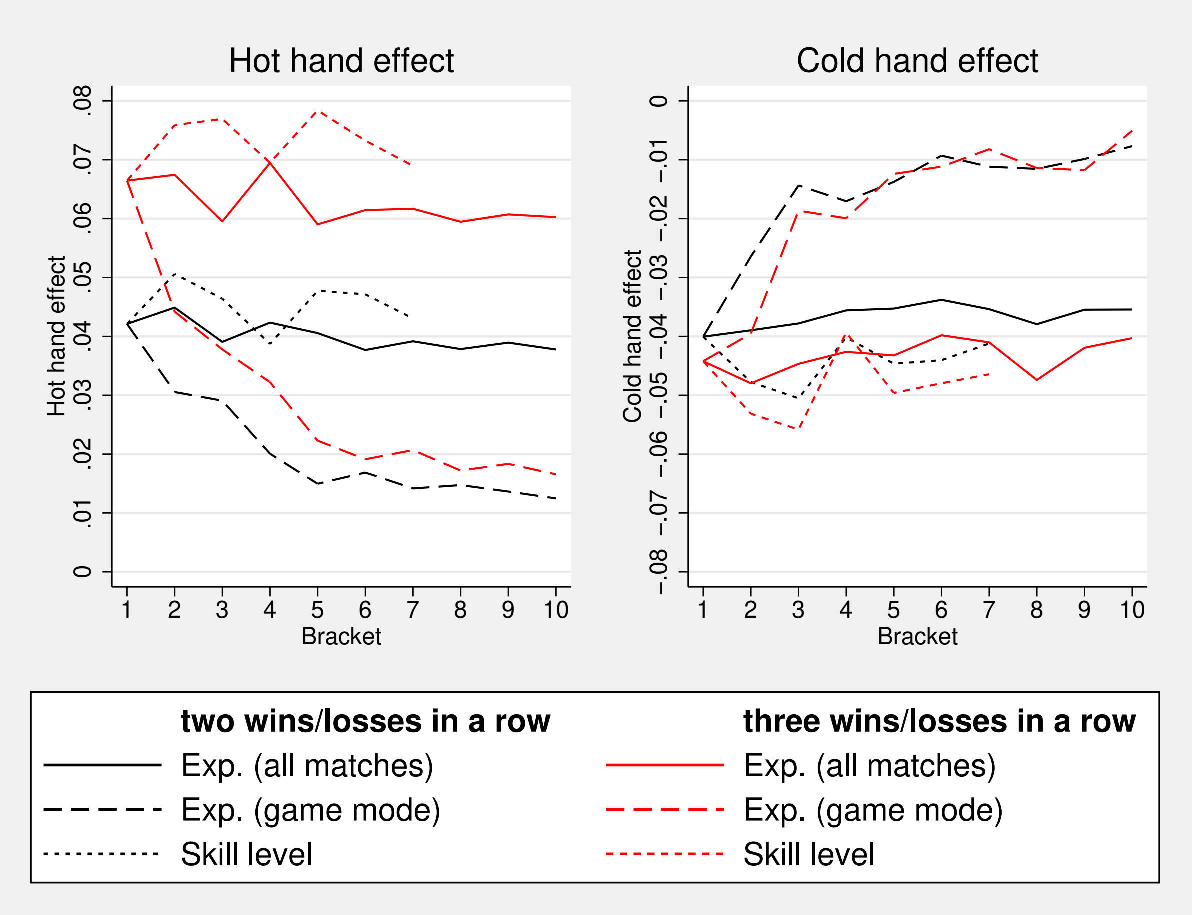
Figure 3.9: Hot hand effect for different skill and experience brackets for specifications (1) and (2) in Table 3.5



effect, with game mode experience having a much higher weight in reducing the hot hand effect than overall experience. A player assigned to the highest (overall and game mode) experience brackets would therefore show almost no hot hand effect – even when considering longer playing histories to determine the hot state. This disappearance of the hot hand effect with higher skill and experience levels is therefore in line with the results found with the simpler model that only conditions on the previous match outcomes.

Specifications (3) and (4) in Table 3.5 show the results of the model that enables the distinction between hot and cold states. Here, specification (3) considers the previous 2 matches ($k = 2$), while specification (4) considers the previous 3 matches (k=3). For $k = 2$, the hot and the cold hand effect are approximately of equal size with an impact of approximately 4 percentage points on winning chances. For $k = 3$, the cold hand effect remains in the same range while the hot hand effect increases substantially by approximately 50%. In both specifications, the differential is considerably larger than

Figure 3.10: Hot and cold hand effect for different skill and experience brackets for specifications (3) and (4) in Table 3.5



the hot hand effect estimated in Table 3.3 but lower than the effects estimated on the recent average winning rates in specifications (1) and (2) in Table 3.5.

The results of the interactions with skill and experience levels are again presented graphically in Figure 3.10 with the left panel displaying the hot hand effect and the right panel displaying the cold hand effect. The black lines correspond to the results of specification (3) in Table 3.5, while the red lines correspond to the results of specification (4) in the same table. Again, similar patterns can be observed with low, medium and high skill levels exhibiting relatively lower hot and cold hand effects and increasing game mode experience driving the hot and cold hand effect towards zero.

Overall, the results of these extensions are very consistent with the results presented in Table 3.3. Even though the estimates of the baseline hot (and cold) hand effects can be substantially larger when considering longer histories of recent successes or streaks of a higher number of consecutive wins or losses, the pattern of the interactions with skill and experience level are still the same. The lowest hot hand effects are found for players playing on low, medium or high skill level (compared to medium-low and medium-high skill levels) and particularly players with a lot of experience in the ranked/all draft game mode in which the matches are played.

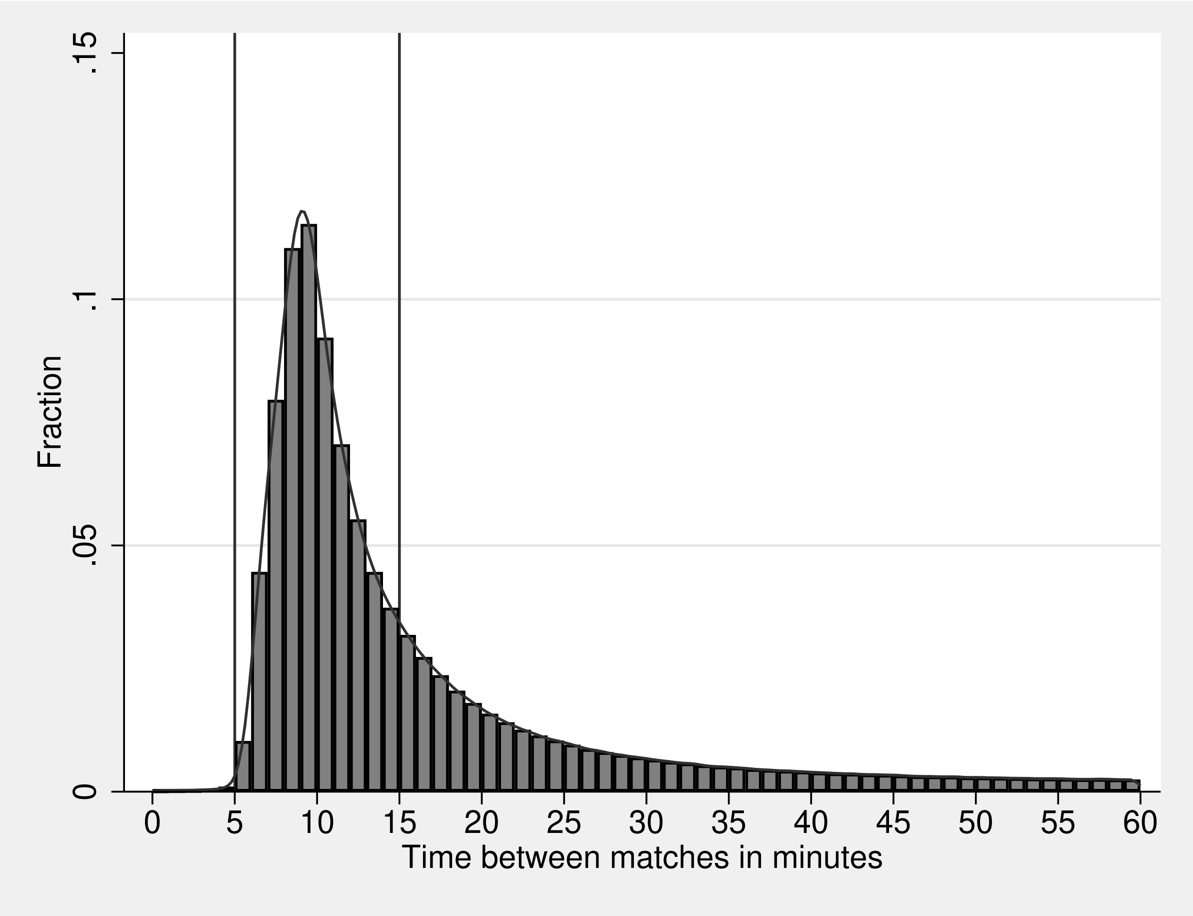
3.6.5 Short Breaks Between Matches

As discussed in Section 3.4.1, I allow for rather long breaks of up to one day in defining a series of matches on which I estimate a hot hand effect. This enables me to also analyze the effects of longer streaks or hot and cold states as presented in the section above. A potential concern with allowing for rather long breaks between matches is the fact that players could “cool down” and leave the hot state during long breaks while remaining in a hot state during shorter breaks. This would then lead to an underestimation of the hot hand effect.

To assess the degree to which such an underestimation occurs and what impact it has on the results presented in Tables 3.3 and 3.4, I will conduct a sensitivity analysis below that uses much shorter maximum break times to define a series of matches. To define an appropriate cutoff, it is helpful to look at the distribution of break times between matches. Figure 3.11 displays the distribution of break time between matches up to breaks of one hour, calculated as the time between the start of the current and the end of the previous match.

It can be seen from the figure that there is almost always at least a small break between two matches. This occurs naturally as some time is spent between matches to re-queue, pick the heroes for the next match and load the actual match. As expected, the distribution of match breaks is right-skewed with the most common time spent between matches being around the 10 minute mark. There is no clear sign of a discontinuity in the distribution that would allow a separation of players which are re-queuing immediately and players taking a break. For the analyses below, I will consider breaks of up to 15 minutes to assign matches to a “playing session”. This seems appropriate, noting that the time difference between the shortest breaks and the median of the distribution in Figure 3.11 is approximately 5 minutes. Categorizing breaks up to 15 minutes as short breaks therefore considers a symmetric interval of plus/minus 5 minutes around the median of the distribution. A playing session, $\tilde{S}$ is therefore defined as a sequence of matches with breaks up to 15 minutes.

Figure 3.11: Distribution of break time between matches up to one hour (only ranked/all draft matches)



Playing session

Matches belong to the same playing session if breaks between all matches are short. The break between two matches is considered to be short if it is less or equal to 15 minutes.

Table 3.6 presents the estimation results of the full model with all interactions with skill and experience levels, using the match outcome and the three individual player performance statistics as the dependent variables. It can be seen from the results that the estimated hot hand effect is considerably larger and between 25 to 30% higher compared to the estimates presented in Tables 3.3 and 3.4, which allow for longer breaks between matches. The interaction effects with skill level as well as game mode and overall expe-

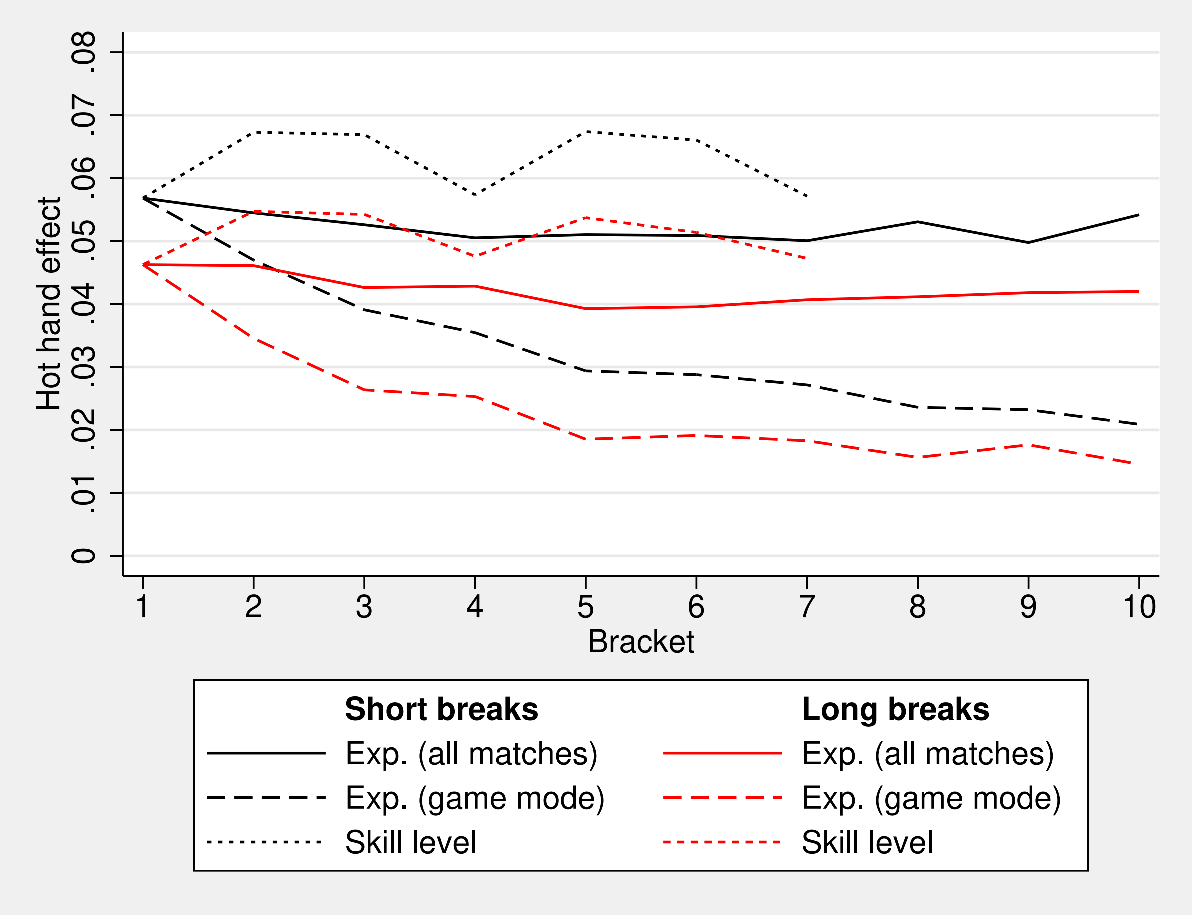
Table 3.6: Regression results: Shorter break times between matches

	Match win		XP per min		Gold per min		KDA	
Last match won	0.057^{***}	(0.005)	21.398^{***}	(1.348)	20.945^{***}	(1.163)	0.309^{***}	(0.028)
Interaction skill level								
× 1.08 to 1.41	0.010^{***}	(0.003)	0.844	(0.763)	1.162^{*}	(0.660)	0.050^{***}	(0.016)
× 1.41 to 1.83	0.010^{***}	(0.003)	0.432	(0.758)	0.978	(0.653)	0.042^{***}	(0.016)
× 1.83 to 2.19	0.001	(0.003)	-0.756	(0.748)	-0.307	(0.644)	-0.013	(0.016)
× 2.19 to 2.66	0.011^{***}	(0.003)	-0.115	(0.723)	0.751	(0.616)	0.026	(0.016)
× 2.66 to 2.95	0.009^{***}	(0.003)	1.485^{*}	(0.791)	1.651^{**}	(0.693)	0.022	(0.017)
× 2.95 and above	0.000	(0.003)	-0.164	(0.852)	-0.007	(0.731)	0.020	(0.017)
Interaction exp all matches								
× 614 to 1168	-0.002	(0.003)	-2.190^{**}	(1.087)	-2.058^{**}	(0.953)	-0.048^{**}	(0.023)
× 1168 to 1725	-0.004	(0.003)	-2.560^{**}	(1.077)	-3.355^{***}	(0.945)	-0.083^{***}	(0.023)
× 1725 to 2233	-0.006^{*}	(0.003)	-3.550^{***}	(1.087)	-3.473^{***}	(0.966)	-0.047^{**}	(0.023)
× 2233 to 2716	-0.006^{*}	(0.003)	-2.963^{***}	(1.086)	-3.517^{***}	(0.959)	-0.077^{***}	(0.022)
× 2716 to 3224	-0.006^{*}	(0.003)	-2.239^{**}	(1.082)	-3.095^{***}	(0.947)	-0.066^{***}	(0.023)
× 3224 to 3809	-0.007^{**}	(0.003)	-2.555^{**}	(1.083)	-3.502^{***}	(0.947)	-0.073^{***}	(0.023)
× 3809 to 4556	-0.004	(0.003)	-3.154^{***}	(1.123)	-3.645^{***}	(0.986)	-0.058^{**}	(0.023)
× 4556 to 5632	-0.007^{**}	(0.003)	-2.884^{**}	(1.125)	-3.768^{***}	(0.988)	-0.070^{***}	(0.023)
× 5632 and above	-0.003	(0.004)	-3.842^{***}	(1.190)	-4.303^{***}	(1.044)	-0.077^{***}	(0.024)
Interaction exp gamemode matches								
× 24 to 51	-0.010^{**}	(0.005)	-5.286^{***}	(1.346)	-5.294^{***}	(1.109)	-0.062^{**}	(0.028)
× 51 to 83	-0.018^{***}	(0.005)	-8.340^{***}	(1.293)	-7.370^{***}	(1.078)	-0.109^{***}	(0.028)
× 83 to 121	-0.021^{***}	(0.005)	-10.209^{***}	(1.268)	-9.736^{***}	(1.068)	-0.145^{***}	(0.027)
× 121 to 165	-0.027^{***}	(0.005)	-11.682^{***}	(1.268)	-10.482^{***}	(1.071)	-0.149^{***}	(0.027)
× 165 to 219	-0.028^{***}	(0.005)	-12.097^{***}	(1.255)	-10.896^{***}	(1.055)	-0.180^{***}	(0.027)
× 219 to 287	-0.030^{***}	(0.005)	-11.515^{***}	(1.252)	-11.460^{***}	(1.051)	-0.163^{***}	(0.027)
× 287 to 384	-0.033^{***}	(0.005)	-12.320^{***}	(1.256)	-11.523^{***}	(1.054)	-0.174^{***}	(0.027)
× 384 to 549	-0.034^{***}	(0.005)	-13.396^{***}	(1.258)	-12.466^{***}	(1.057)	-0.181^{***}	(0.026)
× 549 and above	-0.036^{***}	(0.005)	-12.348^{***}	(1.292)	-12.125^{***}	(1.095)	-0.196^{***}	(0.027)
Skill level								
1.08 to 1.41	0.004^{**}	(0.002)	4.159^{***}	(0.856)	8.351^{***}	(0.764)	0.138^{***}	(0.016)
1.41 to 1.83	0.005^{***}	(0.002)	3.734^{***}	(0.907)	11.014^{***}	(0.817)	0.179^{***}	(0.017)
1.83 to 2.19	0.009^{***}	(0.002)	3.076^{***}	(0.883)	12.836^{***}	(0.820)	0.205^{***}	(0.017)
2.19 to 2.66	0.011^{***}	(0.002)	5.903^{***}	(0.905)	17.944^{***}	(0.873)	0.292^{***}	(0.019)
2.66 to 2.95	0.014^{***}	(0.002)	2.873^{***}	(0.979)	19.509^{***}	(0.919)	0.308^{***}	(0.019)
2.95 and above	0.020^{***}	(0.002)	-2.181^{*}	(1.225)	20.681^{***}	(1.164)	0.372^{***}	(0.023)
Experience over all matches								
614 to 1168	-0.010^{***}	(0.002)	-6.304^{***}	(1.113)	-4.202^{***}	(0.995)	-0.029	(0.020)
1168 to 1725	-0.013^{***}	(0.002)	-7.334^{***}	(1.200)	-4.295^{***}	(1.084)	-0.005	(0.023)
1725 to 2233	-0.015^{***}	(0.002)	-9.369^{***}	(1.197)	-6.614^{***}	(1.085)	-0.035	(0.022)
2233 to 2716	-0.018^{***}	(0.002)	-9.398^{***}	(1.190)	-5.372^{***}	(1.077)	-0.021	(0.023)
2716 to 3224	-0.020^{***}	(0.002)	-10.279^{***}	(1.196)	-5.994^{***}	(1.081)	-0.038^{*}	(0.022)
3224 to 3809	-0.021^{***}	(0.002)	-12.256^{***}	(1.232)	-7.366^{***}	(1.112)	-0.076^{***}	(0.023)
3809 to 4556	-0.026^{***}	(0.002)	-10.732^{***}	(1.279)	-5.687^{***}	(1.162)	-0.067^{***}	(0.024)
4556 to 5632	-0.025^{***}	(0.002)	-13.263^{***}	(1.362)	-6.581^{***}	(1.243)	-0.055^{**}	(0.026)
5632 and above	-0.029^{***}	(0.002)	-12.847^{***}	(1.575)	-3.572^{**}	(1.467)	-0.038	(0.029)
Experience over gamemode matches								
24 to 51	0.025^{***}	(0.003)	15.865^{***}	(0.913)	13.122^{***}	(0.729)	0.139^{***}	(0.018)
51 to 83	0.039^{***}	(0.003)	23.045^{***}	(0.909)	18.889^{***}	(0.734)	0.204^{***}	(0.018)
83 to 121	0.044^{***}	(0.003)	29.840^{***}	(0.904)	24.811^{***}	(0.738)	0.245^{***}	(0.018)
121 to 165	0.056^{***}	(0.003)	35.250^{***}	(0.918)	29.088^{***}	(0.752)	0.288^{***}	(0.018)
165 to 219	0.059^{***}	(0.003)	38.620^{***}	(0.931)	31.967^{***}	(0.765)	0.311^{***}	(0.018)
219 to 287	0.066^{***}	(0.003)	41.859^{***}	(0.947)	34.782^{***}	(0.782)	0.302^{***}	(0.018)
287 to 384	0.068^{***}	(0.003)	44.886^{***}	(0.974)	36.006^{***}	(0.813)	0.289^{***}	(0.019)
384 to 549	0.070^{***}	(0.003)	48.685^{***}	(1.037)	38.257^{***}	(0.881)	0.253^{***}	(0.020)
549 and above	0.072^{***}	(0.003)	52.902^{***}	(1.262)	40.196^{***}	(1.137)	0.221^{***}	(0.023)
Playing on Radiant side	0.063^{***}	(0.001)	2.488^{***}	(0.187)	3.950^{***}	(0.154)	0.125^{***}	(0.004)
Constant	0.501^{***}	(0.011)	620.764^{***}	(1.745)	568.850^{***}	(1.763)	3.073^{***}	(0.037)
Hero FE (all heros)	Yes		No		No		No	
Hero FE (own hero)	No		Yes		Yes		Yes	
Line-up FE	Yes		Yes		Yes		Yes	
adj. R^2	0.05		0.25		0.40		0.05	
Observations	1999244		1999244		1999244		1999244	

Standard errors in parentheses (clustered on the player level)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.12: Hot hand effect for different skill and experience brackets for specifications (4) in Table 3.3 and specification (1) in Table 3.6



rience do show a similar structure but, similar to the baseline hot hand effects, also tend to be slightly larger in size.

For a more intuitive comparison of the results for the interaction terms, Figure 3.12 displays the same graphical representation of the aggregate hot hand effects for different skill and experience levels as Figures 3.9 and 3.10 in the section above. I restrict attention to the match outcome as the dependent variable. However, using player performance statistics as outcome variables shows similar patterns as can be seen from a comparison of the results shown in Tables 3.4 and 3.6. The black line in Figure 3.12 depicts the results using short breaks (specification (1) in Table 3.6) while the red lines correspond to the results allowing for longer breaks (specification (4) in Table 3.3).

While the baseline hot hand effect is substantially larger when only allowing for short match breaks, the patterns of the interactions with skill and experience level remain similar. More match experience in the ranked/all draft game mode leads to substantially lower hot hand effects and players playing on the low, medium and high skill levels also show lower hot hand effects compared to players assigned to the intermediate skill levels. It should be noted that the difference in the baseline hot hand effects for series with shorter and series allowing for longer match breaks shrinks but does not vanish when moving to higher skill and experience brackets. This is somewhat contrary to the results with longer success histories presented in the previous section. In fact, both Figures 3.9 and 3.10 show initial differences in baseline hot hand effects that shrink substantially (or do even vanish for the cold hand effect) when looking at higher (game mode) experience levels. This could be considered evidence that a hot hand effect, if it exists for more experienced players at all, is more likely to be found and of a larger magnitude when looking at series of matches that are not interrupted by long breaks.

3.7 Conclusion

In this chapter, I have used a unique dataset of match outcomes of an online game to shed further light on the existence of the so-called hot hand. I have shown that on average, a player's winning rate is about 2% higher following a win compared to a loss when focusing in matches with short breaks. The rich dataset used for this analysis allows to circumvent the discussion about sufficient statistical power of this test that has frequently emerged in the literature.

Although a hot hand effect can be found in the data, I have provided evidence that there can be stark differences in terms of the degree to which players are susceptible to the hot hand. Players playing on very low, medium and very high skill levels show lower hot hand effects than players on medium-low or medium-high skill levels. However, my results suggest that the real driving force of the hot hand is game experience. In particular, I have found that players with more game experience are much less susceptible to the hot hand. Specifically, what matters is the experience in the particular game mode rather than overall experience with the game. This highlights the importance of the setting and the environment when analyzing the hot hand effect and might be an explanation for the stark variation of results that the literature has found previously. It may also indicate why so many people are susceptible to the hot hand fallacy. Taking the example of professional sports, the experience of people in their daily lives may shape their beliefs about the functioning of a more professional environment. They may therefore believe in a hot hand effect, which they may well experience in their own more casual practice, even though this effect is not present in actual professional environments.

As a result, a casual player may well be susceptible to a strong hot hand effect and believe in it – in contrast to a professional and well-experienced athlete.

The results may therefore help explain the differences in beliefs between experts and a more casual audience. In particular in sports settings, a casual audience may experience a real hot hand in their own play and transfer the beliefs formed to the more professional setting. An expert on the other hand may act professional enough to not be affected as much by her previous performance and thus may neither show hot hand effects in her play nor expect them from other equally professional players. The much quoted hot hand fallacy in sports and other betting markets may therefore be partly due to people factoring in their own experiences too much when assessing professionals' performances.

I have also demonstrated that my results are robust to different definitions of the hot hand effect that have been used in the literature as well as a stricter definition of a match series. Although the hot (or cold) hand effect is estimated to be substantially higher when considering recent average winning rates instead of the outcomes of the previous match, distinguishing between hot and cold hand states or allowing only for shorter match breaks, the key result remains the same. The hot (or cold) hand effect vanishes with more game experience and is also lower for low, medium and very high skill levels. These results also remain robust when considering individual player performance statistics rather than the match outcome itself.

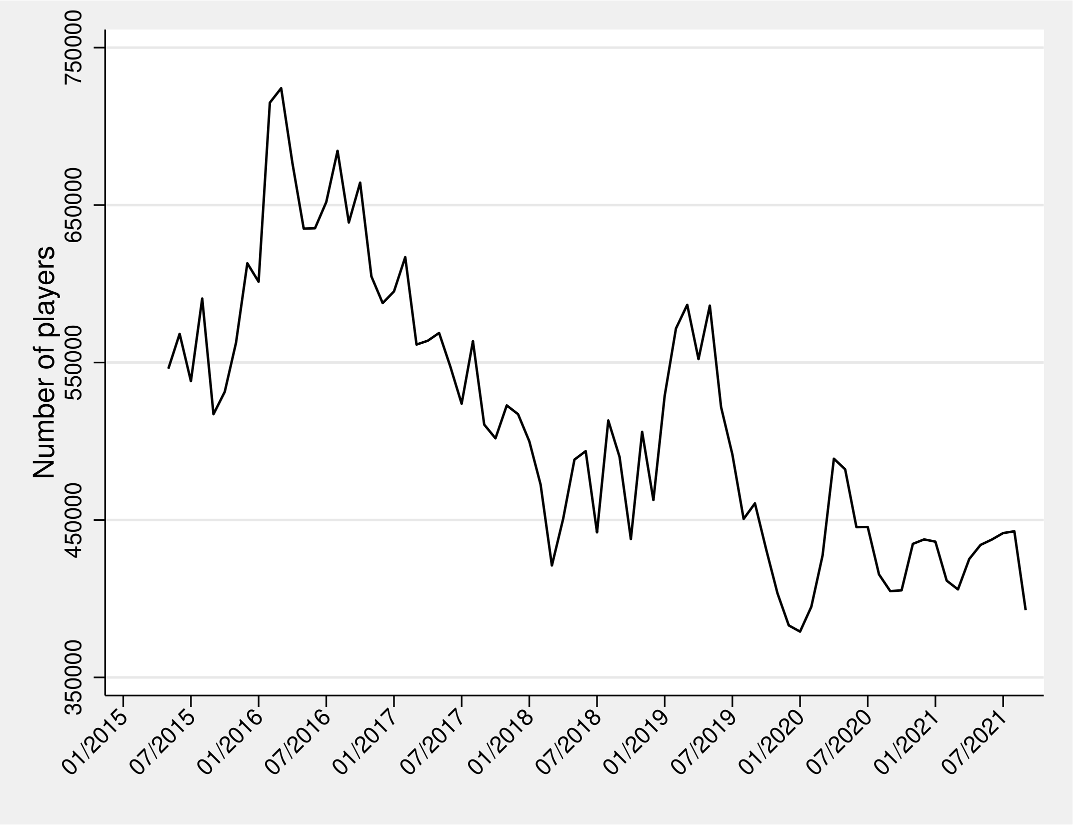
I believe that, despite the long history of the literature, there is ample room for further research in this area. In particular, I think the empirical assessment of the hot hand should be tied more strongly to the particular environment and the sample of players that is investigated. The results of this chapter demonstrate that factors like experience or skill may play a major role both for the existence and the degree to which people are susceptible to the hot hand. At the very least, researchers should be cautious of external validity of their results as the group of players that is investigated may have a major impact on the results. For the wider economic relevance of the hot hand, in particular in financial and betting market settings, it will be important to investigate if my findings transfer to this area and especially if heterogeneity plays a major role in the assessment of the hot hand in these markets.

Beyond the major driving factors identified in this chapter, namely experience with the setting and individual skill, I am convinced that it will be important to shed further light on influencing factors of the hot hand. Most of the literature has focused on providing evidence of the existence or non-existence of the hot hand in specific settings but very little is known about influencing factors that can rationalize the differences in results between settings. Previous research has occasionally pointed out stark heterogeneity of the hot hand effect but, to the best of my knowledge, has not tried to link this heterogeneity to observable characteristics. This chapter has taken a further step by identifying two driving factors of the hot hand. Notably, what is still missing in my view is evidence of a more direct correspondence between people's beliefs about the hot hand and their own experiences, which in turn may influence their beliefs about professional play and could, for example, influence odds on betting markets. This chapter has taken a first step in highlighting that experiences may well differ within the same setting for people playing on different levels of skill and with different experience. However, more direct evidence on the link between own experiences and people's behavior may shed further light on patterns observed in betting and financial markets.

3.8 Appendix

In this appendix I provide a more detailed description of Dota 2 and its rules and mechanics. Dota 2 started out as a community-built map for the real-time strategy game Warcraft III. The game quickly became a fan favorite and later developed into a game of its own under the ownership of Valve software. The evolution of the monthly average number of active players as a measure of popularity is shown in Figure 3.13. Even though popularity has decreased over recent years, the game remains extremely popular nowadays with on average over 400,000 active players at the same time. Dota 2 has also one of the largest professional player scenes with regular tournaments that can have price pools in the range of multiple million dollars.[38]

Figure 3.13: Dota 2 monthly average player numbers over time

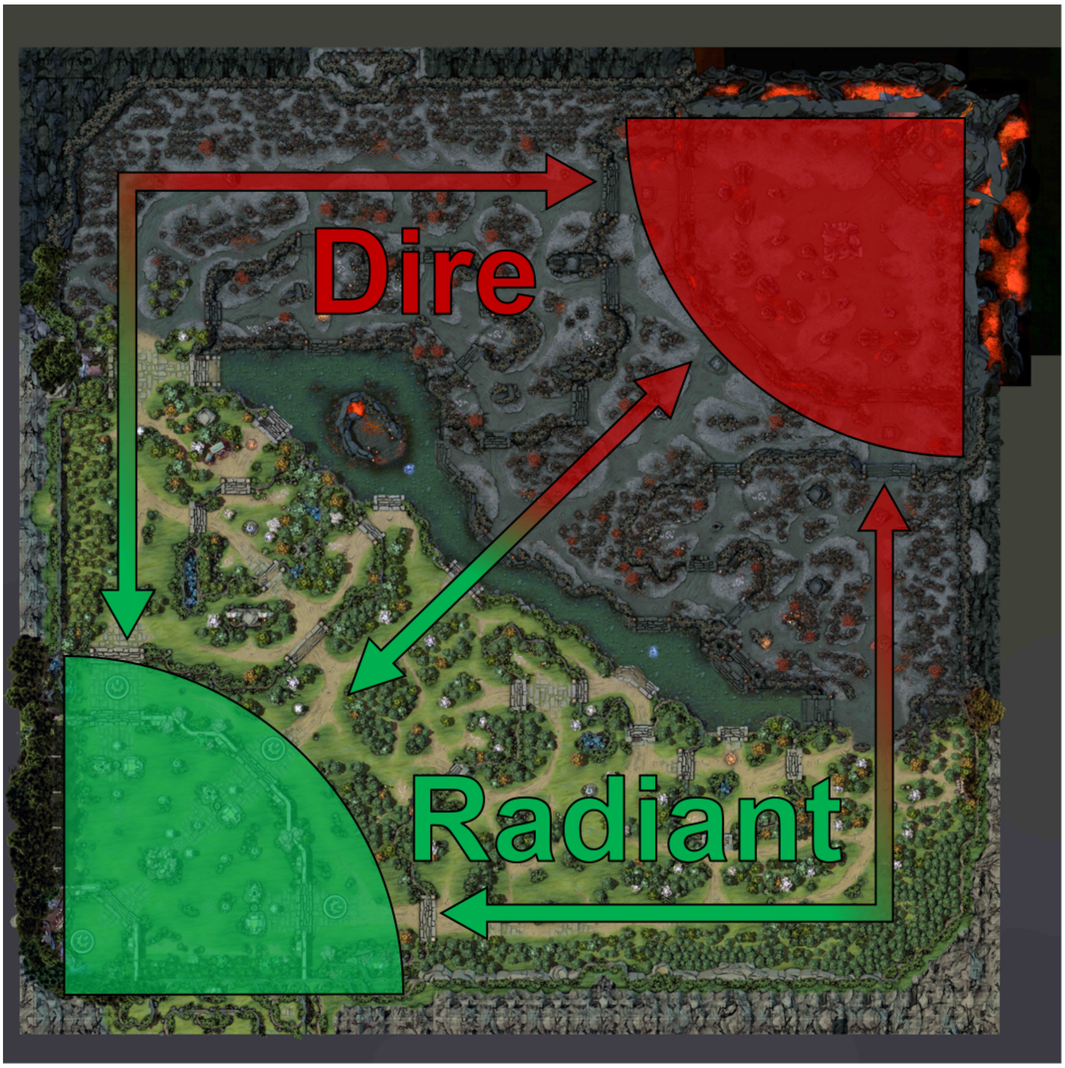


Source: www.githyp.com/dota-2-100988 (accessed on 19/10/2021)

The core concept of Dota 2 (and similar) games is based around arena fights between teams of usually five players with the eventual goal to destroy the opposing team's

[38] The tournament "The International 2019" for instance had a price pool of $34,330,069. See www.esportsearnings.com/comparisons/zcdd-dota-2-vs-lol/largest_tournaments (accessed on 28/06/2022).

Figure 3.14: Illustration of Dota 2 map with team bases and attack lanes



base whilst defending the own. Each player controls a so called "hero" with a unique set of abilities and a certain play-style. During each match players are trying to gain advantages on the map by improving their hero through gathering gold and experience points until eventually one team overcomes the other.[39] The result of each game is either a team loss or team win with no draws being possible. Figure 3.14 displays the map on which matches are played. There are three lines of attack which connect a teams home base with the opponents base. The teams are labeled "Radiant" (green) and "Dire" (red) and are located in opposite corners of the map. To which position a player or teams gets assigned is decided by chance before the match starts. The same is true for

[39] The exact mechanics are hugely complex and do not only evolve around perfecting the play-style with a chosen hero but also around the strength and weaknesses of each hero as well as synergies with heroes on the same team and counters of heroes in opposing teams.

the assignment of players to each other except they deliberately queue for a match as a pre-made group, in which case they always play together in a team.

Whilst a single player's performance certainly affects the probability of success, the outcome will eventually depend on the performance of all other players, which induces a certain degree of randomness in match outcomes from the perspective of the individual player. Independent of the skill players show in the game, the long-run baseline winning rate should be in a relatively narrow range for all players in the sample. This is a result of the "match making algorithm" which aims to match the a player with team members and opponents of equal strength.[40]

[40] The concept operates via an internal "MMR" rating that represents a player's skill. The "MMR" increases with wins and declines with losses and is essentially an Elo rating system. Elo rating systems are a method to calculate the relative strength of players and are frequently used in sports (e.g. in chess). Essentially, a player is matched to relatively weaker players if she wins too little and relatively stronger players if she wins too much.

Bibliography

Anderson, Eric T. and Duncan I. Simester (2014). “Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception.” In: *Journal of Marketing Research* 51.3, pp. 249–269.

Anderson, Michael and Jeremy Magruder (2012). “Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database.” In: *The Economic Journal* 122, pp. 957–989.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press.

Avugos, Simcha, Michael Bar-Eli, Ilana Ritov, and Eran Sher (2013). “The elusive reality of efficacy–performance cycles in basketball shooting: an analysis of players’ performance under invariant conditions.” In: *International Journal of Sport and Exercise Psychology* 11.2, pp. 184–202.

Bachelier, Louis (1900). *Théorie de la spéculation*. Gauthier-Villars.

Bajari, Patrick and Ali Hortaçsu (2004). “Economic Insights from Internet Auctions.” In: *Journal of Economic Literature* 42.2, pp. 457–486. arXiv: arXiv:1011.1669v3.

Bar-Eli, Michael, Simcha Avugos, and Markus Raab (2006). “Twenty years of “hot hand” research: Review and critique.” In: *Psychology of Sport and Exercise* 7.6, pp. 525–553.

Cabral, Luis (2012). “Reputation on the Internet.” In: *The Oxford handbook of the digital economy*, pp. 343–354.

Cabral, Luis and Ali Hortaçsu (2010). “The Dynamics of Seller Reputation: Evidence from eBay.” In: *The Journal of Industrial Economics* 58.1, pp. 54–78.

Cabral, Luis and Lingfang Li (2015). “A dollar for your thoughts: Feedback-conditional rebates on eBay.” In: *Management Science* 61.9, pp. 2052–2063.

Cai, Hongbin, Yuyu Chen, and Hanming Fang (2009). “Observational learning: Evidence from a randomized natural field experiment.” In: *American Economic Review* 99.3, pp. 864–82.

Cain, Michael, David Law, and David Peel (2003). "The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets." In: *Bulletin of Economic Research* 55.3, pp. 263–273.

Camerer, CF (1989). "Does the Basketball Market Believe in the Hot Hand?" In: *The American Economic Review* 79.5, pp. 1377–1386.

Carnehl, Christoph, Maximilian Schaefer, André Stenzel, and Kevin Ducbao Tran (2021a). "Value for Money and Selection: How Pricing Affects Airbnb Ratings."

Carnehl, Christoph, Andre Stenzel, and Peter Schmidt (2021b). "Pricing for the Stars."

Chamberlain, Gary (1982). "Multivariate regression models for panel data." In: *Journal of Econometrics* 18.1, pp. 5–46.

Chen, Pei-Yu, Shin-yi Wu, and Jungsun Yoon (2004). "The Impact of Online Recommendation and Consumer Feedback on Sales." In: *Proceeding of the International Conference on Information Systems* Paper 58, pp. 711–724.

Chevalier, Judith A and Dina Mayzlin (2006). "The Effect of Word of Mouth on Sales : Online Book Reviews." In: *Journal of Marketing Research* XLIII.August, pp. 345–354.

Clark, Russell D (2003a). "An analysis of streaky performance on the LPGA tour." In: *Perceptual and Motor Skills* 97.2, pp. 365–370.

— (2003b). "Streakiness among professional golfers: Fact or fiction?" In: *International Journal of Sport Psychology.*

— (2005). "Examination of hole-to-hole streakiness on the PGA tour." In: *Perceptual and motor skills* 100.3, pp. 806–814.

Connolly, Robert A and Richard J Rendleman Jr (2008). "Skill, luck, and streaky play on the PGA tour." In: *Journal of the American Statistical Association* 103.481, pp. 74–88.

Croson, Rachel and James Sundali (May 2005). "The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos." In: *Journal of Risk and Uncertainty* 30.3, pp. 195–209.

Croxson, Karen and J James Reade (2011). "Exchange vs. dealers: a high-frequency analysis of in-play betting prices." In:

— (2013). "Information and efficiency: Goal arrival in soccer betting." In: *The Economic Journal.*

Cutmore, Adam CJ and William J Knottenbelt (2013). "Quantitative models for retirement risk in professional tennis." In: *Proceedings of the 4th International Conference on Mathematics in Sport.*

De Long, J Bradford, Andrei Shleifer, Lawrence H Summers, and Robert J Waldmann (1991). "The survival of noise traders in financial markets." In: *Journal of Business*, pp. 1–19.

Dellarocas, Chrysanthos (2006). "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms." In: *Management Science* 52.10, pp. 1577–1593.

DellaVigna, Stefano (2009). "Psychology and economics: Evidence from the field." In: *Journal of Economic Literature*, pp. 315–372.

Dickinger, Astrid and Josef Mazanec (2008). "Consumers' Preferred Criteria for Hotel Online Booking." In: *Information and Communication Technologies in Tourism 2008*, pp. 244–254.

Dittrich, Johannes (2013). "Excess Volatility in Belief Streams: Evidence from High Frequency Sports Data." MA thesis. University of Mannheim, pp. 1–33.

Dorsey-Palmateer, Reid and Gary Smith (2004). "Bowlers' hot hands." In: *The American Statistician* 58.1, pp. 38–45.

Dranove, David and Ginger Zhe Jin (2010). "Quality Disclosure and Certification: Theory and Practice." In: *Journal of Economic Literature* 48.4, pp. 935–963.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston (2008). "Do online reviews matter? - An empirical investigation of panel data." In: *Decision Support Systems* 45.4, pp. 1007–1016.

Easton, Stephen and Katherine Uylangco (2010). "Forecasting outcomes in tennis matches using within-match betting markets." In: *International Journal of Forecasting* 26.3, pp. 564–575.

Fama, Eugene F (1970). "Efficient capital markets: A review of theory and empirical work." In: *The journal of Finance* 25.2, pp. 383–417.

Forman, Chris, Anindya Ghose, and Batia Wiesenfeld (2008). "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets." In: *Information Systems Research* 19.3, pp. 291–313.

Forrest, David and Ian McHale (2007). "Anyone for tennis (betting)?" In: *The European Journal of Finance* 13.8, pp. 751–768.

Forrest, David and Robert Simmons (2008). "Sentiment in the betting market on Spanish football." In: *Applied Economics* 40.1, pp. 119–126.

Franck, Egon, Erwin Verbeek, and Stephan Nüesch (2011). "Sentimental preferences and the organizational regime of betting markets." In: *Southern Economic Journal* 78.2, pp. 502–518.

Gilovich, Thomas, Robert Vallone, and Amos Tversky (1985). "The hot hand in basketball: On the misperception of random sequences." In: *Cognitive psychology* 17.3, pp. 295–314.

Green, Brett and Jeffrey Zwiebel (2018). "The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? Evidence from major league baseball." In: *Management Science* 64.11, pp. 5315–5348.

Grimes, Paul W (2002). "The Overconfident Principles of Economics Student: An Examination of a Metacognitive Skill." In: *Journal of Economic Education* 33.1, pp. 15–30.

Guryan, Jonathan and Melissa S Kearney (2008). "Gambling at lucky stores: Empirical evidence from state lottery sales." In: *American Economic Review* 98.1, pp. 458–73.

Hayek, F. A. (1945). "The Use of Knowledge in Society.pdf." In: *American Economic Review* 35.4, pp. 519–530.

Hilger, James, Greg Rafert, and Sofia Villas-Boas (2011). "Expert opinion and the demand for experience goods: an experimental approach in the retail wine market." In: *Review of Economics and Statistics* 93.4, pp. 1289–1296.

Hu, Nan, Indranil Bose, Noi Sian Koh, and Ling Liu (2012). "Manipulation of online reviews: An analysis of ratings, readability, and sentiments." In: *Decision Support Systems* 52.3, pp. 674–684.

Huang, Xinzhuo, William Knottenbelt, and Jeremy Bradley (2011). "Inferring tennis match progress from in-play betting odds." In: *Final year project), Imperial College London.*

Jin, Ginger Zhe and Andrew Kato (2006). "Price, Quality and Reputation: Evidence from an Online Field Experiment." In: *The RAND Journal of Economics* 37.4, pp. 983–1005.

Klaassen, Franc J G M and Jan R Magnus (2001). "Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model." In: *Journal of the American Statistical Association* 96.454, pp. 500–509.

— (2003). "Forecasting the winner of a tennis match." In: *European Journal of Operational Research* 148.2, pp. 257–267.

Klein, Tobias J, Christian Lambertz, and Konrad O Stahl (2016). "Market Transparency, Adverse Selection, and Moral Hazard." In: *Journal of Political Economy* 124.6.

Koehler, Jonathan J and Caryn A Conley (2003). "The "hot hand" myth in professional basketball." In: *Journal of sport and exercise psychology* 25.2, pp. 253–259.

Kou, Yubo, Yao Li, Xinning Gui, and Eli Suzuki-Gill (2018). "Playing with streakiness in online games: how players perceive and react to winning and losing streaks in

League of Legends." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Kuypers, Tim (2000). "Information and efficiency: an empirical study of a fixed odds betting market." In: *Applied Economics* 32.11, pp. 1353–1363.

Lahvicka, Jiri (2014). "What causes the favourite-longshot bias? Further evidence from tennis." In: *Applied Economics Letters* 21.2, pp. 90–92.

Levitt, Steven D (2004). "Why are gambling markets organised so differently from financial markets?" In: *The Economic Journal* 114.495, pp. 223–246.

Li, Lingfang (Ivy) and Erte Xiao (2014). "Money Talks: Rebate Mechanisms in Reputation System Design." In: *Management Science* August 2014.

Luca, Michael (2016). "Reviews, reputation, and revenue: The case of Yelp. com."

Luca, Michael and Oren Reshef (2021). "The effect of price on firm reputation." In: *Management Science* 67.7, pp. 4408–4419.

Luca, Michael and Georgios Zervas (2016). "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." In: *Management Science* 62.12, pp. 3412–3427.

Lucking-Reiley, David, Doug Bryan, Naghi Prasad, and Daniel Reeves (2007). "Pennies from eBay: The Determinants of Price in Online Auctions." In: *The Journal of Industrial Economics* 55.2, pp. 223–233.

Malmendier, Ulrike and Geoffrey Tate (2005). "CEO overconfidence and corporate investment." In: *The Journal of Finance* 60.6, pp. 2661–2700.

— (2008). "Who makes acquisitions? CEO overconfidence and the market's reaction." In: *Journal of Financial Economics* 89.1, pp. 20–43.

Martin, Simon and Sandro Shelegia (2021). "Underpromise and overdeliver?-Online product reviews and firm pricing." In: *International Journal of Industrial Organization* 79, p. 102775.

Mayzlin, By Dina, Yaniv Dover, and Judith Chevalier (2014). "Promotional Reviews : An Empirical Investigation of Online Review Manipulation." In: *American Economic Review* 104.8, pp. 2421–2455.

Melnik, Mikhail I and James Alm (2002). "Does a Seller's Ecommerce Reputation Matter? Evidence from eBay Auctions." In: *The Journal of Industrial Economics* 50.3, pp. 337–349.

Miller, Joshua B and Adam Sanjurjo (2018a). "A cold shower for the hot hand fallacy: Robust evidence that belief in the hot hand is justified." In:

— (2018b). "Surprised by the hot hand fallacy? A truth in the law of small numbers." In: *Econometrica* 86.6, pp. 2019–2047.

Miller, Joshua B and Adam Sanjurjo (2021). “Is it a fallacy to believe in the hot hand in the NBA three-point contest?” In: *European Economic Review* 138, p. 103771.

Moreno, Antonio and Christian Terwiesch (2014). “Doing Business with Strangers : Reputation in Online Service Marketplaces Doing Business with Strangers : Reputation in Online Service Marketplaces.” In: *Information Systems Research* 25.Dezember 2014, pp. 865–886.

Moretti, Enrico (2011). “Social learning and peer effects in consumption: Evidence from movie sales.” In: *Review of Economic Studies* 78.1, pp. 356–393.

Morris, C (1977). “The most important points in tennis.” In: *Optimal strategies in sport* 5, pp. 131–140.

Mundlak, Yair (1978). “On the Pooling of Time Series and Cross Section Data.” In: *Econometrica* 46.1, pp. 69–85.

Narayanan, Sridhar and Puneet Manchanda (2012). “An empirical analysis of individual level casino gambling behavior.” In: *Quantitative Marketing and Economics* 10.1, pp. 27–62.

Nelson, Phillip (1970). “Information and consumer behavior.” In: *Journal of Political Economy* 78.2, pp. 311–329.

Odean, Terrance (1999). “Do Investors Trade Too Much?” In: *The American Economic Review* 89.5, pp. 1279–1298.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T Hancock (2011). “Finding deceptive opinion spam by any stretch of the imagination.” In: *arXiv preprint arXiv:1107.4557*.

Page, Lionel (2009). “Is there an optimistic bias on betting markets?” In: *Economics Letters* 102.2, pp. 70–72.

Raab, M (2002). “Hot hand in sports—The belief in hot hand of spectators in volleyball.” In: *ECSS proceedings* 2, p. 971.

Rabin, Matthew (2002). “Inference by believers in the law of small numbers.” In: *The Quarterly Journal of Economics* 117.3, pp. 775–816.

Rabin, Matthew and Dimitri Vayanos (2010). “The gambler’s and hot-hand fallacies: Theory and applications.” In: *The Review of Economic Studies* 77.2, pp. 730–778.

Reinstein, David A and Christopher M Snyder (2005). “The Influence of Expert Reviews on Consumer Demand for Experience Goods : A Case Study of Movie Critics.” In: *The Journal of Industrial Economics* 53.1, pp. 27–51.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood (2006). “The value of reputation on eBay: A controlled experiment.” In: *Experimental Economics* 9.2, pp. 79–101.

Sauer, Raymond D et al. (1998). "The economics of wagering markets." In: *Journal of economic Literature* 36.4, pp. 2021–2064.

Shampanier, Kristina, Nina Mazar, and Dan Ariely (2007). "Zero as a special price: The true value of free products." In: *Marketing Science* 26.6, pp. 742–757.

Smith, Gary (2003). "Horseshoe pitchers' hot hands." In: *Psychonomic bulletin & review* 10.3, pp. 753–758.

Snowberg, Erik and Justin Wolfers (2007). ""The Favorite-Longshot Bias: Understanding a Market Anomaly." In: *Efficiency of Sports and Lottery Markets, edited by Donald Hausch and William Ziemba. Elsevier: Handbooks in Finance series.*

— (2010). "Explaining the favorite–long shot bias: Is it risk-love or misperceptions?" In: *Journal of Political Economy* 118.4, pp. 723–746.

Sundali, James and Rachel Croson (2006). "Biases in casino betting: The hot hand and the gambler's fallacy." In: *Judgement and Decision Making* 1.1, p. 1.

Thaler, Richard H and William T Ziemba (1988). "Anomalies: Parimutuel betting markets: Racetracks and lotteries." In: *The Journal of Economic Perspectives* 2.2, pp. 161–174.

Tversky, Amos and Thomas Gilovich (1989a). "The "hot hand": Statistical reality or cognitive illusion?" In: *Chance* 2.4, pp. 31–34.

— (1989b). "The cold facts about the "hot hand" in basketball." In: *Chance* 2.1, pp. 16–21.

Vaughan Williams, Leighton and David Paton (1998). "Why are some favourite-longshot biases positive and others negative?" In: *Applied Economics* 30.11, pp. 1505–1510.

Vermeulen, Ivar E. and Daphne Seegers (2009). "Tried and tested: The impact of online hotel reviews on consumer consideration." In: *Tourism Management* 30.1, pp. 123–127.

Wooldridge, Jeffrey M (2010). *Econometric analysis of cross section and panel data.* MIT press.

— (2019). "Correlated random effects models with unbalanced panels." In: *Journal of Econometrics* 211.1, pp. 137–150.

Wooldridge, Jeffrey Marc (2005). "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity." In: *Journal of Applied Econometrics* 20.1, pp. 39–54.

Xie, Sihong, Guan Wang, Shuyang Lin, and Philip S Yu (2012). "Review spam detection via temporal pattern discovery." In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 823–831.

Ye, Qiang, Rob Law, Bin Gu, and Wei Chen (2011). "The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings." In: *Computers in Human Behavior* 27.2, pp. 634–639.

Yuan, Jia, Guang-Zhen Sun, and Ricardo Siu (2014). "The lure of illusory luck: How much are people willing to pay for random shocks." In: *Journal of Economic Behavior & Organization* 106, pp. 269–280.

Zegners, Dainis (2019). "Building an Online Reputation with Free Content: Evidence from the E-book Market."

Curriculum Vitae

2013–2022	Ph.D. *Economics*	University of Mannheim
2011–2012	Visiting Student *Economics Department*	University of California at Berkeley
2011–2013	Master of Science *Economic Research*	University of Mannheim
2008–2011	Bachelor of Science *Economics*	University of Mannheim