UNIVERSITÀ COMMERCIALE LUIGI BOCCONI
PhD School

PhD Program in Economics and Finance
Cycle: 32
Disciplinary Field (code): SECS-P/06

Essays in Applied Economics

Advisor: Marco OTTAVIANI

PhD Thesis by: Marco WALLNER
ID: 3029065

Year 2022

Abstract

During my studies and research I have always been fascinated by the variety of problems economists can study using empirical and theoretical methods. Understanding how incentives, the (non)-availability of information, or strategic interaction shape outcomes in many areas of life is the key goal of my academic work. Naturally, many of these ideas emerged from studying the functioning of markets or situations in which market failures lead to inefficiencies. However, applying economic insights to various aspects of society has opened up a rich set of research questions. The very fact that many of these aspects have an impact on important areas of society has motivated me to research in different areas, applying economic thinking and methodology. This view is reflected in the choice of research questions for this thesis. It comprises a study of how the support of crowds, in theory mostly irrelevant to a fully rational agent, affects the performance of professional football players. I argue that choking under pressure is a relevant dimension in such environments. I consequently study the role of betting markets in pricing novel insights on the performance of such athletes in the absence of crowds due to the coronavirus pandemic. I also study a crucial field of modern societies that is, due to the presence of externalities, largely organized outside of markets: academic research. I contribute to the quantitative understanding of the role of information acquisition across different scientific fields.

In the first chapter, I show that crowd presence may be a two-edged sword not benefiting all teams equally – some instead seem display what is known as choking under pressure. This finding is important for labor markets, incentive theory and other fields where human agents are usually incentivized under the assumption that stronger incentives lead to higher effort and better outcomes. In Chapter 2, I document that this effect is not adequately incorporated into market prices by bookmakers, contradicting market efficiency – one of the building blocks of financial theory. In the third chapter, I discuss the important policy problem of allocating scarce resources to different research projects of ex-ante unobservable quality. Commonly, expert knowledge is used to identify projects considered fund-worthy. I document empirically that experts across scientific fields differ with respect to their ability in reliably evaluating the merit of their peers' work. In particular, scholars in natural sciences display higher agreement and less noisy evaluation ability than their colleagues from other fields.

Chapter 1

Performance, Pressure and the Role of the Crowd: An Empirical Investigation of the Home Advantage in European Football during the Covid-19 Pandemic[†]

[†]Author: Marco Wallner, wallner.marco@phd.unibocconi.it

Abstract

As a consequence of the Covid-19 pandemic, a sizable fraction of European football matches of the 2019/2020 season was played in empty stadiums. This unique natural experiment-like situation allows to identify the effect of supporters on home-team performance in a controlled way, distinguishing it from a home-advantage induced by simply playing in one's own stadium. Perhaps surprisingly, I do not find a negative effect of matches behind closed doors on the performance of the home team in all leagues. Instead, some teams made the most out of crowd-absence and preserved the home field advantage while others did suffer from an almost complete loss thereof. Looking into match statistics such as shots or goals reveals that the effects are almost exclusively driven by performance measures of the home side. Regarding the heterogenous findings, a possible explanation is a negative effect of crow-induced social pressure on players. Strikingly, in Spain where the home field advantage is most pronounced before the Coronavirus outbreak, the effect of crowd-absence is estimated to be the highest corroborating that these teams might have better dealt with the pressure before and hence do not benefit equally from a lack of it. Another finding supporting this hypothesis is the decline in shots for some teams that is not followed by a decline in (expected) goals indicating that shots are only undertaken in more promising situations. This hypothesis is further substantiated by the conversion of penalties. Penalties under high pressure result in goals almost 10 percentage points less likely compared to penalties under lower pressure. This effect is, however, only present when supporters are admitted to the stadium.

1 Introduction

Professionals, scholars and fans around the world alike, have long taken for granted the existence of a home advantage in sports competitions, in particular in football. The literature has previously suggested various explanations for this phenomena. An exhaustive overview can be found in Pollard (2008) . Possible causes can largely be classified into two channels: (i) the advantage of playing at home, hence, enjoying a familiar environment, short travel etc. and (ii) the dominance of home supporters in the stadium, creating a performance-enhancing atmosphere through, for example, increased motivation for the home side or biasing the referee towards the home side.

This paper aims at empirically disentangling the relative importance of the latter channel exploiting a unique natural experiment-like situation. Due to the Covid-19 pandemic, most countries prohibited mass gatherings including crowded stadiums. As a consequence, major football competitions were forced to hold matches behind closed doors.

I use data on the English Premier League, the Italian Serie A, the Spanish La Liga, and the German Bundesliga in order to analyze how home team performance has changed following the introduction of matches behind closed doors. Given the experimental character of this policy, arguably introduced independently of performance in sports competitions, I can identify the effect of the crowd using different regression techniques.

I find mixed evidence on the relative importance of the components for the home advantage. In the Italian and English leagues, merely playing at home has a positive and significant effect on the prospects of the home team, while additionally the presence of supporters does not significantly improve performance. Put differently, the ban of crowds did, perhaps surprisingly, not impact the hosting team negatively

in all cases. In Spain and Germany on the other hand, the home advantage almost entirely disappeared when home teams played without their supporters. In these leagues, factors such as playing at a familiar territory or the lack of possible fatigue due to travel did only improve the performance to a relatively small extent. In England and Italy rather the opposite is true: supporters did not matter significantly and after the ban of crowds, English and Italian home teams continued to outperform their opponents by a similar magnitude as before.

I suggest that the pressure on players, created by the atmosphere of a crowded stadium, might be an important driver of this result. If teams experience different levels of mostly harmful pressure by the audience, or alternatively are able to cope with such pressure differently, this may explain the observed differences. Overall, the heterogenous findings make it clear that crowd-support is not unambiguously positive for performance. Rather country specific, demographic or cultural differences may be required to understand why some teams seem to benefit from a crowd while others do not (or are even harmed by it).[1]

The suggested role of pressure is also reflected in the conversion rate of penalties. Penalties taken under high pressure conditions are converted significantly less likely than penalties with lower pressure. Interestingly the effect is more pronounced for home teams and is not present when supporters are banned from the stadium.

Related Literature

The home advantage has been studied by scholars from different fields over many decades. In an early article, Schwartz and Barsky (1977) document home advantage

[1]The topic has also received some attention in the media and simple descriptive results show strikingly heterogenous effects. See for example `https://football-observatory.com/IMG/sites/b5wp/2020/wp304/en/`

for a variety of sports in the US and reason that crowd support as opposed to familiarity to the stadium or fatigue of the away team is the main driver of it. Over the years, an enormous amount of papers has taken a closer look at the underlying causes of the repeatedly confirmed home advantage. Overall, there exists evidence for both previously described channels: a home advantage due to crowd-support as well as due to the comfort/ benefit of simply playing at home.[2]

Regarding crowd-support, Ponzo and Scoppa (2018) do find evidence for its importance using same city derbies, thereby eliminating any territorially related factors . Garicano, Palacios-Huerta, and Prendergast (2005), among others, show that this might at least partially be due to referee bias induced by the crowd. At the same time there exists also evidence that the players themselves are affected, e.g. display altered levels of testosterone (see Neave and Wolfson (2003)).

On the other hand, various studies also document evidence in support of the second channel. Oberhofer, Philippovich, and Winner (2010) find that travel distance plays a role for the magnitude of the home advantage in Germany, with a stronger home advantage materializing when the away team needs to travel longer distances. To sum up, the home advantage is a well documented phenomena with plenty of underlying, yet (quantitatively) poorly understood causes. A broad summary of the empirical and theoretical debate can be found in Pollard (2008).

It is important to stress that both channels may well exist simultaneously. Indeed, there exists convincing evidence for both explanations to matter. With studies being conducted using mainly observational data, it is naturally hard to generalize the findings and in particular to quantify their relative importance. Same city derbies for example are particular types of matches whose findings may not be externally valid. Having said this, it is of little surprise that the historically unique widespread

[2]It is possible to further disentangle these two channels, as for example in Fischer and Haucap (2020), but in the rest of the paper I distinguish only between these two channels.

introduction of matches behind closed doors due to the Covid-19 pandemic has attracted a lot of interest in the literature aimed at understanding precisely the effect of the crowd as opposed to other sources of home advantage.

Fischer and Haucap (2020) document a significant reduction in the home advantage for teams from the first German division. Lower tier teams from the second and third division, however, did not experience a comparable reduction in the home advantage. They ascribe this to the previously lower attendance in these leagues that enabled a faster habituation to the new conditions. In particular, the first division league teams, after some match days and with growing experience, managed to re-establish the old level of home advantage.

Similarly, Dilger and Vischer (2020) document that the bias in favor of the home team disappeared in Germany's first division when playing without crowd. While they do not find significant effects on measures of performance (e.g. passing accuracy), they additionally are able to document a change in the referee's behavior. Home teams are found to no longer receive fewer yellow cards than the away team and that extra time (in the decisive second half) is reduced.

On a related note, Bryson, Dolton, Reade, Schreyer, and Singleton (2020) in a sample of 23 leagues from 17 countries find a reduction of the home-away gap in the distribution of yellow cards. The home team receives more yellow cards while the away team receives less. Effects on outcomes such as the final result or the goal difference are, however, overall not significant. Similar results regarding the referee bias are obtained in a study by Endrich and Gesche (2020) controlling for some dimensions of player behavior.

That social pressure is not only relevant for referees but also for (home team) players, is argued for in Scoppa (2020). He documents that across five major European leagues home teams secure around 0.2 points less and score 0.15 less goals relative to the away side when playing without crowd.

My paper adds to this existing literature in two important dimensions. First, I emphasize cross-country differences and find heterogenous effects contradicting a general disadvantage from crowd absence as previously documented. Second, I explicitly take into account some outcome measures other than the pure final result. This allows me to explore in greater detail the role of the crowd going beyond solely match results, which are a quite noisy measure due to the sizable portion of randomness in scoring/ defending goals. This allows me to discuss why teams were affected differently and some of its possible causes. Doing so, I am able to show that crowd absence affected, if at all, the outcomes almost entirely due to altered performance measures of the home team. Away teams are characterized by identical performance measures as before.

Contrarily to some of the related literature, I do not take into account the role of the referee explicitly. I recognize the importance of a referee bias for the outcome of matches and its relation to a possible disappearance of the home bias. Given that all four leagues in my sample use a Video Assistant Referee, the role of the referee is limited by the fact that decisive and wrong decisions are corrected after having been reviewed.[3] Further, it remains unclear whether differences in observed referee actions are due to altered player behavior or actual change in the referee's behavior.

The paper is organized as follows. Section 2 briefly describes the data used throughout the remaining sections. Section 3 estimates the home advantage separately from supporter presence controlling for a variety of other factors such as team quality. In Section 4, I address one possible explanation of the reults and further decompose the findings of Section 3. To substantiate the claim of social pressure, I analyze data on penalty conversion in section 5. Section 6 concludes.

[3]Importantly, red cards, goals, and (possible) penalties are always checked by reviewing the respective scene. Therefore, all such decisions should be decided correctly when abstracting from mistakes in the review process.

2 Data

Data on football matches are easily available. I use data on match outcomes from the four first divisions in Germany, England, Italy, and Spain. These four leagues are, together with the first division from France, regularly considered the most competitive leagues and certainly attract the most attention from supporters as well as media. Besides being comparable in terms of size of the country, these leagues were also found to be characterized by a generally similar level of home-advantage (see Pollard and Gómez (2014)). I focus mainly on the 2019/2020 season. Match results together with additional information on each match are downloaded from `https://www.football-data.co.uk/data.php`. Importantly, the data contain the final scores and shots (*shots*) by the home and away teams.

Because football matches are characterized by a high degree of unpredictability and a general low number of goals, I include another dimension of outcome: expected goals (xG), as computed by 538.com.[4] This variable, computed ex-post, attempts to capture the expected number of goals a team should have scored in a given match according to the course of the match. It is computed using the positioning of the attacking and defending team, the distance to the goal, and other relevant factors of each single goal-scoring attempt. As such, it is less subject to randomness. A promising chance is recorded as such irrespective of its outcome (goal/ no-goal). Similarly, a lucky strike that counts as one entire goal is given a relatively low weight. It is not an ex-ante prediction on the expected value of goals prior to the match but rather can be considered a sophisticated way of counting shots after (or during) the match, giving high weights to promising chances and vice versa. Put differently, this measure better captures how teams have actually

[4]Data were retrieved from `https://github.com/fivethirtyeight/data/tree/master/soccer-spi`.

performed because it abstracts from the random component in goals scored but rather counts the attempts, weighted by how promising they are. With respect to simply counting the number of shots, it is more advanced due to its consideration of each shot's quality.

These outcome-data for each match are complemented with additional information regarding each team's quality. Following other studies (e.g. Fischer and Haucap (2020)), I include information from `https://www.transfermarkt.com` regarding a team's market value (*market value*). The market value of a team is the combined market value of their players. A player's market value is a community-based estimated about how much a player is worth, given the industry convention. Using a stock market analogy, it is derived from obtaining dividends (his performances) for the duration of a contract/ over his career. Due to the rare transaction of players, market values rely on expert judgement compared to actual transaction date. Real transaction data play the role of informing about the industry convention.[5]

For the season of 2019/2020, I consider market values declared on August 15, 2019 – approximately the start of the respective seasons.[6]

Additionally, I compute the current form of a team as the points obtained over the last three matches ($form_{recent}$).[7] I also compute the average points per match so far ($form_{season}$) to capture the overall form and the days since the last league match (*rest*). In order to capture diverging incentives to win a match, I also include

[5]There has been a (slight) decrease in the overall price level of market values after the pandemic shock due to limited revenue from clubs and high uncertainty. I ignore this aspect as I collect data once and for all in the beginning of the season. Eventual changes in market prices are assumed to not alter the performance of players.

[6]I do not continuously update the market value throughout the season. Given the experimental character of my variable of interest, this seems to be a minor problem. Further, the changes in market value throughout the season (due to performance or transfers) are rather small. I do address the issue of continuously changing quality in the Appendix using SPI (described in the end of section 2) as a measure of quality.

[7]One might argue that this is too simplistic as it does not account for the strength of the opponent over the last three matches. The variable SPI does that and hence can be used likewise. The results on the coefficients of interest do not change in a meaningful way.

a match-specific index for importance of the match for each team (*importance*).[8] In the Appendix, I use the continuously updated variable Soccer Power Index Soccer Power Index (*SPI*) from 538.com to control for quality differences as an alternative to market value.[9]

In order to proxy the atmosphere and size of the various stadiums, I include the capacity of all stadiums (*capacity*) as well as the average attendance (*attendance*) over the last season.[10] Data are again downloaded from `transfermarkt.de`.[11]

Another common explanation of the home advantage refers to the discomfort faced by the away team. The inconvenience and fatigue from traveling and playing under possibly different climatic conditions has previously been identified as an important factor for the home advantage. To proxy such possible discomfort, I collect the location of each side's stadium and compute the linear distance between the home team's and the away team's stadium (*distance*).[12]

Lastly, I use a binary variable (*crowd*) indicating whether a match was played with (=1) or without crowd (=0). Likewise the variable $empty = 1 - crowd$ identifies matches without crowd. This variable plays the role of the exogenous treatment variable allowing identification of a supporter-presence advantage as opposed to a

[8]The index is computed by 538.com and quantifies the chance of a team to bring about important position changes (e.g. finishing first instead of second or avoiding a relegation spot) by winning the match. It takes values between 0 and 100.

[9]Details on the computation can be found here: `https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/`. Basically, this measure initially combines a team's performance over the last season and its market value. Throughout the season it becomes updated according to data on team's performance throughout the season.

[10]I use the attendance from one season ago for two reasons: (i) all matches were played with crowd and (ii) the attendance in the 2019/2020 is possibly endogenous to the performance of a team in that season. Suppose a team performs surprisingly well for some reason outside of supporters' control (e.g. a new coach being a good fit). Such a success may attract more visitors into the stadium, causing some spurious correlation between performance and attendance.

[11]For promoted teams where attendance is observed for a different (likely less attractive) division, I nonetheless assign last year's attendance. Results are robust to excluding matches involving promoted teams or assigning the 2019/2020 attendance.

[12]Geographical locations are downloaded from geonames.org, distancefrom.com, or google.com/maps depending on availability.

mere home-side advantage (see section 3). I created this variable according to match-specific information and news on `kicker.de`, one of the leading sport platforms in Germany.

In order to measure and estimate the home advantage, I consider the following outcome variables of a match:

- Points obtained by the home team, i.e. 3 for a win, 1 for a draw, or 0 for a loss – *points*
- Goal Difference = Goals of the Home Team - Goals of the Away team – $\Delta goal$
- Shot Difference = Shots of the Home Team - Shots of the Away Team – Δ *shots*
- Expected Goal Difference = Expected goals of the Home Team - Expected Goals of the Away Team – $\Delta\ xG$
- A Dummy variable for the home team winning – *win*.

Finally, my sample consists of all 1,446 matches from the top four leagues (Germany, England, Italy, Spain) in 2019/2020 of which 416 were played behind closed doors. All matches without crowd occurred between March and August 2020. Between mid of March and mid of May no matches were played due to the severe outbreak of the Coronavirus in large parts of Europe that forced leagues to pause their operations . In mid of May the German Bundesliga was the first to restart its operation under newly developed hygienic protocols, followed by the other three leagues over the following weeks in a similar fashion.

3 Estimation of the Home Advantage

The main contribution of this paper consists in identifying the importance of the crowd-effect for the overall home-advantage.

I do so by exploiting the unique situation in which a large number of games takes place at the home teams' stadiums without spectators.[13] This allows me to identify separately the effect of supporters for the prospects of home teams.

I begin with presenting some descriptive evidence on the effect of empty stadiums. Given the experimental character of the considered treatment, such descriptive evidence can already be considered useful for identifying causal effects.

Descriptive Evidence

In Table 1, summary statistics on outcome-related variables are compared for matches with and without crowd.

The most significant dimension of a football match's outcome is certainly its result and the consequently awarded points. Before the Covid-19 pandemic, hence with a crowd, the home side secured an average of 1.568 points per match, while playing in an empty stadium resulted in slightly less 1.481 points for the host team.

In the 1,030 matches with fans, home teams scored on average 0.296 goals more than their opponent. In the 416 matches without spectators this number decreased to 0.180. In particular, home teams scored less goals (1.517 vs. 1.568), created fewer chances (11.728 vs. 13.257 shots per game), and consequently won only 41.5% of their matches compared to 44.1% before the ban of supporters.

Notably, shots and expected goals of the away side remained pretty much at the same level with and without fans. One might argue that such a pattern is no surprise. Away support is often quite insignificant and the atmosphere is largely determined by home supporters. Yet, it is important to consider that the perfor-

[13] Such situations usually occur very rarely when fans are banned from the stadium for prior misconduct. Vice versa, some matches are played at neutral grounds, yet with supporters, when the usual stadium of a team cannot be used. However, these events are not well-suited for documenting the desired effect for at least two reasons: (i) misconduct is possibly endogenous to success, and (ii) they occur rather infrequently.

mance of the away team's attackers is also impacted by the performance of the home team's defense, and vice versa. Hence, there is no ex-ante reason to assume only the measured home-team performance (e.g. shots or expected goals) is altered as a consequence of the crowd absence. It is a complex interaction between players from both sides that should in one way or another tend to be dominated by the home team.

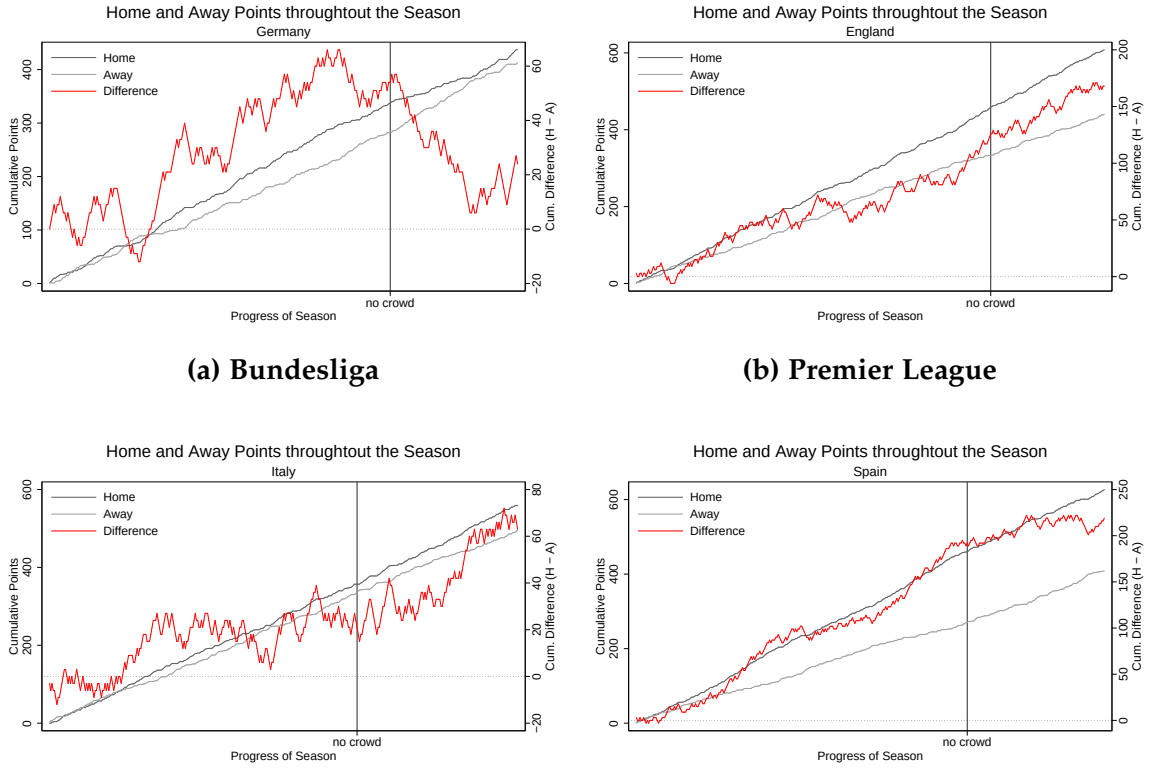
	crowd		**no crowd**		**Overall**	
	mean	median	mean	median	mean	median
points	1.568	1	1.481	1	1.543	1
win	0.441		0.416		0.434	
$\Delta goal$	0.296	0	0.180	0	0.263	0
goals (H)	1.568	1	1.517	1	1.553	1
goals (A)	1.272	1	1.337	1	1.290	1
shots (H)	13.257	13	11.728	11	12.817	12
shots (A)	10.817	10	10.726	10	10.790	10
xG (H)	1.641	1.51	1.479	1.34	1.595	1.45
xG (A)	1.353	1.195	1.348	1.21	1.351	1.2
N	1,030		416		1446	

Notes: For the variables xG(H) and xG(A) the overall number of observations are 1,443 since for three matches the values are not reported in mz sample.

Table 1: Descriptive Measures with and without Crowd

Figure 1 plots the sum of obtained points by home and away teams (in gray) throughout the season for each league. That is, if a home team wins a match the respective curve increases by 3, for a draw both curves by one, and if the home team loses/ the away team wins the respective away curve grows by 3. The red line is simply the difference between overall points collected by home and away teams and is measured on the vertical axis on the right. Across all 4 leagues, the home teams secure more points over the season, even though in the Italian league the home-advantage appears to be relatively small compared to the other three

leagues. The horizontal line denotes the introduction of games behind closed doors for each league. In Germany there is a striking visible effect. Away teams reverse the previous trend, they secure more points when playing in empty stadiums and bring the cumulative difference almost down to zero towards the end of the season. Contrarily, in England as well as in Italy no such effect is immediately visible. In Spain, the home advantage continues to prevail but appears to have been slowed down considerably by the absence of supporters.



(a) Bundesliga **(b) Premier League**

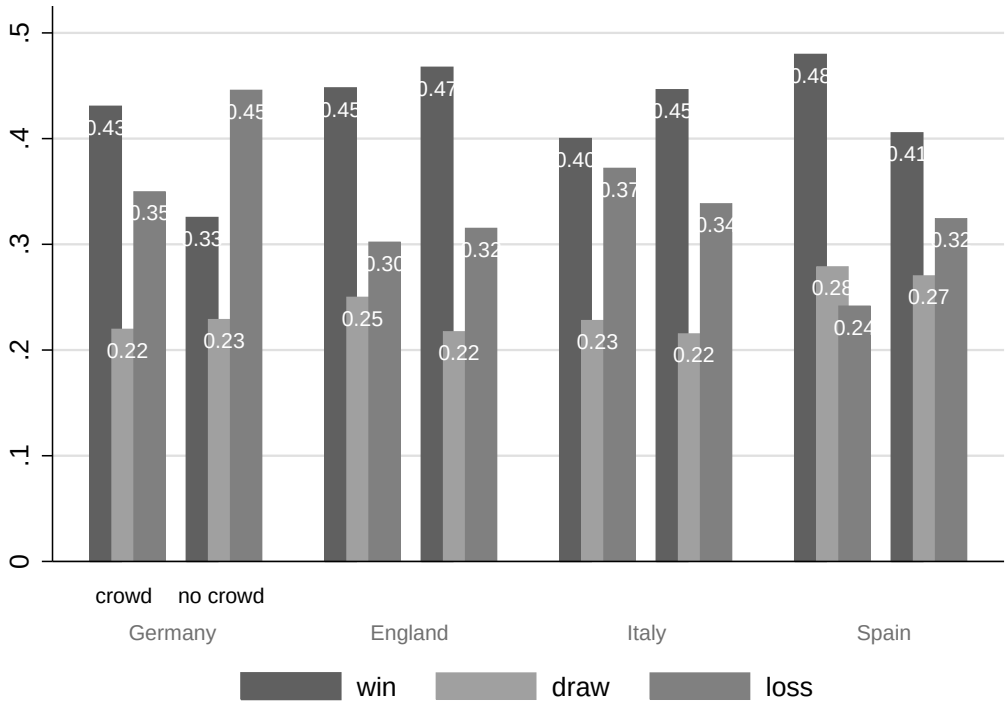
(c) Serie A **(d) La Liga**

Notes: On the left axis I depict cumulated points by home teams and away teams respectively. The right axis displays the excess points that home teams secured cumulatively over time. The vertical line indicates nationwide prohibition of spectators.

Figure 1: Evolution of Points won by Home and Away Teams

Naturally, this pattern emerges also in the distribution of final match results, compared before and after the ban of crowds. Figure 2 plots the distribution of the three possible match outcomes for the four different leagues with and without

crowds. Certainly, in all leagues the home team wins more often than the away team when fans are admitted to the stadium. In Germany, the opposite is true when fans are absent and away teams secure a win in 45% of the matches. In England and Italy, the absence of supporters does not change the prospects of the home teams for worse, rather home teams are found to win (slightly) more frequently. In Spain, home teams do win more often than away teams even without a crowd, yet the advantage appears to be narrowed down. This is in line with the evidence from Figure 1 regarding the collected points. Notably, the frequency of draws is stable across leagues and not visibly affected by the crowd.



Notes: Each set of the eight sets of bars represents the distribution of results for a particular league with or without crowd. The numbers on top of the bar represent the frequency of a given match outcome for one of the eight subgroup of matches. For example, 0.43 matches in the German Bundesliga are won by the home team when a crowd is admitted.

Figure 2: Distribution of Results

To sum up this part, a first glance at how teams performed after the forced absence of supporters indicates that, indeed, crowd-support does (positively) matter

for the prospects of the home side. Unconditionally, home teams perform significantly worse after the re-start of leagues. Distinguishing between the different leagues, however, reveals that for Italian and English teams such an effect is not directly observable.

Given the plausibly exogenous intervention in stadium-attendance, one might be tempted to give these results a causal interpretation. This would, however, ignore the match schedule that could coincidentally favor home or away sides for the fixtures without crowd. Suppose away teams in Germany during the ban of spectators were predominantly of higher quality than the home teams. Then even without any effect of supporters, a pattern like the one in Figure 1 would emerge. Such spurious correlation would then falsely ascribe a causal effect to crowd-presence. Also a prolonged period of luck on the away sides' favor could drive the pattern in the German and Spanish leagues.

Therefore, I next use econometric models that estimates the effect of crowd-presence, controlling for a variety of other factors such as team quality. I also include the dependent variables *shots* and ΔxG, subject to less randomness than scored goals or obtained points.

Econometric Strategy

Before presenting the results, I introduce the identification strategy of my estimation. For simplicity, I use a linear model and the dependent variable $\Delta goal$ to elucidate the underlying principle. The idea remains however the same along richer specifications and non-linear models with different dependent variables.

Assume that the true model for the dependent variable (here: Goal Difference) is given by the following linear model

$$\Delta goal_i = \beta_0 + \beta_1 empty_i + \gamma X_i + \epsilon_i \quad (1)$$

where the vector X_i contains controls for factors that are correlated with the outcome of the match: difference in market values/ quality, difference in form, traveled distance of the away team, stadium characteristics, difference in the importance of the match for each side, and difference in rest days.

In a regular season, the variable $empty_i$ takes the value 0 for all i because matches are – besides very few exceptions – played exclusively with a crowd. Hence, this model cannot be estimated.

The ban of mass gathering, however, induces the unique situation in which, for a sizable amount of matches (416 in my sample), $empty_i$ is set to 1. In addition, it seems natural to assume that their occurrence, i.e. the epidemiologically motivated ban of crowds, is exogenous to all sorts of football-related outcomes. This allows to estimate β_1, the crowd-presence induced home-advantage, consistently and separately from β_0.

The next part presents estimation results from models in the spirit of (1) using the different dependent variables introduced in section 2. For the binary outcome variable of a home victory (*win*), I use a Linear Probability Model while the remaining models are estimated using OLS.

Estimation Results

Baseline Model

Table 2 summarizes the results from estimating the parameter β_1 from model (1). Throughout all specifications, i.e. for the number of awarded points to the home team (1), the excess goals (2), the excess shots (3) as well as the excess expected goals

of the home team (4), and also the probability for a win of the home team (5) the estimated coefficient is negative as suggested by previous literature, common sense, and the descriptive analysis from above. However, in all but the specifications with $\Delta shots$ and ΔXG as dependent variable, the estimated coefficient is not different from zero using standard significance levels. Regarding the overall home-advantage (β_0), I find a positive and significant effect on $\Delta shots$, ΔxG as well as on $\Delta goals$ that, in all specifications, is quantitively larger than the effect of supporters. So even without crowd, the home-team is – setting all other other factors to zero – on average outperforming the away-team (by a small margin).[14] For the arguably most reliable measure of performance, the difference in expected goals from column (4), I find that home teams achieve on average 0.29 more expected goals than their opponents when all other factors are set to 0. This advantage is approximately halved, i.e. decreases by 0.157, when supporters are banned from stadiums.

Note that I excluded stadium related control variables in this analysis in order to give the constant β_0 a handy interpretation: if two (hypothetically) identical teams in terms of market value, importance of the match, form, and rest days faced each other, the home team would score β_0 goals more due to the home-advantage. If supporters were not admitted, the home-advantaged shrinks by β_1.[15] The coefficient of interest, β_1, is not altered in a meaningful way when these controls are added (see Appendix). Results are also robust to using the more sophisticated and continuously updated measure of quality SPI (see Appendix). Overall, the identified effects are quantitatively by and large in line with the descriptive evidence, pointing towards only a negligible effect of the scheduled match pairings.

[14] Consider the goal difference from column (2): To read off the home advantage when playing without crowd one needs to add the coefficient of *empty* to the constant. In this case, the advantage from playing at home, yet in an empty stadium, would be 0.297 - 0.115 = 0.182.

[15] In contrast, if for example capacity is added, β_0 has the inconvenient interpretation of this effect for a stadium with capacity of 0.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (LPM)
empty	−0.087	−0.115	−1.434***	−0.157**	−0.025
	(0.071)	(0.102)	(0.395)	(0.067)	(0.027)
$\Delta market\ value$	0.001***	0.001***	0.006***	0.001***	0.000***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
$\Delta importance$	0.002	0.004**	0.020***	0.003***	0.001
	(0.001)	(0.002)	(0.007)	(0.001)	(0.001)
$\Delta rest$	−0.001	−0.004	−0.013	−0.001	−0.000
	(0.004)	(0.005)	(0.015)	(0.003)	(0.001)
$\Delta form\ recent$	0.002	0.005	0.045	0.008	0.002
	(0.012)	(0.018)	(0.069)	(0.012)	(0.004)
$\Delta form\ overall$	0.151**	0.281***	1.394***	0.279***	0.046*
	(0.06)	(0.083)	(0.333)	(0.056)	(0.024)
Constant	1.569***	0.297***	2.447***	0.290***	0.441***
	(0.039)	(0.053)	(0.209)	(0.036)	(0.041)
N	1,446	1,446	1,446	1,443	1,446
R^2	0.137	0.173	0.222	0.260	0.112
adj. R^2	0.133	0.17	0.219	0.257	0.109
F	49.099	46.89	65.487	64.511	41.868

*Notes: Robust standard errors in parentheses. Clustering standard errors at the team level yields basically identical standard errors, which is no surprise as every match-pair is in some sense a unique observation. Therefore, I only use robust standard errors throughout the paper. Since all outcomes are highly positive correlated and aim at measuring the same thing, performance, no Bonferroni correction is applied despite 5 different dependent variables. * p<0.1, ** p<0.05, *** p<0.01.*

Table 2: Baseline Estimation of the Home Advantage

Regarding the insignificant effect on goals and points, a possible explanation could be the arguably high degree of by which football match outcomes are characterized. Goals occur relatively infrequently – the mean of total goals per match

amounts to only 2.84. As a consequence, with still only around 400 matches played without crowd, I might not find significant effects as the crowd induced home advantage, or respectively the lack thereof, does not translate into actual changes of outcome frequently and consistently enough. Put differently, the noise in the outcome variables (points, goals, and wins) might be too high for establishing a significant effect. In line with this explanation are the significant coefficients found when looking at shots or expected goals as a (less noisy) proxy for performance as well as the higher R^2 for those models (column (3) and (4)).

These first estimation results suggests that home teams perform worse when playing without supporters. So far, however, I have assumed that there exists one coefficient, β_1 across all teams and leagues. Considering the heterogenous trends from Figure 1, I next allow for the effect of the crowd to differ by league, by capacity and attendance of the stadium, and the experience teams have in playing without crowd. This aims at identifying why some teams may have suffered more from the ban of crowds than others. Fischer and Haucap (2020) for example argue that the low attendance in lower tier leagues in Germany leads to a smaller decline of the home advantage because players of such teams are already used to poorly filled stadiums.

League Differences

In the spirit of Figure 1, I start with emphasizing league differences in the effect of crowds. Table 3 contains coefficients from estimating the following models for different dependent variables

$$\Delta goal_i = \beta_0 + \alpha^l \cdot league_i + \beta_1^l crowd_i \cdot league_i + \gamma X_i + \epsilon_i \tag{2}$$

where *league* represents a set of dummies for the four different leagues (l). The respective coefficients $alpha^l$ measure the difference in overall home-advantage with respect to the baseline league Germany (β_0). The effect of the crowd for each league is simply the corresponding β_1^l. In all five specification, the same control variables as in Table 2 are included.

Comparing the different intercepts reveals that there is exist very little differences between the leagues when it comes to the overall advantage of playing at home. Only Spanish home teams are subject to a significantly higher advantage when playing at home with crowd. Bundesliga home teams are estimated to score slightly more than 0.2 goals more per match compared to the away team when all other factors are set to zero, hence abstracting from quality differences. The coefficients α^l for English and Italian clubs are all insignificant and close to 0, indicating no relevant differences with respect to the magnitude in the baseline specification Germany. Only Spanish teams benefit significantly more from playing at home (prior to the ban of crowds). The estimated α^{Spain} in column (4) is highly significant and of a similar magnitude as β_0, doubling the home-advantage to an excess of more than 0.4 expected goals for home teams in the Spanish LaLiga. The effect on actual goals (column (2)) is of a similar magnitude, yet not statistically significant. From column (1), however, it is apparent that Spanish home teams also do secure more points than home teams in the other leagues and are more likely to win (column (5)). Notably they do so while without displaying a significantly better shot balance (column (3)).

Turning to the main coefficients of interest – the effect of the crowd – I find mixed evidence. In Germany, playing in one's own but empty stadium, significantly decreases the points of the home team by an estimated -0.305. Similarly, Spanish home teams are harmed by an average loss of 0.227 points when they have to play without supporters. In England in Italy, however, I do not find a significant effect and the

estimated coefficients are even positive. This is again in line with the unconditional evidence from Table 1.

Interestingly, for the German Bundesliga the effect on shots and expected goals (column (3) and (4)) is not significant. While the estimated coefficients are both negative, it is worth mentioning that the actual decline in goals of 0.465 (column (2)) is surprisingly large when compared to the estimated decline in expected goals of only 0.211 (column (4)). I will emphasize these differences in more detail in section 4, investigating their composition and providing a possible explanation.

Regarding the English and Italian league, only the effect on $\Delta shots$ is negative and significant. Other coefficients are close to zero, even positive, and not significant. There appears to be no relevant effect of supporters for these leagues.

In this light it is surprising to find such a strong effect on $\Delta shots$. While in Germany, where the home-advantage in terms of $\Delta goals$ shrinks the most when crowds are banned from stadiums, the point estimate in column (3) is insignificant and comparably small (around 0.7). By comparison, English and Italian home teams achieve significantly fewer excess shots (around 1.7 and 1.3, respectively). That does, however not translate into worsened expected goals nor into actually scored goals or obtained points. Spanish teams suffer a decline in ΔxG of around 0.3, the same magnitude as the actual decreases in scored goals $\Delta goals$. As a result home teams in Spain secured an average 0.227 points less after controlling for differences in their quality when playing without crowd.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (LPM)
England	0.08	0.048	0.07	0.04	0.017
	(0.112)	(0.167)	(0.629)	(0.105)	(0.040)
Italy	−0.085	−0.09	−0.054	0.026	−0.031
	(0.116)	(0.17)	(0.576)	(0.111)	(0.044)

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (LPM)
Spain	0.206*	0.244	0.609	0.212**	0.049
	(0.113)	(0.161)	(0.610)	(0.100)	(0.041)
empty · Germany	−0.305**	−0.465*	−0.748	−0.211	0.105
	(0.154)	(0.255)	(0.905)	(0.167)	(0.065)
empty · England	0.034	0.091	−1.742*	−0.119	0.023
	(0.147)	(0.212)	(0.911)	(0.151)	(0.059)
empty · Italy	0.119	0.145	−1.278**	−0.039	0.044
	(0.135)	(0.185)	(0.614)	(0.122)	(0.049)
empty · Spain	−0.227*	−0.298*	−1.859**	−0.286**	−0.073*
	(0.134)	(0.175)	(0.793)	(0.114)	(0.042)
Constant	1.514***	0.242*	2.281***	0.218***	0.431***
	(0.084)	(0.133)	(0.449)	(0.077)	(0.031)
Controls			✓		
N	1,446	1,446	1,446	1,443	1,446

*Controls include the same variables as in Table 1, but for better readability their coefficients are omitted. Robust Standard Errors are reported in parentheses. Star-levels follow the usual convention; * p<0.1, ** p<0.05, *** p<0.01.*

Table 3: Estimation Results, separately by League

Given that the above results regarding $\Delta goals$, $\Delta shots$ or ΔxG are composed of the respective measure for the home and away team, it is worth stressing that the observed differences can be either due to a change in the home teams' performance, the away team's performance, or both. Even in the following more granular analysis, an identification of the true underlying cause is not fully possible. A worsening of the observed home team's attacking performance can likewise be caused by an improved performance of the away team's defense. Arguably this is true for all

possible measures, even passing accuracy as for example used in Dilger and Vischer (2020). An altered passing accuracy of the home can be a result of the home team playing less accurate passes or the away team better intercepting passes.

Stadium Characteristics

Next, I turn to the dimensions of home stadiums and their attendance. Suppose crowd-presence was important due to the created atmosphere. Then, the dimension of the stadium, its attendance, or the percentage of seats sold might well affect the magnitude of the effect. Intuitively, clubs with larger stadiums or higher attendance should suffer more when crowds are not permitted.

To estimate such an effect, I use a models in the spirit of (1) where an interaction term between the binary treatment variable and stadium characteristics (e.g. attendance) is added

$$\Delta goal_i = \beta_0 + \beta_1 crowd_i + \beta_2 crowd_i \cdot attendance_i + \gamma X_i + \epsilon_i \tag{3}$$

and test the hypothesis whether β_2 is different from zero. In Panel B and C of Table 4, I replace $attendance_i$ with $capacity_i$, and their ratio $\frac{attendance_i}{capacity_i}$ ($occupancy$). Attendance and capacity are measured in tens of thousands.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)
PANEL A - Capacity					
empty	−0.201	−0.267	−1.337**	−0.088	−0.201
	(0.130)	(0.190)	(0.655)	(0.111)	(0.138)
empty · attendance	0.037	0.049	−0.031	−0.022	0.042
	(0.034)	(0.052)	(0.181)	(0.033)	(0.037)
Constant	1.569***	0.297***	2.447***	0.290***	−0.160***

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)
	(0.039)	(0.053)	(0.209)	(0.036)	(0.041)
PANEL B - Attendance					
empty	−0.300**	−0.305	−1.622**	−0.098	−0.323**
	(0.148)	(0.216)	(0.772)	(0.134)	(0.161)
empty · capacity	0.053*	0.047	0.047	−0.014	0.063*
	(0.032)	(0.048)	(0.17)	(0.031)	(0.035)
Constant	1.569***	0.297***	2.447***	0.290***	−0.160***
	(0.039)	(0.053)	(0.209)	(0.036)	(0.041)
PANEL C - Occupancy					
empty	0.01	−0.076	−0.821	−0.121	0.041
	(0.221)	(0.308)	(1.02)	(0.173)	(0.220)
empty · occupancy	−0.127	−0.052	−0.8	−0.047	−0.146
	(0.270)	(0.385)	(1.305)	(0.226)	(0.272)
Constant	1.569***	0.297***	2.447***	0.290***	−0.160***
	(0.039)	(0.054)	(0.209)	(0.036)	(0.041)
Controls			✓		
N	1,446	1,446	1,446	1,443	1,446

*Controls include the same variables as in Table 2, but for better readability their coefficients are omitted. Robust Standard Errors are reported in parentheses. Attendance and capacity are measured in tens of thousands. So a one unit increase in attendance has to be interpreted as an additional 10,000 attendees. Star-levels follow the usual convention; * p<0.1, ** p<0.05, *** p<0.01.*

Table 4: Stadium Characteristics as Explanatory Variables

Throughout almost all specifications the estimated coefficients are not statistically significant different from zero. Further, there is no consistency regarding their sign, and their magnitude is relatively small. Hence, I do not find evidence for the effect of crowd to depend on the attendance in the stadium. Also the size of the stadium as well as a measure of atmosphere (average percentage of seats sold) are

not found to be associated with different effect sizes of the crowd. Given that for top-tier clubs attendance is often very close to 1, it is no surprise to find such effects. The argument made serves mainly to exclude that league differences are driven by different pre-pandemic levels of occupancy.

It seems therefore unlikely that the heterogeneity regarding the effect across the four leagues is driven by different levels of stadium attendance or the like. In contrast, Fischer and Haucap (2020) ascribe some of the differences across the first and lower tier German divisions to the difference in occupancy. On the basis of data for the top four first tier leagues in Europe this, however, has to be rejected. [16]

Experience with Empty Stadiums

Lastly, I address the experience teams have playing in (their) empty stadiums. Potentially the identified effects could be driven by a short-lived shock of empty stadiums. If after a while players get used to their home stadium not being populated by supporters, the effect might disappear. Fischer and Haucap (2020) find that the disadvantage of absent crowd disappeared for German teams after some match-days. Indeed, in Figure 1 the data for Germany and England do point towards such a learning process that makes the home advantage disappear only for a while before returning.

If in addition not only the own experience, but also learning from observing other matches behind closed doors is relevant, experience could explain some cross-league differences that started at different points in time.

I therefore add the following two variables to my regression.

[16]The finding for the first three German leagues could be driven by other relevant differences between the first and lower division. Since such differences are correlated with occupancy, which is highest in the first league, omitting them introduces a bias in the estimate of occupancy.

- The amount of home matches without fans a certain home team has played previously (in that season) exp_i
- The amount of matches with empty stadiums that have been played previous to a certain match across the four leagues in consideration exp_o.[17]

In Panel A of Table 5, I add an interaction term between the binary variable $empty$ and individual as well as overall experience to the regression and report the respective coefficients.

	(1) $points$	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) win (Probit)
PANEL A					
$empty$	−0.201	−0.267	−1.337**	−0.088	−0.201
	(0.130)	(0.190)	(0.655)	(0.111)	(0.138)
$empty \cdot exp_i$	0.037	0.049	−0.031	−0.022	0.042
	(0.034)	(0.052)	(0.181)	(0.033)	(0.037)
$empty \cdot exp_o$	0.037	0.049	−0.031	−0.022	0.042
	(0.034)	(0.052)	(0.181)	(0.033)	(0.037)
PANEL B					
$empty$	−0.004	−0.095	−0.587	−0.097	0.019
	(0.221)	(0.309)	(1.044)	(0.175)	(0.222)
$empty \cdot occupancy$	−0.197	−0.147	0.328	0.065	−0.253
	(0.281)	(0.398)	(1.350)	(0.236)	(0.286)
$empty \cdot exp_i \cdot occupancy$	0.04	0.055	−0.653 ∗ ∗	−0.064	0.061
	(0.048)	(0.073)	(0.259)	(0.046)	(0.050)
Controls			✓		
N	1,446	1,446	1,446	1,443	1,446

[17] This is certainly an imperfect measure for learning. Possibly, teams learn more from matches in the same league that are broadcasted more prominently in the local media. Also entirely ignoring matches behind closed doors from other leagues, e.g. Portugal, is certainly not fully justified. Hence, I strongly suggest not to quantitatively interpret the estimated coefficients for these variables. I am mainly interest in their sign and significance. To construct this variable for each match, I count matches behind closed doors that were played on previous days. This implies that I ignore matches that were played behind closed doors on the same day yet prior to a given match.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)

*Controls include the same variables as in Table 1, but for better readability their coefficients are omitted. Robust Standard Errors are reported in parentheses. In Panel B, occupancy is computed as the average percentage of seats sold in the last season. Star-levels follow the usual convention; * p<0.1, ** p<0.05, *** p<0.01.*

Table 5: Stadium Characteristics as Explanatory Variables

Individual experience seems to play no significant role. The estimated coefficients are not significant and negative across all 5 specifications. The sign is the opposite of the one expected under the assumption of learning to deal with crowd-absence. Besides the actual outcome, and hence awarded points and probability of winning, overall experience is also not significantly associated with better outcomes (goals, shots, expected goals) for the home team. Given that English and Italian leagues re-started their leagues late, this is no surprise because it is precisely those teams that are assigned high values of overall experience. At the same time, these teams are found to suffer less from the ban of crowds. When repeating the regressions for the four leagues separately, overall experience turns out to be insignificant within each league.

Interestingly, Fischer and Haucap (2020) find a positive effect of the interaction between the three terms *empty*, *experience individidual*, and *occupancy*.[18] They interpret it as an habituation effect – teams who usually play in packed stadiums do require some experience in order to acclimatize to the new conditions entirely without crowd. In Panel B, I add this triple interaction. If such a habituation effect

[18]They explicitly use match-day instead of experience, but given that my sample consists of various leagues with different number of matches and conventions of scheduling matches, I use the amount of matches played behind closed doors. The underlying idea of acclimatizing after some experience remains the same.

was present, the coefficient on the interaction should be positive. In contrast, using my sample of the four top European divisions, I do not find such an effect.

In summary, using four top leagues, deemed comparable in terms of professionalism, level of ability and various other dimensions, I find home teams to be affected by the ban of crowds very differently. Previously highlighted explanations in the literature for differences in the first three German divisions (occupancy and experience) however, fall short in explaining these cross-country differences. Therefore, I next suggest to consider a different channel of crowd-support to reconcile these findings.

4 A Possible Explanation of the Findings

The above findings are partially surprising and deliver novel insights into the role of supporters in the stadium. While my analysis confirms the decreasing home-advantage for German teams to some extent, only Spanish home teams are found to suffer consistently when playing without supporters. These effects are neither driven by a general, unrelated to crowd, deterioration of the home-advantage towards the end of season (see Appendix A.2) nor by observable differences in the teams' abilities or other relevant factors. Contrarily, in England and Italy, the home advantage is not diluted when supporters are banned from the stadiums. Surprisingly, this is true despite home teams achieving (relative to the away side) significantly less shots compared to matches with supporters across all leagues. While the effect on shots is of similar magnitude across leagues, it translates into a significant decrease of expected and actual goals only in Spain and partially in Germany.

To understand better, how these differences arise, I analyze home and away performance measures separately. The descriptive analysis (Table 1) indicates that observed changes are mainly driven by altered levels of performance of the home side.

To understand what has driven the rather different outcomes across the four leagues in my sample, I repeat a similar analysis using as dependent variables home and away team measures separately – not their difference (e.g. $\Delta goals$) as above.

I begin with estimating the following model

$$shots_i^j = \beta_0 + \alpha^l \cdot league_i + \beta_1^l crowd_i \cdot league_i + \gamma X_i + \epsilon_i \tag{4}$$

that is identical to models like 2 from section 3, except for the dependent variable being not the difference in shots but rather the level of the home side ($j = H$) and away side ($j = A$), considered separately.

In Panel A, I report the coefficients $beta_1^l$ from regressions using $\Delta shots$, $shots^H$, and $shots^A$ as dependent variables. By construction, the estimated effect on $\Delta shots$ in column (1) is just the difference of the effects in (2) and (3).

	(1) Δ	(2) H	(3) A
PANEL A - Shots			
Germany	−0.748	−1.239**	−0.492
	(0.905)	(0.632)	(0.561)
England	−1.742*	−1.845**	−0.103
	(0.911)	(0.523)	(0.581)
Italy	−1.278**	−0.921**	0.357
	(0.614)	(0.420)	(0.405)
Spain	−1.859**	−1.599**	0.260
	(0.793)	(0.495)	(0.479)
PANEL B - Expected Goals			
Germany	−0.211	−0.195*	0.016
	(0.167)	(0.105)	(0.114)
England	−0.119	−0.167*	−0.048
	(0.151)	(0.097)	(0.104)

	(1) Δ	(2) H	(3) A
Italy	−0.039	−0.022	0.017
	(0.122)	(0.086)	(0.078)
Spain	−0.286**	−0.293***	−0.007
	(0.114)	(0.081)	(0.077)
PANEL C - Goals			
Germany	−0.465*	0.307*	0.158
	(0.255)	(0.164)	(0.169)
England	0.091	0.037	−0.055
	(0.212)	(0.149)	(0.130)
Italy	0.145	0.217*	0.072
	(0.185)	(0.131)	(0.123)
Spain	−0.298∗	−0.242∗	0.056
	(0.175)	(0.128)	(0.155)
N	1,446	1,446	1,446

*The same control variables as in Table 1 are included. Robust Standard Errors are reported in parentheses. Star-levels follow the usual convention; * p<0.1, ** p<0.05, *** p<0.01.*

Table 6: Decomposition of the Observed Effects

With respect to shots the estimated effects of crowd-absence are quite homogenous across the four leagues. Besides for Germany, $\Delta shots$ decreases significantly, as documented above. In line with the descriptive evidence from Table 1, the decline is almost entirely driven by a reduction in the shots of the home team. Away teams display no significant effect and the point estimates are close to zero.

Panel B reports the effects on expected goals. Naturally, the coefficients are much smaller in magnitude. Surprisingly, however, the effects differ considerably across leagues. The decline in shots of the home team in Spain (1.599 less when playing without crowd), leads to a compatible decline in the expected goals of almost 0.3 causing the observed drop in ΔxG. In Italy ΔxG does not change in a meaningful

way when playing without supporters, nor does the home team display lower levels of expected goals despite achieving fewer shots. In England, the home team scores 0.167 fewer expected goals but the magnitude is well below the one found in Spain, despite shots of English home teams decreasing more. The reduction in shots hence did not translate into a comparable deterioration of expected goals as it is observed in Spain. Together with the simultaneous slight reduction in away team expected goals, the effect on ΔxG is estimated negative yet not significantly different from zero. Similarly, the estimated effects for Germany are in line with the worsened shot statistics for the home team, but do not translate into a significant decrease in ΔxG.

Throughout all leagues, however, the observed differences in relative performance measures are largely driven by home team statistics. Given the complex interaction between attackers and defenders, this is a noteworthy observation and not necessarily expected.

In Panel C I report the effects on actual goals. In Spain, the effect on actual goals is in line with the one found for expected goals. Similarly, in Italy and England where the effect on expected goals was relatively small, no significant effects are found for actual goals. In Germany on the other hand, home teams did perform worse than what the effects on expected goals implied. While expected goals for the home team decreased by an estimated 0.195, actual goals decreased by more than 0.3. Similarly, the away side displayed almost no effect in terms of expected goals, yet secured on average 0.158 additional goals when crowds were banned from stadiums. This divergence is interesting because the way expected goals are measured, they should not systematically over- or underestimate actual goals. This can in principle have two causes. German teams might have been particularly unlucky (home sides) or lucky (away sides). Alternatively they could have managed to outperform

the expected goal measure (away sides) or fall short with respect to it (home sides) due to other reasons.

To sum up, I find that home teams consistently shoot relatively less often when playing without crowd. In Spain the fewer shots translate into considerably less expected goals and consequently into fewer actual goals and less awarded points. The effects in England and Germany follow a similar pattern, but are much less pronounced. This is what causes the expected goal balance, ΔxG, to not change significantly. In Italy, the altered shots have no effect on expected goals (and actual outcomes).

A possible explanation that could reconcile these heterogenous findings is the following: Suppose crowd-presence was a two-edged sword and teams can leverage it differently. One the one hand, crowd presence could certainly increase player's motivation and performance through the established channels. On the other hand, it could also exert a certain pressure on players urging them to shoot more often, even in unpromising situations. The presence of this latter channel could well explain why despite fewer shots, the effects on expected goals are relatively small in England and Germany, or even zero in Italy. Under this hypothesis the absence of a crowd, would then not only deteriorate home players' motivation and eventually worsen their performance, but contrarily also allow them to act in a more serene way choosing more promising actions. As a consequence shots decrease while expected goals do not (as much). In Spain, the pressure from the crowd might play a smaller role, leading to a domination of the first effect and hence home teams being hurt the most. In line with this argument is the initially much more pronounced home-advantage in Spain.

5 Penalties

Explaining the different effect of the crowds across leagues with an existing social pressure created by the usually good-intended home supporters may seem far fetched. While it could reconcile the rather sizable disadvantage in Spain with the different magnitudes in other leagues, it is a novel argument not considered in the literature so far. To the best of my knowledge, previous studies have only emphasized the positive effects of the crowd. Social pressure has been discussed as a driver of the home-advantage in, e.g. Scoppa (2020), focussing mainly on its upsides. In order to substantiate the claim of a negative effect of the crowd, I present further evidence from data on penalty conversion, with and without supporters.

Being a complex and highly dynamic, hence hard to fully capture statistically, game with a non-negligible role of luck, the previous analysis suffers from a rather poor understanding of what really happened on the pitch. There exists, however, a situation that is almost perfectly comparable across all matches. Every penalty follows the exact same protocol and hence is highly comparable across different matches and leagues. Additionally, it seems reasonable to assume that the serenity of the penalty-taker plays an important in role in the likelihood of conversion. I exploit the occurrence of penalties in a variety of matches with and without crowd, to shed further light on the role of the crowd and possibly resulting social pressure.

Data

I collect data on all penalties taken in the 2019/2020 season in the four leagues from above. Overall, 499 penalties were awarded, of which around 80% resulted in a goal. 92 penalties, on the other hand, were missed. I observe the penalty-taker

and the opponent's keeper, the minute in which the penalty was taken, the score at the moment of the penalty as well as the final result of the match. [19]

Around 41% of the penalties were awarded in the first half (45min) of the game and 59% in the second half of the game. Matches without crowd saw 340 penalties awarded while in the matches behind closed doors 159 penalties were taken. Slightly more than half (264 penalties or 52.9%) were awarded to the home side. This is, however, not necessarily a result of a possible referee bias but can likewise be the result of the home team finding themselves more often in respective situations, i.e. attacking in the opponent's box.

The Role of Pressure

As argued above, the serenity of the penalty-taker is arguably an important input in the outcome of a penalty. Such pressure naturally emerges from the situation in which a single player is responsible for the entire team's goal count. In addition, following the argument from above, the crowd may exert additional pressure and thereby harm the prospects of the penalty-taker.

Such an effect is indeed observable in the data. I define as 'high pressure' a penalty occuring within the last 30 minutes of a game and in which the current score is not differing by more than one goal between the two teams.[20] In such a situation the game is still on the rocks while the low remaining playing time makes the outcome of the penalty particularly relevant.

In a Linear Probability Model of the following form

[19] Data are available on `https://www.transfermarkt.com/premier-league/elfmeterstatistik/wettbewerb/GB1/saison_id/2019/plus/1`

[20] I consider also a situation with a one goal lead as a high pressure situation because it would somewhat 'decide' the match. It is of course true true that when missing such a penalty, the penalty-taking team is still leading.

$$Pr(Goal_i = 1|X_i) = \beta_0 + \beta_1 HighPressure_i$$

with $Goal_i = 1$ when a penalty results in a goal, and 0 otherwise. I find the coefficient β_1 to be -0.082 and statistically significant different from zero. A penalty that is taken within the last 30min of a still undecided (goals differ at most by one) match is therefore 8 percentage points less likely to be converted into a goal when compared to a penalty that is taken in a situation with lower pressure, i.e. earlier in the game or with an already clearer match outcome (or both).[21] This is in line with the idea that penalty-takers do perform worse when taking the penalty under high pressure and corroborates that pressure can have a negative effect on the performance of players.

The Role of Pressure and the Crowd

In order to study the effect of the crowd, I allow the coefficient of high pressure situations to differ between matches with and without fans and between home and away teams.

Column (2) depicts estimation results with a distinction introduced between penalties taken by home and away teams. Interestingly, home teams appear to convert their penalties 6 percentage points more frequently than away teams, broadly in line with an existing home advantage. In column (3), however, it becomes apparent that once penalties with high and low pressure are distinguished, the home team suffers slightly more from the additional pressure. While not reaching significance at conventional levels, these findings are again in line with the idea that the

[21] The probability to score is reduced from 84% to 76%. Using a Probit model instead, yields coefficients of the same sign and quantitively identical effect sizes.

home side experiences potentially more pressure and hence does perform relatively worse in this high pressure situations.

In order to estimate the effect of the crowd on this effect, I allow the effect size of pressure to differ depending on the presence of supporters. The results provide further evidence for the newly introduced role of social pressure due to crowd presence.

In column (4), I only control for crowd presence and find no effect. In column (5), however, a relatively large negative effect is found for the interaction between pressure and fan presence. Penalties under high pressure (in the absence of the crowd) are converted at a non statistically significant different rate. Also penalties in matches with crowd do not differ significantly – the estimated coefficient is even positive. However, pressure seems to have a sizable negative effect in those matches where the crowd is present. Compared to a low pressure penalty in a match with supporters, penalties under pressure with crowd are found to be converted 18 percentage points less likely, while the effect of high pressure (in matches without crowd) is not significant. The overall effect (column (1) of High Pressure is exclusively driven by matches in front of crowds. In the absence of crowds, high pressure penalties display no difference. Also the crowd per se does not lead to worse conversion rates. Both coefficients are actually slightly positive. What really seems to lower the success probability for a penalty, is the combination of crowd and pressure.

	(1)	(2)	(3)	(4)	(5)
High Pressure	−0.082**		−0.077		0.042
	(0.039)		(0.062)		(0.063)
Home		0.062*	0.069		
		(0.035)	(0.040)		
High Pressure · Home			−0.013		
			(0.275)		
Crowd				−0.003	0.056
				(0.037)	(0.049)
Crowd · High Pressure					−0.184**
					(0.080)
Constant	0.842***	0.783***	0.806***	0.818***	0.804***
	(0.020)	(0.027)	(0.031)	(0.031)	(0.039)

*Notes: The reported numbers are the coefficients of Linear Probability Models. Due to linearity, marginal effects are identical to the coefficients. The sample consists of all N=499 penalties of the 2019/2020 season.; * p<0.1, ** p<0.05, *** p<0.01.*

Table 7: Estimation of Penalty Conversion Probabilities

This finding is in line with the suggested role played by social pressure for the serenity of players. Players experience a drastic drop in their ability to convert penalties into goals when the penalty is taken under high pressure and with a crowd. The effect of pressure is, however not observed, when crowds are banned from the stadium.[22] Lastly, the effect is quantitatively larger for home teams than for away teams (see column (5) and (6)).

In summary, I am able to document that (even) professional footballers' performances suffer under high pressure. Penalties are converted significantly less likely into goals when they occur towards the end of undecided matches. The effect is stronger for home teams whose players might feel even more intimidated and pres-

[22]Naturally, this observation is based on the relatively small sample of 52 penalties awarded in matches behind closed doors. By collecting more data from recent matches being played behind closed doors and increasing statistical power this problem can be overcome.

surized by the crowd. Interestingly, such an effect is not present when there is no crowd. This finding supports the above hypothesis that the ban of crowds has also positive effects. In particular, the relation between shots and expected goals as well as the conversion of decisive penalties suggest that empty stadiums allow (home team) players to act more serene. This new component may well explain why a possible loss of the home advantage is not consistently found in all leagues.

6 Conclusion

I conclude by summarizing the key results of my analysis. In contrast to common sense and some of the previous literature, I find a much less consistent decrease in home team performance for matches behind closed doors across the four biggest European leagues. Only Spanish and to some extent German teams seem to suffer from the ban of crowds, intended to slow the spread of the novel Coronavirus. When looking at shots as a measure of performance, the four leagues display similar effects. However, actual outcomes or expected goal measures show that in England and Italy reduced shots did not translate into significantly worse outcomes for the home team.

To reconcile these differences, I suggest to consider a new dimension of crowd-presence: harmful social pressure induced by the crowd on home team players. To the best of my knowledge, previous literature has so far considered only positive effects of crowd presence. Such an explanation, together with differing degrees of resilience to it, may well explain why Spanish teams suffered the most from crowd absence. This is in line with the most pronounced home advantage in Spain prior to the ban of crowds. Possibly, some teams are better able to cope with the pressure from the crowd (or alternatively in some stadiums the pressure is less relevant), allowing them to leverage the fan support the most. Eventually, in the

absence of a crowd, these teams experience the largest decrease in performance. Contrarily, some teams, i.e. the less resilient ones, may experience no deterioration of performance because the lack of pressure enables them to act in a more serene way compensating the otherwise positive effect of crowds. That the effect seems less relevant in Spain (and Germany) fits the anecdotal evidence of a more tactically sophisticated and well organized playing style in these leagues.

Analysis of penalty conversion and the role of pressure and crowd presence therein supports this idea. I am able to document a significantly lower conversion rate for penalties under high pressure only for matches in which a crowd is admitted to the stadium. In addition, the effect is quantitatively larger for home teams than for away teams. This opens a new direction for future research. In particular with richer data on player behavior it would be possible to analyze player decisions in greater detail. For example, analyzing which passing options players choose and how this is altered when fans are absent can deliver novel insights into the human decision making process.

Additionally, the partially surprising results and the piecewise learning thereof also raise interesting research questions regarding the efficiency of betting market where bookmakers are required to ex-ante form an expectation about the magnitude of the home advantage.

References

Bryson, A., P. Dolton, J. J. Reade, D. Schreyer, and C. Singleton (2020): "Experimental effects of an absent crowd on performances and refereeing decisions during Covid-19," *Available at SSRN 3668183*.

Dilger, A., and L. Vischer (2020): "No home bias in ghost games," *Discussion Paper of the Institute for Organisational Economics 7/2020*.

Endrich, M., and T. Gesche (2020): "Home-bias in referee decisions: Evidence from Ghost Matches during the Covid19-Pandemic," *Economics Letters*.

Fischer, K., and J. Haucap (2020): "Does crowd support drive the home advantage in professional soccer? Evidence from German ghost games during the COVID-19 pandemic," *CESifo Working Paper No. 8549*.

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

Neave, N., and S. Wolfson (2003): "Testosterone, territoriality, and the home advantage," *Physiology & behavior*, 78(2), 269–275.

Oberhofer, H., T. Philippovich, and H. Winner (2010): "Distance matters in away games: Evidence from the German football league," *Journal of Economic Psychology*, 31(2), 200–211.

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

——— (2008): "Home advantage in football: A current review of an unsolved puzzle," *The open sports sciences journal*, 1(1).

Pollard, R., and M. A. Gómez (2014): "Components of home advantage in 157 national soccer leagues worldwide," *International Journal of Sport and Exercise Psychology*, 12(3), 218–233.

Ponzo, M., and V. Scoppa (2018): "Does the home advantage depend on crowd support? Evidence from same-stadium derbies," *Journal of Sports Economics*, 19(4), 562–582.

Schwartz, B., and S. F. Barsky (1977): "The home advantage," *Social forces*, 55(3), 641–661.

Scoppa, V. (2020): "Social Pressure in the Stadiums: Do Agents Change Behavior without Crowd Support?," *IZA Discussion Paper*.

A Appendix

A.1 Further Regression Analyses and Robustness Checks

In this part I report some further regression analyses and robustness checks.

In Table 2, I excluded stadium characteristics in order to give the constant an intuitive interpretation. Here, I report estimation results from the same regressions but add *distance*, *capacity* and *attendance* in order to control for stadium characteristics and the travelled distance of the away team. Importantly, the size of the coefficient of interest is not altered in a meaningful way. Further, all additional control variables themselves are not significantly different from zero.[23]

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)
empty	-0.085	-0.112	-1.431^{***}	-0.157^{**}	-0.07
	(0.071)	(0.102)	(0.396)	(0.068)	(0.075)
ΔMV	0.001^{***}	0.001^{***}	0.006^{***}	0.001^{***}	0.001^{***}
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
$\Delta importance$	0.002^{*}	0.004^{***}	0.020^{***}	0.003^{***}	0.002
	(0.001)	(0.002)	(0.007)	(0.001)	(0.001)
$\Delta rest$	-0.001	-0.004	-0.013	-0.001	-0.001
	(0.004)	(0.005)	(0.015)	(0.003)	(0.004)
$\Delta formrecent$	0.003	0.007	0.043	0.008	0.007
	0.012	0.018	0.069	0.012	0.013
$\Delta formoverall$	0.148^{**}	0.275^{***}	1.400^{***}	0.279^{***}	0.123^{*}
	(0.06)	(0.083)	(0.333)	(0.056)	(0.065)

[23] This remains true also when estimating league-specific effects. The estimation results are omitted in this Appendix.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)
distance	0.000	0.000*	−0.001	0.000	0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
capacity	0	0	0	0	0
	0	0	0	0	0
attendance	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	1.396***	0.032	2.609***	0.305***	−0.304***
	(0.102)	(0.144)	(0.541)	(0.095)	(0.106)
N	1,446	1,446	1,446	1,443	1,446
R^2	0.139	0.177	0.222	0.260	
adj. R^2	0.133	0.172	0.217	0.256	
F	34.400	32.447	43.800	48.643	

*Notes: * p<0.1, ** p<0.05, *** p<0.01.*

Table 8: Robustness Check of the Baseline Estimation including stadium characteristics

The most important benefit of using regression techniques instead of simply comparing outcomes before and after the introduction of matches behind closed doors concerns the possibility of controlling for the quality of the encountering teams. Throughout the paper, I use market values as reported by transfermarkt.de. These market values are in many aspects not perfect measures of quality. The market value of a young but talented, promising player may be high due to his future impact and not necessarily due to his current ability.[24] Another concern is related

[24]Consider the case of Cristiano Ronaldo who at the time of writing this paper is already 35 years old and has a reported market value of 60 Million EUR. His teammate Matthijs de Ligt, who is currently just 21 years old has a reported market value of 67.5 Million EUR (despite being a defender). Many experts, however, would agree that the former player currently has a bigger impact on the

to the development of market values throughout the season. I use market values from the beginning of the season in order to ensure they are exogenous to the performance of a team during the 2019/2020 season. This could, however, introduce a problem. Teams may have added players to their squad, or some talented yet unknown players may have increased their market value over the season. All this leads to an imprecise measurement of actual ability. As an alternative, I use the ESPN Soccer Performance Index that uses a combination of market values and recent performances to proxy a team's strength throughout the entire season. Again, the coefficient of interest does not change and I therefore use the easier to interpret and widely used market values. The R^2 however does increase moderately.

	(1) *points*	(2) $\Delta goals$	(3) $\Delta shots$	(4) ΔxG	(5) *win* (Probit)
empty	−0.076	−0.099	−1.346***	−0.141**	−0.063
	(0.070)	(0.100)	(0.374)	(0.066)	(0.076)
ΔSPI	0.039***	0.059***	0.306***	0.049***	0.037***
	(0.003)	(0.005)	(0.018)	(0.003)	(0.004)
Controls			✓		
N	1,446	1,446	1,446	1,443	1,446
R^2	0.164	0.204	0.304	0.306	
adj. R^2	0.160	0.201	0.301	0.303	
F	60.872	55.761	105.765	90.479	

*Notes:Control variables as in the baseline regression are included but not reportes * p<0.1, ** p<0.05, *** p<0.01.*

Table 9: Robustness Check with SPI as a measure of quality

performance of his team. In that sense the market value may be a misleading measure of current ability.

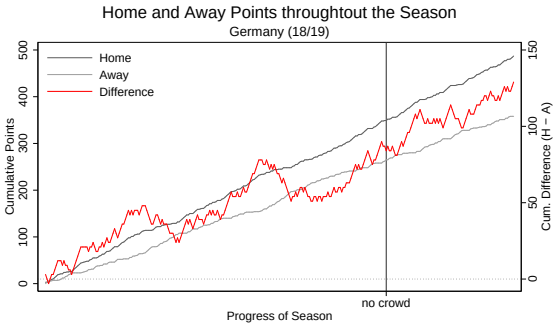
A.2 Placebo Tests

Even though it is plausible that the introduction of the public-event ban is not linked to outcomes of football matches, there remains one concern about the validity of the results. All matches in empty stadiums were played towards the end of the season. Hence, evidence that in these matches the home-advantage is in some cases diminished and ascribing it to the exclusion of supporters could be premature. Another possible cause for the above findings could be a general deterioration of the home advantage towards the end of the season, arguably the period with the decisive matches taking place. Then my estimation method would spuriously estimate a negative effect of empty stadiums and falsely ascribe the decline in home-advantage to the exclusion of supporters.

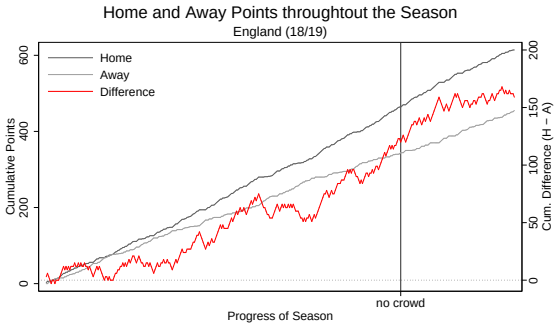
To control for this concern, I repeat the same analysis with data from the two seasons prior to the one in my analysis (2017/2018 and 2018/2019) and pretend that in these seasons matches from march onwards were played behind closed doors.

No such effects are however visible towards the end of a regular season. In Figure 3, I conduct the identical analysis as in Figure 1 and assign a hypothetical ban of crowds to the last matches (as it was the case in the 2019/2020 season). It is also reassuring that in a *normal* season the home advantage evolves similarly across the leagues and the 2019/ 2020 season is prior to the ban of crowd not particularly different.

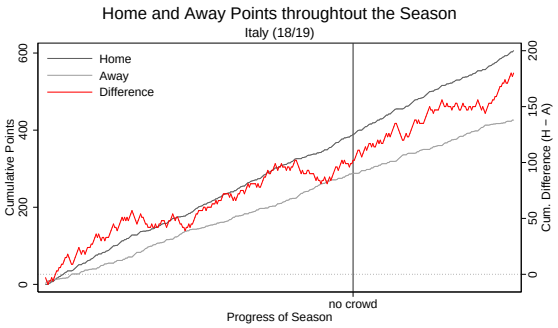
Figure 3 shows the respective graphs for the 2018/2019 season. The 2017/2018 season shows no relevant differences in all four leagues and is therefore omitted. Further regression analysis are not reported.



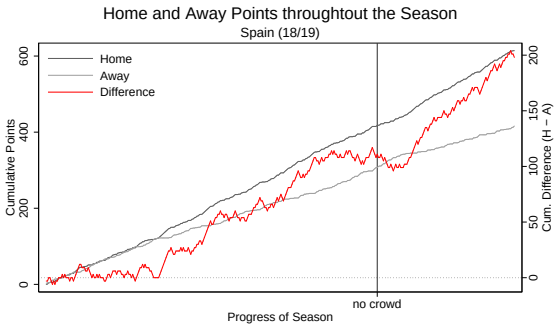
(a) Bundesliga



(b) Premier League



(c) Serie A



(d) La Liga

Notes: On the left axis I depict cumulated points by home teams and away teams respectively. The right axis displays the excess points that home teams secured cumulatively over time. The vertical line indicates a hypothetcial nationwide prohibition of spectators in line with what happened in the 2019/2020 season.

Figure 3: Evolution of Points won by Home and Away Teams with a Placebo Ban of Crowds

Chapter 2

Betting Market Inefficiencies in the Presence of Heterogenous Treatment Effects †

†Author: Marco Wallner, wallner.marco@phd.unibocconi.it

Abstract

Building on the evidence in chapter 1, I show that the reduction in the home field advantage due to the removal of the crowd-component is affecting matches differently. Favorite teams continue to outperform their opponents at identical rates, while underdog teams fail to achieve the pre-pandemic level of home performance. This is in line with the idea of choking under pressure from chapter 1. Favorites may well experience greater pressure and hence it is only natural to see them performing at similar or even better levels once the pressure is released. A detailed look into shots and expected goal conversion confirms this view.

In contrast to the correct pricing of the average reduction in the home field advantage, bookmaker (fixed) odds do not account for this heterogeneity. Based on this observation that is in contrast to at least the semi-strong form of market efficiency, I show that profitable betting strategies can be constructed. Exploiting the mispricing of favorite prospects in home matches without crowd yields positive net returns from placing according bets. The returns are, however, relatively small and mostly offset by the bookmakers' margins. Correcting for the margin and using fair prices, considerable positive returns could be achieved in matches without crowd.

1 Introduction

In early 2020, Europe and most other parts of the world faced a public health crisis due to the spread of the novel coronavirus SARS-CoV-2. To slow down transmission and protect the population and the hospital system, restrictions on social gatherings were introduced. Sectors faced with such limitations faced a variety of different problems, ranging from organizing new ways to sell their products, collaborating without physical contact, or dealing with the financial consequences of missing revenue. The professional sports industry is no exemption. Football league competitions were paused for a couple of weeks to then only restart with matches being played behind closed doors. The obvious financial consequences of which are lower revenues due to ticket sales and other sources of related revenue.

Beyond this direct consequences, the question about the magnitude of the home field advantage in the absence of a crowd troubled many. In chapter 1, I have documented that home team performances did not unequivocally suffer in the weeks following the re-start of the major European football leagues. This finding goes against the everyday wisdom of supporters cheering for their team and by doing so improving the athletes' performance.

In this chapter, I incorporate further evidence including the 2018/2019 and the newly played 2020/2021 season. The former was played with no restrictions on the attendance of supporters while the latter was almost exclusively played behind closed doors due to the ongoing health crisis. The additional data allow me to estimate richer models in order to understand how the absence of crowds affected performance. This evidence is presented in this chapter 2 and builds on some of the analysis in chapter 1. The league-specific differences appear to be less pronounced when taking into account also the 2020/2021 season. Instead, I uncover a different, but stable across leagues, pattern that can explain why teams reacted so differently

to the ban of crowds. Allowing the effect of no supporters to depend on the relative strengths of the teams, reveals that only relatively weaker clubs suffered from the restriction. Teams that can be considered superior to their opponent, displayed no effect. Put differently, only the relatively smaller/ weaker clubs benefit from their supporters when it comes to performance. Favorites on the other hand perform equally well also in the absence of the crowd, some even better. I argue that such a pattern is consistent with the idea of choking under pressure from chapter 1.

Next, I address the question of how bookmakers dealt with this previously unknown situation. Besides evidence on an absolute minority of matches, their pricing/ odd-setting historically is based on home matches that have supporters admitted to the stadiums. Borrowing from the finance literature on market efficiency one might expect that betting odds reflect all the (publicly) available information and hence are set correctly. Since basically no information on the home field advantage without supporters has ever been observed, setting the correct price remains (ex-ante) a challenging task.

I find that bookmakers do properly account for the average decrease in home advantage relatively fast. They do, however, not treat superior home sides, for which the drop in home field advantage is not existent, differently from others. This form of market inefficiency can be exploited by bettors in order to achieve slight positive returns.

The paper is organized as follows. In section 2, I summarize the relevant literature on the crowd-component of the home field advantage and the incorporation of this evidence by bookmakers. In section 3, I describe the data used in this paper. Section 4 presents evidence on the heterogenous effects depending on the favorite or underdog status of the home team. In section 5, I explore to what extent the betting market has incorporated the emerging evidence. At the end of that section, I

show that the arising inefficiencies allow for betting strategies with positive returns. Section 6 concludes.

2 Literature

The exogenous ban of crowds offers a unique opportunity to study questions centered around performance effects of the crowd – the crowd component in the home field advantage (see chapter 1). Previous studies investigating the matter find an effect rooted in altered refereeing behavior in the absence of the crowd (e.g. Scoppa (2020) or Bryson, Dolton, Reade, Schreyer, and Singleton (2020)). They document that part of the home field advantage may be ascribed to more favorable refereeing decisons towards the home side when playing in front of a crowd (e.g. less yellow cards for the home team, more for the away team). Others document that performance suffers due to unfamiliarity with the environment (Fischer and Haucap (2020)), but only for a relatively short period of time to then recover once players have familiarized themselves with the new circumstances. While overall, the reduction in the home field advantage seems relatively clear for at least the top leagues in European football, the underlying mechanisms remain unclear despite such a unique controlled experiment-like situation. In other sports, there exist also mixed evidence. Higgs and Stavness (2021) find that in North-American basketball (NBA) and ice-hockey (NHL) the home advantage decreased, while it did not or only very little in American football (NFL) and baseball (MLB).

An overview, also on how such studies contribute to economic research can be found in Singleton, Bryson, Dolton, Reade, and Schreyer (2021) . Beyond the pure lesson on athletes' performance the natural experiment-like situation offers possibilities to study risk-taking on the supporters side (see Reade and Singleton (2021)), the reaction to incentives and choking under pressure (see chapter 1) and promi-

nently also to study (betting) market efficiency. The latter aspect is also the main contribution of this paper.

Economist have long used betting markets to study information processing in markets (e.g. Thaler and Ziemba (1988)). In particular, theories originating in finance are suitable for being tested in betting markets. The relatively easy environment where betting behavior has no effect on outcomes and the short time until uncertainty about the outcome is revealed make them attractive compared to actual financial markets, where often uncertainty is revealed at best in the long run.

The pandemic situation is no different in this regard. The necessity to offer odds for home matches without crowd provides a natural setting for testing the efficient market hypothesis and its different versions according to Fama (1970). Fischer and Haucap (2021) document inefficiencies in the pricing of the home field advantage when crowds are banned in the Bundesliga. A similar point is made by Winkelmann, Ötting, and Deutscher (2020) who however argue that temporary mispricing might simply be a result of noisy outcomes and is to a comparable extent also featured in Monte-Carlo simulations under the assumption of efficiency. Hegarty (2021) even finds improved accuracy of betting markets without crowd.

Outside the world of European football, Orefice (2021) finds that for the NHL the betting market incorporated the new evidence efficiently and hence no profitable betting strategies were possible.

I contribute to the first literature on the home field advantage by uncovering a new channel that relates the crowd component to the strength of the home team. This heterogenous effect should, under the assumption of market efficiency, be incorporated in the offered odds (at least at some point). By investigating the price reaction with respect to this new effect, I contribute also to the second literature of betting market efficiency exploiting the exogenous ban of crowds. Contrary, to some of the other studies I find that the average reaction of bookmakers is in line

with the shrinkage of the home field advantage. I document, however, that betting markets fail to adequately adjust their odds to the heterogenous effect the ban of crowds exerts, thereby contradicting market efficiency.

3 Data

Variables and Data Sources

In the spirit of chapter 1 as well as other studies (e.g. Hegarty (2021)) my analysis focuses exclusively on the four major European football leagues in Germany, England, Italy and Spain (1. Bundesliga, Premier League, Serie A, and Primera Division). I use data on the 2018/2019 season (the last season played without restrictions in attendance), the 2019/2020 season (where restrictions were introduced for the first time towards the end of the season), and the 2020/2021 season (almost entirely played behind closed doors). Following the idea of a balanced panel data set, I only consider matches that involve teams present in all three seasons. [1]

In total the data used throughout the paper comprise 2526 matches of which 1041 are played without any supporters. An additional 44 are are played with a minor crowd (on average less than 5,000 visitors). Those matches occurred mainly towards the end of the 2020/2021 season when low numbers of infections and progress in vaccination rates allowed stadiums to be filled up to a certain level.[2]

I have obtained match results as well as betting odds from the website `https://www.football-data.co.uk/data.php`. The odds are collected on Friday afternoon for matches played on the weekend, for midweek games on Tuesday afternoon.

[1] Otherwise the dataset would include teams that had their matches played in full stadiums only or, vice versa, never played in front of supporters.

[2] Throughout the analysis I will consider those matches as being played behind closed doors as well. The results are robust to changing that classification or excluding those matches.

There exist minor differences to the closing odds right before the matches that may be driven by betting behavior and hedging motives of the bookmakers. The data include quotes from different popular bookmakers such as Bet365, Interwetten, or William Hill. [3]

I have downloaded from transfermarkt.com data on each team's market value prior to the beginning of the respective season as a proxy for its ability. Further, I have collected data on the actual attendance of each match, also from transfermarkt.com.

Besides that I include more sophisticated metrics computed by 538.com. These contain expected goals, as an alternative measure of performance compared to the mere result of the match. I also include the SoccerPerformanceIndex provided by 538.com to measure each team's ability. The idea is similar to that of market values but it combines market values with actual performance throughout the season to capture some within-season dynamics in the strength of a team (e.g. a new striker outperforming the expectations and hence contributing more to ability than suggested by its pre-season market value). Lastly, 538.com provides a measure of importance of a match to each participating team reflecting different incentives to win the match (e.g. the deciding match for the championship has higher importance than a late game for a team that is already sure to finish on a mediocre position).

A comment on market values

The market value of a player, as reported by transfermarkt.com, is a measure of the financial value of a player. It can best be thought of as a stock price where future dividends are performances on the pitch (and other relevant benefits of signing a player such as media attention or merchandise sales). Unlike stocks, players are not

[3] For a full list, see https://www.football-data.co.uk/notes.txt. The analysis is in large parts conducted using the odds offered by B365. The results are robust to using other bookmakers' quotes.

traded frequently and hence the market values are not market prices based on actual transaction, but rather expert judgements by the transfermarkt.com community and its operators. It is also important to note that market values are not necessarily good predictors of actual transaction prices as the latter depend on various factors such as contractual details, personal preferences or a club's financial situation.

Naturally, older players tend to have lower market values as the stream of dividends (their performance and the like) is only obtainable for fewer years compared to younger players. Hence, this measure may be biased against older teams. Nonetheless, it can be considered as the standard measure of a team's (or player's) ability in the literature and is used throughout most of the related studies.

It is also true that as a result of the restriction on attendance and the uncertain future to clubs' revenues, market values dropped during the pandemic. The drop was, however, relatively small and since in all further analysis I consider differences only, I do not account for changes in the overall price level but use 'nominal market values'. [4]

A comment on expected goals

Before I begin with the actual analysis, let me also clarify how the variable expected goals is constructed. It is not an ex-ante expectation of the random variable goals. Rather it can be considered a sophisticated way of counting shots/ attempts during a match. Instead of simply counting each shot it gives weights to each shot reflecting the probability that the attempt would result in a goal; e.g. high values (closer to 1) to shots near the goal and low values (closer to 0) to shots far away from the goal.[5] It therefore should be considered an ex-post computed variable rather than an actual

[4] In the 2018/2019 season the average market value across all teams in my sample was 328 million Euro. In 2019/2020 (before the season) it amounted to 389 million Euro, while it dropped to 359 before the 2020/2021 season.

[5] Of course also other parameters such as the positioning of the players etc. are considered.

ex-ante expectations on the number of goals a team will score. The reason why I use this variable is the reduced level of noise it contains. Scored goals depend (a lot) on luck, the goalkeeper performance and other circumstances that are not directly related to the actual performance of a team. [6]

Descriptive Statistics

In the 1,442 matches with supporters before the pandemic, home teams on average outperformed the away side by 0.3 goals or 0.32 expected goals. They won 43.9% of the matches, compared to only 29.8% won by the away team (the remaining ones end with a draw). As a consequence, a home team playing in front of a crowd, on average secured 1.57 points. Away teams only end up securing 1.15 points. All this is generally known under the term home field advantage.

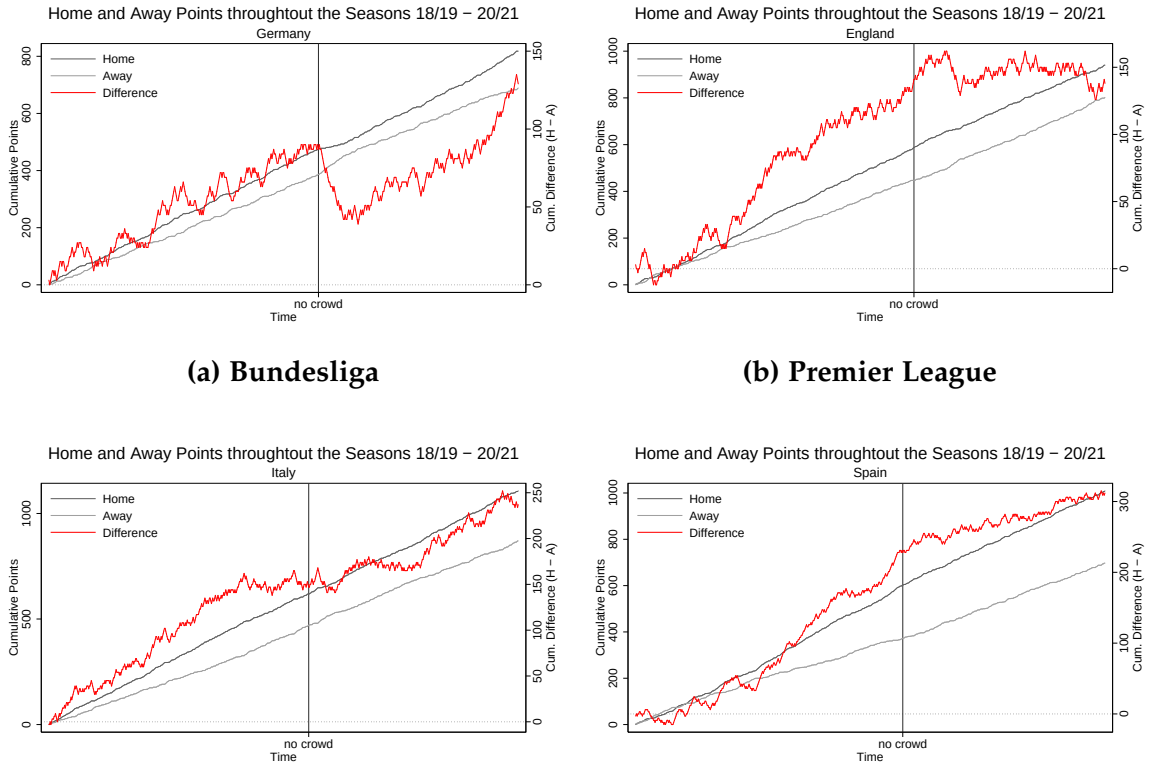
When supporters are banned from the stadiums, the excess goals and expected goals decrease by 0.11 to 0.19 and by 0.18 to 0.15. The share of matches that are won by home teams decreases to 40.9 %, while away teams win more frequently (34.5 % instead of 29.8%). The points awarded to the home side decrease by 0.1 to 1.47, while those secured by the away side increase by 0.13 to 1.28.

This overall trend can be understood as a consequence of the ban of supporters from stadiums. The crowd component, according to this descriptive analysis, can quantitatively account for roughly half of the home field advantage.[7]

An easy way to depict the home advantage is to simply compare the points collected by home teams and those collected by away teams over time. Needless to say that most matches are not won due to the home advantage but due to differences

[6] If one estimates the following model: $Goals = beta_1 xGoals + \varepsilon$, the estimated coefficient is 0.94 and highly significant, showing how closely the two measures are related. The R^2 amounts to around 0.7.

[7] With supporters home teams receive 1.57 - 1.15 = 0.42 more points than away teams. This difference narrows down to 1.47 - 1.28 = 0.19 points, which is approximately half of the prior level. A similar argument can be made for the winning probability and the excess goals.



(a) Bundesliga **(b) Premier League**

(c) Serie A **(d) La Liga**

Notes: On the left axis I depict cumulated points by home teams and away teams. The right axis displays the excess points that home teams secured cumulatively over time. The vertical line indicates the point in time from which onwards matches were played behind closed doors.

Figure 1: Evolution of Points won by Home and Away Teams

in ability, form, importance and other relevant factors as well as luck. The overall evolution of that curve however is indicative by how much on average home teams outperform away teams. Looking at it over many matches, gives a hint at the home field advantage and the evolution of it over time; in particular whether a change occurs once supporters are removed. If no home field advantage existed, the two gray curves would rise at identical rates and the red difference would oscillate around 0. Instead, the red line keeps consistently growing over time.

These trends are plotted in Figure 1, separately for the four different leagues. The right axis measures the difference in points between home and away teams. If no home field advantage existed, that difference should move around zero. Clearly, in

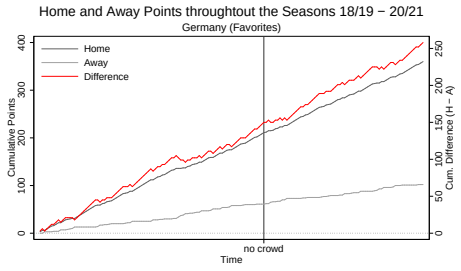
all 4 leagues it has a positive slope/ trend indicating that more points are awarded to home teams than to away teams – the classical home field advantage.

In Germany a huge initial drop in the (relative) performance of home teams is observed after the ban of crowds that almost completely vanishes after a while bringing back the slope close to the pre-pandemic level. In England such initial drop is not evident. It is rather true that the red curve flattens considerably, causing the home field advantage to vanish almost entirely. In Italy the slope flattens out, yet by a smaller amount and no initial drop is evident. Lastly, a similar pattern is found for the Spanish first division: a clear reduction in the slope of the red line that however keeps increasing slightly.

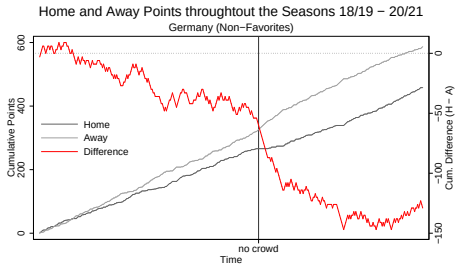
It becomes apparent that with the ban of crowds (represented by the vertical lines), the slope of the difference becomes flatter, best visible for England and Spain. In Germany, there is an initial big drop in the home advantage (even becoming a disadvantage), but after a while home teams manage to regain their advantage and the slope of the red line is (almost) back to its usual trend. Fischer and Haucap (2020) interpret this as a familiarity effect: on impact home teams suffer but after a while they get used to playing in empty stadiums and hence again outperform the away sides. Such pattern is however not visible for other leagues. Concluding this first descriptive look on the data, leaves the following take-away: prior to the pandemic there exists a considerable home advantage and home teams collect (in sum) more points than away teams. This trend is slowed down once supporters are banned from stadiums – a back-of-the-envelope calculation suggests that the home advantage is reduced by approximately half. In Germany there exists a very unique and particular initial big drop that is not present in the other leagues.

In this chapter of the thesis, I argue that the difference in strength of the home and away team plays a key role for the effect of banned crowds. To uncover this discrepancy, in Figure 2, I plot the pre- and post-Covid trend in point differences

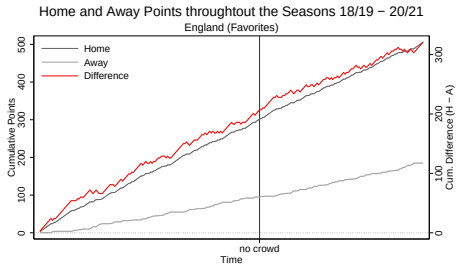
between home and away teams split by the following criterion: the left panel contains only those matches where the home team is considerably stronger than the away team (market value of the home team > market value of the away team + 100 , in million). The right panel plots the same variables but for all other matches, i.e. those in which the home team is similar to the away team in terms of market value or weaker.



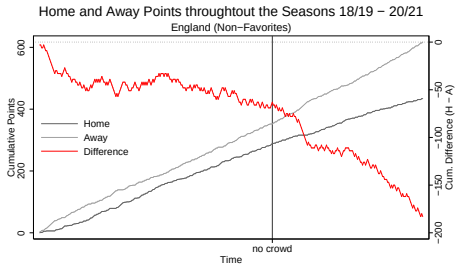
(a) Bundesliga, Favorites



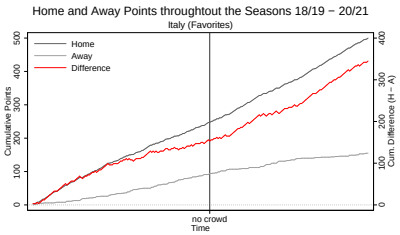
(b) Bundesliga, Non-Favorites



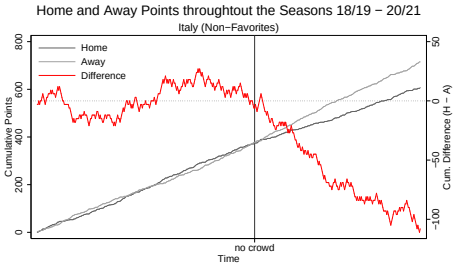
(c) Premier League, Favorites



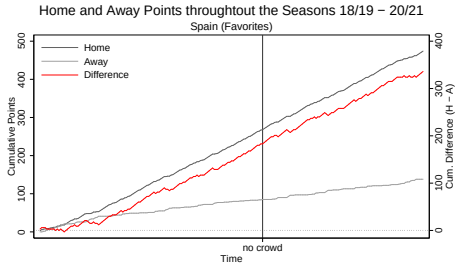
(d) Premier League, Non-Favorites



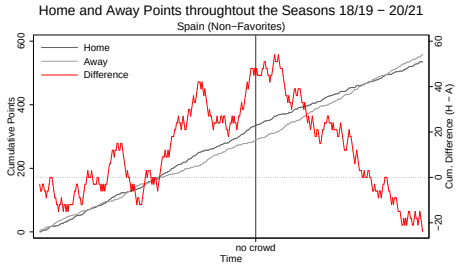
(e) Serie A, Favorites



(f) Serie A, Non-Favorites



(g) Primera Division, Favorites



(h) La Liga, Non-Favorites

Notes: As before, on the left axis I depict cumulated points by home teams and away teams. The right axis displays the excess points that home teams secured cumulatively over time. The vertical line indicates the point in time from which onwards matches were played behind closed doors. The left column contains the trend for all those matches in which the home team is the favorite according to the criterion of its market value exceeding that of the away team by more than 100 million EUR.

Figure 2: Evolution of Points won by Home and Away Teams for Favorites and Non-Favorites

The differences are striking. Obviously, the stronger teams (left panel) collect more points and the difference (in red) grows over time. This is for two reasons. First, they are by definition stronger and due to their superior ability have an edge in the competition for points. Second, they enjoy the home advantage.

In the right panel, the home teams (the 'underdogs') do not manage to secure more points than the away team. Obviously this is due to them being the underdogs. However this time, the ban of crowds does matter a lot. Across all four leagues the performance literally collapses. This puts the findings from chapter 1 and the overall unclear situation regarding the crowd component in a different light.

Favorites apparently do not rely on the crowd to secure points when playing at home. They collect points at the same rate irrespective of the presence of supporters. Contrarily, relatively weaker clubs actually do leverage the presence of their supporters. Once removed, their performance is radically different and they do not manage to keep up with their previous performance in front of supporters. For them the exclusion of supporters mattered and had strong negative effects.

4 Estimation of Treatment Effects

BEFORE turning to a possible explanation of the observed pattern, let me confirm the findings from the descriptive analysis using regression analysis controlling for a variety of otherwise ignored factors, such as the exact level of market value difference or the relative importance of each match to the home and away side, respectively. Also outcome variables other than points will be considered.[8]

[8]Since the introduction arguably happened exogenously at a (from a sports perspective) random point in time, other factors should not be correlated with the match outcomes and the estimates from regression analysis, as in chapter 1, are close in size to the descriptive ones.

Estimation

To identify the effect of an empty stadium on performance and outcomes, I use a standard procedure to establish treatment effects in the presence of an exogenous shock to the treatment variable (here: presence of supporters).

Assume the variable of interest (outcome of the match) follows

$$y_i = \beta_0 + \beta_1 Ban_i + \beta_3 \text{Market Value Differences}_i + \text{controls} + \varepsilon_i$$

where ban_i is a dummy that takes value one for matches played in the absence of supporters.

β_0 then represents the overall home-advantage controlling for differences in market value and other variables that affect the performance of a team. Consider the variable of interest excess goals of the home team, β_0 then describes the excess goals to be expected from the home team for two otherwise identical sides (with equal market values and importance of the match etc.). The crowd component, $-\beta_1$, can then only be identified due to the presence of matches with ban on supporters, which is identical to the econometric strategy exploited in chapter 1.

By adding an interaction term between Ban_i and Market Value Differences$_i$, I allow the treatment effect of empty stadiums to differ according to the relative strength of the involved teams (see discussion at the end of section 3). [9]

$$y_i = \beta_0 + \beta_1 Ban_i + \beta_2 (Ban_i \cdot \text{Market Value Differences}_i) + \beta_3 MVD_i + \text{controls} + \varepsilon_i$$

[9] More sophisticated approaches, such as allowing a quadratic or cubic functional form of the marginal effect, do not lead to different conclusions and the higher-order coefficients are all insignificant. For this reason, I keep working with a linear dependency.

The marginal effect of a crowd-ban would then be given by $\beta_1 + \beta_2 * MVD_i$. β_1 is expected to be negative, indicating an average decrease in the home advantage (see Figure 1). In the light of the narrative from above with underdogs being hurt more, the coefficient $/beta_2$ is expected to be positive, yielding a less severe reduction (or even increase) for favorite teams, i.e. whenever the market value of the home team exceeds that of the away team.

In columns (1) -(3) of Table 1, such models are estimated using excess points and excess goals of the home team as well as a binary variable for a home win as dependent variables. In all specifications β_1 is estimated to be negative and significantly different from zero. Quantitatively the estimates imply that for two otherwise identical teams, the home advantage shrinks by around 30-50% – approximately the magnitude of the crowd component argued for in section 3. Whenever β_2 is positive, however, the crowd-component also depends on the relative market values of the two involved teams. This leads to heterogenous estimates for the crowd component – smaller (even zero or negative) for dominant teams, larger for the underdogs.

	(1) $\Delta points$	(2) $\Delta goals$	(3) *win* (LPM)	(4) $\Delta shots$	(5) ΔxG
$Ban(\beta_1)$	-0.2245^{***}	-0.1065^{*}	-0.0289^{*}	-1.3330^{**}	-0.1682^{***}
	(0.0838)	(0.0577)	(0.0170)	(0.2940)	(0.0422)
$Ban \cdot \Delta MV(\beta_2)$	0.0005^{**}	0.0004^{**}	0.0001^{**}	-0.0000	0.0001
	(0.0003)	(0.0002)	(0.0000)	(0.0008)	(0.0001)
Constant (β_0)	0.4116^{***}	0.2936^{***}	0.4371^{***}	2.7334^{***}	0.3158^{***}
	(0.0610)	(0.0455)	(0.0121)	(0.2530)	(0.0336)
Controls	Yes	Yes	Yes	Yes	Yes
N	2,520	2,520	2,520	2,520	2,518
R^2	0.1735	0.1956	0.1365	0.3048	0.3153

*Notes: Controls include the market value difference of the two teams, the SPI difference, and the importance index difference between the home and away team whose coefficient estimates are all highly significant and positive as expected. Standard errors are robust and clustered at the Home Team level. * p<0.1, ** p<0.05, *** p<0.01.*

Table 1: Estimates for the Effect of Banning Crowds

It becomes apparent that the above suspected pattern also holds in a more careful regression analysis. The negative effect of banning crowds from stadiums is only present for teams that are not more than approximately 250 million stronger than their opponent. [10] Stronger teams are even estimated to benefit and this holds also when allowing the marginal effect to be a higher-order polynomial of market value.

Choking under pressure?

What drives this difference between favorites and underdogs? In line with the arguments from chapter 1, I argue that choking under pressure can be a relevant

[10] The threshold level of market value differences for the marginal effect to equal zero, depends on the exact specification and the variable of interest one is looking at. For all specifications it ranges by and large from 250 million to 450 million.

factor. There exist some studies on such motives in other sports (e.g. Toma (2017) or Böheim, Grübl, and Lackner (2019) for free-throws in basketball) as well as in football (e.g.Ferraresi and Gucciardi (2021)). This paper is to the best of my knowledge the first relating the magnitude of the crowd-component in the home advantage to different levels of pressure (measured through the favorite status).

In order to reconcile the heterogenous findings with respect to the crowd component, one would need to assume that for favorite teams the pressure dominates the supportive dimension of the crowd, while for underdog teams the opposite holds true. Possibly the supporters of a favorite team exert higher pressure due to higher expectations or a higher demand for entertaining and spectacular football (leading to suboptimal risk-taking by players of the home team). Supporters of underdog teams may well be less full of expectations and support their team more or less irrespective of the outcome.

To substantiate this claim, I investigate what happens to shots (and expected goals) of the home team after the introduction of matches behind closed doors.[11]

In column (4) and (5) of Table 1 one shots and expected goals are used as dependent variables. It becomes apparent that this time, no heterogenous effect with respect to the difference in market value can be established. All home teams, independent of how they compare to the away team, are estimated to attempt around 1.3 fewer shots than the away team once supporters are banned, approximately halving the previous home field advantage of around 2.7 shots. Similarly, the difference in expected goals decreases irrespective of market value differences by around 50% from 0.32 to 0.15.

[11]Since the match outcome is always a product of the performance of both teams, one may object that the home advantage is driven by a worse performance of the away team rather than an improved performance by the home team. The data do not support this claim and the changes in home team variables are the driving forces. It is empirically, however, impossible to ultimately ascribe that to the performance of the home team as better attacking statistics can likewise be caused by worse defending behavior of the away team.

This confirms the picture drawn in chapter 1. All home teams seem to suffer in the absence of the crowd when looking at shots. Some teams, however, keep on scoring the same amount of goals despite lower shots (and expected goals) and eventually do not suffer in terms of relevant performance measures. That can well be explained by released pressure that allows teams to play more serene, not being forced to entertain the crowd and hence not facing a disadvantage out of it. A more detailed look with more matches being played without crowd sheds light into what characterizes these teams. They are the favorite teams and if assuming that favorites experience higher pressure, it is only natural to see them performing no worse when the pressure is released by banning supporters.[12]

In order to substantiate this claim, I look at the conversion rate of expected goals into actual goals by the home team and the away. First, I run the simple regression

$$Home\ Goals = \beta_1 * Home\ Expected\ Goals + \varepsilon_i$$

for matches with crowd and obtain the following coefficients.

[12]It is also true that this can explain some of the cross-country differences observed in chapter 1. For example, the home teams in Italy were on average relatively stronger than their counterparts after the introduction of the ban of crowds.

	(1) Entire Sample	(2) $\Delta MV > 100$	(3) $\Delta MV < 100$
β_1	0.9146^{***}	0.9389^{***}	0.8912^{***}
	(0.0150)	(0.0240)	(0.0196)
N	1,442	492	950
R^2	0.7200	0.7569	0.6852

*Notes: * p<0.1, ** p<0.05, *** p<0.01.*

Table 2: Conversion Rates of Expected Goals into Goals with Crowd

For an additional expected goal in the match statistics, the actually scored goals increase by around 0.9. Home teams that exceed the away team's market value by 100 million EUR do convert chances into goals at a slightly higher rate.

For the underdog teams this is no different after supporters are banned from the stadiums. Yet, looking at home teams that are stronger than their opponent, the coefficient increases substantially to above 1.[13]

	(1) Entire Sample	(2) $\Delta MV > 100$	(3) $\Delta MV < 100$
β_1	0.9744^{***}	1.0231^{***}	0.9175^{***}
	(0.01864)	(0.0300)	$(.02413)$
N	1,082	383	699
R^2	0.7165	0.7524	0.6745

*Notes: * p<0.1, ** p<0.05, *** p<0.01.*

Table 3: Conversion Rates of Expected Goals into Goals without Crowd

[13] A very similar pattern is found for the conversion of shots into goals.

Hence, precisely the teams that are expected to benefit from the release of pressure display a considerable improvement in the conversion of their chances into actual goals. This explains well why shots and expected goals are estimated to decrease by roughly the same amount irrespective of differences in strength while the eventual performance does only decrease for underdogs. They actually benefit the most from their supporters and hence suffer the most once they are removed. Home teams that are considerably stronger on the other hand do not suffer despite fewer shots as they convert them at higher rates, which I argue to be a consequence of the released pressure (see also chapter 1).

5 A Betting Market Perspective

As mentioned above, the (sport) betting market many times acts as a sort of laboratory for economic theories. The underlying reasons are evident. Actual outcomes are typically independent of betting market behavior unlike in financial markets where financing decisions may eventually have an impact on the underlying ventures. Additionally, outcomes are easily and quickly observable compared to financial markets where uncertainty about, for example, a firm's value is at best resolved in the long run.

In particular, betting markets offer a natural setting to test (different versions of) the efficient market hypothesis according to Fama (1970). The weak version postulates that prices (here: betting odds) reflect all past information. The semi-strong form requires all public information, current and past, to be incorporated into prices while under the strong version all public as well as private information is reflected in prices.

Fischer and Haucap (2021) argue that the price reaction of betting odds violates market efficiency. I do not find evidence for persistent mispricing in the betting

odds due to the absence of the crowd-induced home field advantage. Instead, I show that such a possible inefficiency disappeared relatively quickly. Instead, I document that bookmakers did not account for the heterogenous pattern identified in section 4. Such ignorance would still violate market efficiency, yet in a more subtle way. Towards the end of the section, I show how such inefficiencies can potentially be exploited by according betting strategies in order to obtain positive average returns. The inefficiencies are however relatively small in magnitude causing potential gains to be largely eaten up by the margins bookmakers charge.

Before beginning with the empirical analysis, I briefly introduce the reader to the basic functioning of (fixed-odds) betting markets.

The Functioning of (fixed-odds) Betting Markets

Consider a football match with three possible outcomes $j = \{Win, Draw, Loss\}$ from the perspective of the home team.[14] For every single match i, each of the three events is associated with a (fixed) odd, which I call $O_{i,j}$. A bettor willing to gamble on such a bet places an amount A on outcome j and if outcome j realizes, he receives $A \cdot O_{i,j}$. Otherwise he receives nothing and simply loses the initial investment of A.

When bets are priced fairly, i.e. bookmakers (and bettors) make zero-profits in expectations, the $O_{i,j}$ reflect implied winning probabilities by the market. In practice, bookmakers will offer (slightly) lower odds to obtain positive profits and cover (transaction) costs. Let $p_{i,j}$ denote the probability of event j occurring in match i, then

[14] In the betting market industry the three events are usually referred to as Home, Draw, Away respectively. Needless to say that today's bookmakers offer possibilities to place bets on a variety of further outcome variables such as number of yellow cards, the exact result, or who scores a goal etc. that are ignored here.

$$Pr_{i,j}^{book} = \frac{1/O_{i,j}}{\sum_j 1/O_{i,j}}$$ [15]

In an efficient market the implied probabilities then would correspond to the (unobserved) true ex-ante probabilities of event j in match i $p_{i,j}$.

A consequence of this consideration is that in a regression that uses match outcomes as dependent variable, the implied bookmaker probabilities should be the only significant explanatory variable (with coefficient equal to 1 in a Linear Probability Model). Indeed, that is overall the case in the dataset I am using (see below). Neither market values, relative importance of the match nor importantly the ban of crowds turns out to be significant in a Linear Probability Model (see Table 4) supporting the idea of market efficiency. Instead the significant estimates documented in the literature for the initial matches after the ban of crowds disappear relatively quickly. [16]

This shows that (on average) bookmakers adjusted their odds correctly to what can be viewed as a decrease in the home field advantage. Possible inefficiencies were, if at all, relatively short-lived.

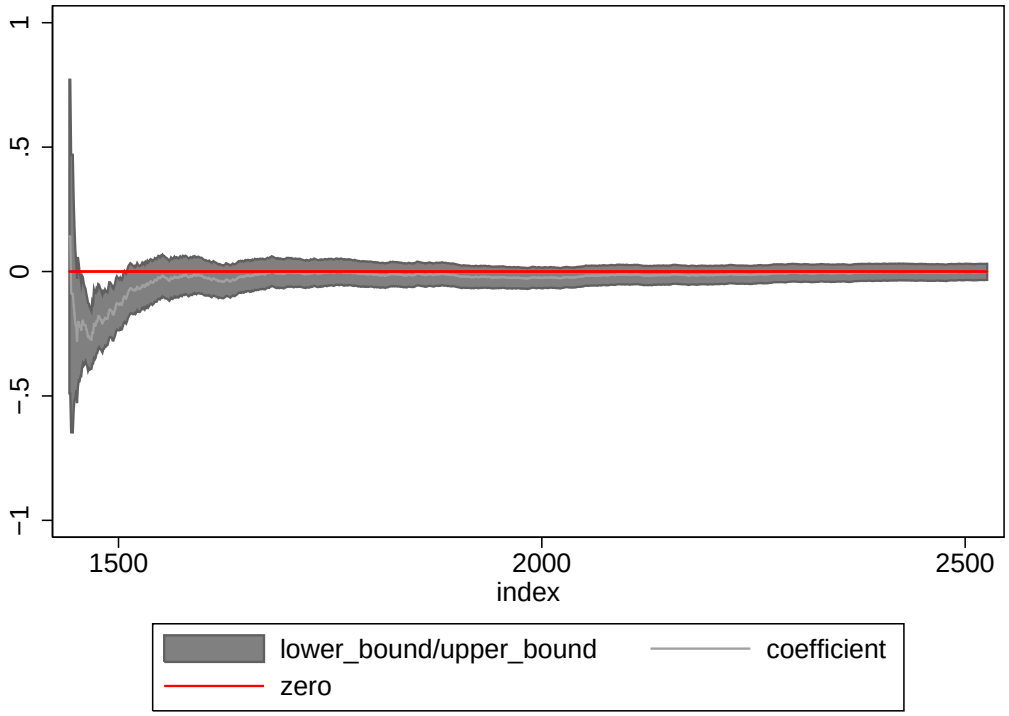
For the remainder of the paper, I mainly estimate versions of the following Linear Probability Model

$$Pr_i(j = Win) = \beta_1 Pr_{i,j=win}^{book} + \beta_2 Ban_i + \varepsilon_i$$

[15] In a perfectly competitve market without arbitrage or transaction costs, $\sum_j 1/O_{i,j}$ should equal 1. This is empirically, however, not the case and hence the above normalization ensures that the probabilities over all possible j sum to 1.

[16] There exist competing explanations for this phenomena in the literature. Either bookmakers learned from the gradually arising data and incorporated the new evidence quickly, or alternatively the streak of bad luck of home teams ended and the market appears again efficient but in fact has always been so.

which is identical to the one in column (1) of Table 4 below. I do so once after each single match behind closed doors is played, gradually adding the emerging evidence, and plot the obtained coefficient β_2 (over time) and the respective confidence interval in Figure 3. If the betting market were efficient, β_2 would be zero as such easily available information – whether a match is played with or without crowd – should have no predictive power regarding its outcome beyond what is implied by the bookmaker odds.



Notes: The index on the horizontal axis denotes the order in which the matches took place. at index = 1,442 the first game behind closed doors is played since from that point onwards such models can be estimated.

Figure 3: Estimate of β_2 with gradually arriving evidence

It reveals that the confidence interval does not include zero only in the very beginning. Eventually, the mispricing entirely disappears since after relatively few matches the point estimate approaches zero and the narrowing-down confidence interval is consistently including the theoretically hypothesized value of zero (red line).

Overall, the market seems to have incorporated available/ emerging evidence on the crowd component of the home field advantage relatively well. A possible under-reaction in the beginning (abstracting from a series of bad luck on the home teams' side) may as well be explained by a (conservative) updating of the available information. Indeed, possibly profitable betting strategies based on this inefficiency only yield a positive returns for the first matches after the introduction of the ban. This is in line with all versions of the efficient market hypothesis, as it can be reasonably argued that the actual effect size was unknown to everyone before the introduction of a wide-scale ban of crowds.[17]

(Not) Accounting for Heterogenous Effects

The main analysis in this chapter is centered around the question whether bookmakers did adjust their odds in line with the evidence on heterogenous effects across matches from above. If so, the interaction between the indicator for a match behind closed doors and the market value difference should as well be insignificant in the above mentioned regression. Therefore, in column (2) and (3) I run similar regressions adding as a candidate explanatory variables the interaction between the ban of crowds and market value differences. If markets were efficient all coefficients but β_1 would be insignificant.

Yet, it turns out that this is not the case. Across the specifications in Table 4, the coefficient on the interaction term is positive and significantly different from zero while all other coefficients are not.

[17] Previous studies have documented effects only under very particular circumstances and in relatively few matches, at least for the major leagues considered in this study (see discussion above and in chapter 1).

	(1) LPM	(2) LPM	(3) LPM
$Pr^{book}_{i,j=win}(\beta_1)$	0.9743***	0.9604***	0.9723***
	(0.0239)	(0.0251)	(0.0272)
Ban	−0.0010	0.0049	−0.0004
	(0.0166)	(0.0171)	(0.0177)
$Ban \cdot \Delta MV$		0.0001**	0.0001**
		(0.0000)	(0.0000)
ΔMV			0.0000
			(0.0000)
ΔSPI			−0.0012
			(0.0012)
$\Delta Importance$			0.0002
			(0.0003)
N	2,525	2,525	2,519

*Notes: The dependent variable in all specifications is a binary variable taking value 1 whenever the home team wins, 0 otherwise. * p<0.1, ** p<0.05, *** p<0.01.*

Table 4: Testing for Market Efficiency

The positive sign is in line with what one would expect if the market were to ignore the different effects between favorites and underdogs. Suppose a highly able (hence high market value) team is facing an underdog (with low market value) at home. The probability that the home team wins is high for mainly two reasons: it is more skilled and it also enjoys the home advantage. The latter does not disappear despite the exclusion of supporters (see section 3 and 4). The bookmakers, however,

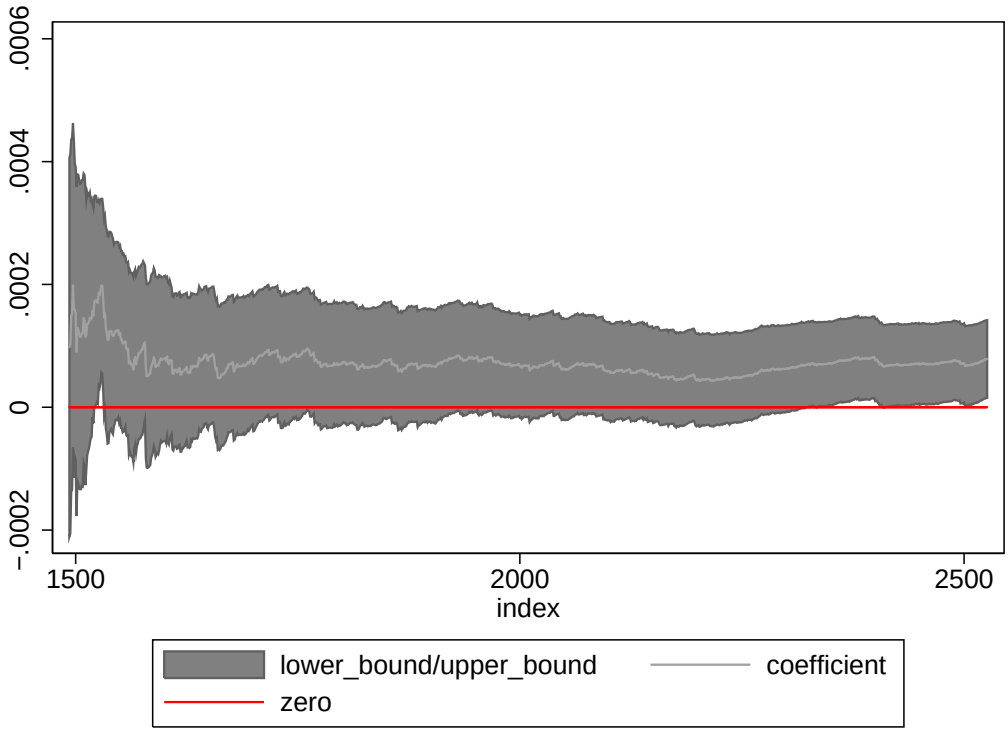
act is if the home team was negatively affected by the ban of crowds and consequently assign a lower winning probability to the home team. As a result the actual probability is above the implied one by the bookmaker. Vice versa, whenever the market value of the home side lies below that of the away team, the home team is actually hurt more than the bookmakers (who appear to use an average treatment effect) assume. Therefore in situations with the difference in market values being negative the actual probability of winning lies below the one of the bookmaker. This is line with the positive and significant coefficient of around 0.001 in Table 4. Across the various specifications, all other potentially predictive variables such as the SPI index difference or the difference in importance are insignificant, as suggested by the theory of efficient markets.

The estimated effect size suggests that a 100 million EUR difference in market value, would lead to an underestimation of the winning probability by 1%.

Next, I analyze the dynamics of this inefficiency. Does it behave as the supposed inefficiency regarding the pricing of the reduced home field advantage and becomes smaller over time and will eventually vanish or is it a stable pattern the betting markets overlook. For this purpose I repeat the regression from column (2) after each match that is played behind closed doors and again plot coefficient and confidence intervals over time.[18]

In contrast to the general home field advantage, the ignorance of heterogenous effects prevails even after hundreds of matches being played behind closed doors and does not decline/ approach zero with more evidence becoming available to bookmakers (and the public). At the end of the sample the confidence region does not include zero and the point estimate is significantly positive. This is in sharp

[18]I cut off the first 50 matches where confidence intervals are so large that they impair visibility of the rest of the graph. No significant estimates are obtained for these omitted matches and they are included as observations in what is depicted in Figure 4.



Notes: The index on the horizontal axis denotes the order in which the matches took place. At index = 1,442 the first game behind closed doors is played and from then on such models can be estimated. Here only matches from index= 1,492 are depcited for better readability. The estimated model is the Linear Probability Model from column (2) in Table 4.

Figure 4: Estimate of the Interaction Coefficient with gradually arriving evidence

contrast with even the semi-strong form of market efficiency. Easily available information on the relative strengths of the teams does posses predictive power for the match outcome that goes beyond that contained in the market prices.

Lastly, let me emphasize that this inefficiency is neither vanishing over time, which would be in line with market efficiency where bookmakers learn only over time, nor are the above regressions using insights/ information from future matches. Every single observation consists of the bookmaker implied probability, the eventual outcome, and (easily available) information on pre-season market values and Covid-related crowd bans.

Profitable Strategies

Having observed this, it seems natural to construct betting strategies that exploit this inefficiency and hence yield (on average) positive net returns. Intuitively, one should bet on matches behind closed doors in which the home team is (highly) favored, i.e. has higher market value than the away team. Likewise, one could bet on away teams when the home teams is characterized by a relatively lower market value than the away team. Overall, this strategy suggests to place bets on the favorites and resembles the advice under the presence of a favorite-longshot-bias. It is important to note that such bias is not present in the data prior to the introduction of crowd bans (see below).

First, I consider a strategy that places 1 EUR on a home win in all matches that are played behind closed doors and in which the home team has a market value that exceeds that of the away side by 100 million.[19]

In terms of market odds, the strategy yields a positive average return of 0.6%. If instead, the odds are corrected for the bookmaker's margin and *fair* odds are being used, the strategy yields an average return of 6.4%.[20]

This slightly positive return is not driven by a general favorite-longshot bias or other forms of mispricing in the market. In fact, applying the same strategy to the 492 relevant matches played with crowd yields negative returns.

[19] This threshold is of course arbitrary. If willing to assume that the bookmakers adjusted precisely according to the average effect, one should place a bet on all home teams with above average (= 0, due to symmetry) market value differences. Choosing 100 million can therefore be considered a relatively conservative strategy. Needless to say that the higher such threshold, the less bets that are actually placed.

[20] The correction happens according to the logic of the computation of implied probabilities. The *fair* odd is constructed as the reciprocal of the implied probability.

	Ban = 0		Ban = 1	
	market odds	fair odds	market odds	fair odds
Betting on Home Favorites	-10.9%	-6.5%	0.6 %	6.4%
Number of Bets	492		383	
Betting on Away Favorites	-8.4%	-3.9%	-0.7%	4.8%
Number of Bets	492		383	

*Notes: Favorites are defined as home teams that exceed the market value of the away team by more than 100 million EUR. The return is computed as that on placing a single bet with a fixed nominal value, 1 EUR for example, on all bets that meet the criterion of the strategy. Fair odds refer to odds that are computed such that under the implied probabilities expected profits equal zero. * p<0.1, ** p<0.05, *** p<0.01.*

Table 5: Returns of Different Strategies

Similarly, the market inefficiency suggests to place bets on away teams whenever the home team is weaker. Again, in my strategy I use a conservative threshold of requiring the market value difference to be 100 million apart, i.e. market value difference < -100 million. In terms of market odds, the return is close to zero, while it becomes clearly positive when using fair odds. Again a sizable improvement compared to the performance of the identical strategy prior to the ban.

Overall, the inefficiency allows to construct simple betting strategies that yield a slight positive return. The mispricing is, however, relatively small.[21] In terms of actual market odds that is certainly a bar to actually achieving profitability. Once, market odds are converted into profit-free fair odds, the returns turn positive.

[21] After all, the ban of crowds does not turn match prospects completely upside down. Hence, ignoring the heterogeneity still ensures more or less correct odds. Yet, they systematically miss the differences for balanced and unbalanced matches.

6 Conclusion

The contribution of this chapter is twofold. First, I show how the crowd-component depends on the relative strength of the involved teams. Favorite home sides maintain their previous performance-level even when the home crowd is removed for exogenous reasons. Underdogs, in contrast, experience a severe decline in their performance. I argue that this can be explained by the different level of pressure that such teams experience from the crowd. Favorite teams may well exert what is referred to as choking under pressure when their crowd is present and hence not leverage their full potential. Once removed, I observe heterogenous effects. Relatively stronger teams convert shots and chances at improved rates and hence do not display a worsened performance. The opposite is true for relatively weaker teams. They do not improve their shot conversion and overall perform poorer.

Second, I document that bookmakers do not account for this phenomenon. Instead they only correctly price the average reduction in the home field advantage due to the removal of the crowd component. Initial inefficiencies disappear relatively quick and market efficiency should not be rejected based on this factor. However, even after more than 1000 matches played behind closed doors in the top four leagues, the relative strength contains relevant information on the match outcome beyond the market implied probabilities, thereby violating even the weak form of market efficiency.

I show that this inefficiency allows bettors to construct slightly profitable betting strategies. Given the margins bookmakers have, the net return is however relatively small.

I leave for future research to discuss possible reasons for this inefficiency. In times of easily available statistical tools that allow to uncover such heterogenous effects

without deep understanding of the underlying mechanics, i.e. Machine Learning, it is at least surprising to observe such consistent mispricing. With possibly more matches being played behind doors there is also new evidence becoming available on learning about this effect by bookmakers. As I showed, the mispricing of the general home advantage disappeared after very few matches and might have initially been driven by a series of bad luck on the home teams' side. Finally, it would be worthwhile to analyze the effects in smaller leagues where inequality across teams might be much less pronounced and every team is somewhat similarly strong. In such environments the decrease in the home field advantage should be much more consistent across teams.

Besides that it is also an interesting challenge to investigate in greater detail how the crowd affects players' decision-making. With more sophisticated match-statistics one could analyze how risk-taking, concentration, physical effort and other more granular measures of performance reacted to the ban of crowds. It is at least to some extent surprising to see the documented effect on expected goals. Apparently some teams (and hence players) outperform their previous level of converting chances into goals when the crowd is not present. This gives rise to a sizable psychological/ behavioral dimension of performance. Situations are not entirely objectively comparable according to the physical circumstances. Rather, even elite football players seem to be affected by psychological factors in different ways. Some teams leverage the crowd-support leading to a considerable crowd-component in the home field advantage. But on the other hand, the crowd might lead players to choke under pressure and negatively affect their ability to convert chances into goals. Data on penalty conversion, where almost any physical differences are eliminated, could be an interesting opportunity if more detailed match-statistics are not easily available (see chapter 1).

In a broader context, such insights can have important implications for the design of incentive schemes. Increasing the stakes, traditionally associated with higher effort and hence performance, might result in unintended consequences. As even professional athletes seem to be affected by such adverse effects, it seems plausible that similar phenomena are present in a variety of other settings, too.

Regarding market efficiency, there are many directions for future research. The evidence in this paper is ambiguous. After some initial mispricing, the market learned well about the average decrease in the home field advantage. On the other hand, the heterogenous effect is more or less ignored. An interesting approach for future research would be to estimate team-specific home field advantages and observe whether they are previously, and eventually also after the pandemic, priced correctly.

References

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

Bryson, A., P. Dolton, J. J. Reade, D. Schreyer, and C. Singleton (2020): "Experimental effects of an absent crowd on performances and refereeing decisions during Covid-19," *Available at SSRN 3668183*.

Fama, E. F. (1970): "Efficient capital markets: a review of theory and empirical work', Journal of Finance, 25," .

Ferraresi, M., and G. Gucciardi (2021): "Who chokes on a penalty kick? Social environment and individual performance during Covid-19 times," *Economics Letters*, 203, 109868.

Fischer, K., and J. Haucap (2020): "Does crowd support drive the home advantage in professional soccer? Evidence from German ghost games during the COVID-19 pandemic," *CESifo Working Paper No. 8549*.

——— (2021): "COVID-19, Home Advantage in Professional Soccer, and Betting Market Efficiency," .

Hegarty, T. (2021): "Information and price efficiency in the absence of home crowd advantage," *Applied Economics Letters*, pp. 1–6.

Higgs, N., and I. Stavness (2021): "Bayesian analysis of home advantage in North American professional sports before and during COVID-19," *Sci Rep*, 11(14521).

Orefice, M. (2021): "Betting Market Efficiency during the COVID Crisis: Evidence from the 2020-2021 NFL Season," Ph.D. thesis.

Reade, J. J., and C. Singleton (2021): "Demand for public events in the COVID-19 pandemic: a case study of European football," *European Sport Management Quarterly*, 21(3), 391–405.

Scoppa, V. (2020): "Social Pressure in the Stadiums: Do Agents Change Behavior without Crowd Support?," *IZA Discussion Paper*.

Singleton, C., A. Bryson, P. Dolton, J. Reade, and D. Schreyer (2021): "What Can We Learn About Economics from Sport during Covid-19?," *Available at SSRN 3770193*.

Thaler, R. H., and W. T. Ziemba (1988): "Anomalies: Parimutuel betting markets: Racetracks and lotteries," *Journal of Economic perspectives*, 2(2), 161–174.

Toma, M. (2017): "Missed shots at the free-throw line: Analyzing the determinants of choking under pressure," *Journal of Sports Economics*, 18(6), 539–559.

Winkelmann, D., M. Ötting, and C. Deutscher (2020): "Betting Market Inefficiencies in European Football–Bookmakers Mispricing or Pure Chance?," .

Chapter 3

Quantifying Agreement in Peer Review Evaluation Across Disciplines[†]

[†]Author: Marco Wallner, wallner.marco@phd.unibocconi.it

Abstract

Evaluation through expert peers is a widely used method to assess projects of unknown quality, in particular in academia. Despite being so dominant, its critics claim that it does not produce accurate results, is unnecessarily using up resources, and is to some extent arbitrary and not reliable. I focus on the last point, namely to what extent peer review assessment of scholarly articles is reliable. Using data from a nationwide university evaluation exercise in Italy, I compute measures of inter-rater reliability and find that more scientifically oriented fields can be considered more reliable than social sciences and humanities. The commonly accepted intuitive distinction between high-paradigm and low paradigm fields seems to be supported by the data. In particular, methodological rigor is highly consensual in the high-paradigm fields pointing towards a lack of commonly accepted and utilized methodologies on other fields. The results can have important implications for the design of institutions that aim at guaranteeing a level playing field to a variety of research disciplines.

1 Introduction and Related Literature

Peer review has a dominant role regarding the evaluation of research and research proposals. Journals use peer review to select papers they publish, while funding bodies rely largely on the expertise of peer reviewers when deciding whom to award research grants. Guaranteeing efficient allocation of scarce resources (e.g. journal pages or funding) requires that the process of peer review produces accurate evaluations based on the merit of the item under review. Given the highly uncertain nature of the benefit of different types of research, this is naturally a challenging task including a variety of possible biases. It is therefore fundamental to understand in which dimension the evaluation through peer reviewers does not work perfectly well and how it can be improved upon.

Indeed, peer review is often criticized for not working well. Smith (2006) argues that despite being expensive and slowing down the publication process considerably, it is not evidently improving the quality of decisions or detecting flaws in academic work. Further, peer reviewers are found to be biased against certain characteristics or might use their power to their own advantage by blocking or stealing ideas from competing researchers. Importantly, he also argues that the *black box* of peer review is inconsistent and decisions partly resemble a lottery with the outcome being determined by who reviews a paper.

In this paper, I focus on this last point and, mostly descriptively, analyze to what extent peer review is unreliable and how this differs across disciplines. From a welfare point of view, the accurate selection of the best scholars and papers is instrumental for an efficient use of resources within science. Low levels of accuracy and reliability in peer review evaluation would clearly prevent an efficient allocation of resources across scientist and research projects. The truly most promising ideas might not be recognized as such and hence never become published or financed.

Moreover, the randomness in the selection of successful candidates or projects can have wide-ranging implication for the incentives of researchers to apply for funding or exert high effort in competitive contests. In a general equilibrium context such distorted incentives can push the allocation across fields towards the most inaccurate ones. Fields themselves then might face perverse incentives to lower the reliability of evaluation procedures to attract more funds and consequently award funds in less and less accurate and reliable ways. This point is theoretically discussed in Ottaviani (2020).

Hence, quantifying the level of reliability in peer review evaluation is a first step in order to tackle some of the problems mentioned by its critics. In addition, documenting differences between disciplines – which is the key contribution of this paper – can point out sources of heterogeneity across scientific fields that should be addressed by respective institutions in order to provide a level playing field for researchers from different backgrounds. This would also help tackling some of the unintended consequences implied by vastly different incentives across fields.

For example, fields with low levels of agreement among reviewers, might use more reviewers in order to reduce the randomness of their decision process. Similarly, fields with unreliable reviewers could improve their decisions by giving more weight to objective criteria.

However, objective criteria are limited in their availability, especially for decisions that are to be made ex-ante i.e. before the actual research is carried out. Ex-post, the success of a research project can (at least partly) be measured using bibliometric measures. Yet, institutions that award research grants need to evaluate the prospects of a project in advance and before eventual success can (imperfectly) be observed. Having said this, there exist few alternatives to using expert judgment in evaluating the prospects of a project in advance. Therefore, despite its critics, peer review will most likely continue to be a fundamental pillar in awarding and allocating research

grants and publication decisions. Taking into account the level of reliability of this process can help improving upon the realized allocations and partly address some concerns of the critics centered around reliability. In this paper, I aim at quantifying heterogeneity in reliability across different disciplines to provide an empirical basis for such a debate.

The underlying problem of unreliable expert judgment is of course relevant to a wider scope of problems than solely peer review in academia. Financial investors have to decide based on personal judgement which projects to finance, firms hire employees based on personal evaluations and universities admit students to their programs based on screening by faculty members. Naturally, every such expert judgement is to some extent subjective and imprecise. Similar methodologies can uncover heterogeneity in ratings also in these important areas.

This paper addresses to what extent disagreement is present in the evaluation of a paper across different scientific disciplines. Levels of agreement are likely to differ across disciplines, based on the ease of evaluating work in this area or the availability of established quality-criteria. In their seminal contribution, Zuckerman and Merton (1971) describe such institutional differences across disciplines based on varying rejection rates. They document that more experimentally and observationally oriented fields exhibit lower rejection rates than more humanistically oriented ones. They ascribe these differences partly to the different consensus on scientific standards across fields.

I use data from a national university evaluation exercise conducted in Italy between 2011 and 2014 (VQR). In particular, disagreement is computed between two independent reviewers evaluating the same article as part of the quality assessment of university departments on a scale from 3 to 30. The underlying idea is, that by looking at an identical item, any difference in evaluation must be due to subjective factors and noise; not due to varying quality of the underlying item. Intuitively, when

evaluating a paper in the absence of noise and subjective factors, two (unbiased) reviewers should reach exactly the same conclusion regarding its merit. However, since evaluation is noisy and subjective, their evaluation will usually differ to some extent. The magnitude of disagreement across the two reviewers is used to construct measures for the level of reliability in the evaluation process.

The outlined analysis adds to the existing literature in two important aspects. First, to the best of my knowledge, it uses less aggregated scientific areas (16 different fields) than prior studies. This allows for a more detailed analysis and can possible uncover important differences that vanish when using highly aggregated fields. Second, by considering data from a university evaluation, the problem of sample selection is mitigated. Sample selection can occur when using data from peer review procedures where applicants actively had to apply at a cost. For example, anticipating the high level of accuracy and agreement, controversial and less promising candidates might be discouraged from applying and would consequently not show up in the sample. In the VQR each researcher in the Italian university system is reviewed, independent of factors under her control.

The studies closest to this analysis are Pina (2015) and Mutz, Bornmann, and Daniel (2012). Mutz, Bornmann, and Daniel (2012) using data from the Austrian Science Fund find that agreement varies across research areas, with humanities displaying the highest level of agreement. It is worth noting that the patterns are found to be more or less stable over time suggesting that heterogeneity in agreement is something inherent to each research area, at least in the short(er) run.

Pina, Hren, and Marusic (2015), even though not their main research question, document correlations among reviewer scores for different fields in the Marie Curie Actions peer review process. They find relatively high levels of agreement and only small differences across disciplines. This rather high level of agreement could, according to them, be a result of the high quality of applicants. Indeed, in their data,

agreement is lower for proposals with lower final scores pointing towards more agreement for high quality proposals. This is in contrast to other studies who find higher levels of agreement on lower quality items (e.g., Cicchetti (1991) finds more agreement on rejections than on acceptance).

Further studies documenting different levels of agreement across disciplines are Nicolai, Schmal, and Schuster (2015) who find low levels of agreement in both Management Journals as well as in the high-paradigm disciplines chemistry and physics. Traag, Malgarini, Cicero, Sarlo, and Waltman (2018), using a similar dataset (VQR), find varying levels of correlation between the two reviewers' scores. They find the level of correlation to increase when aggregating items on an institutional level rather than on individual paper level. An early reference on field differences can be found in Jayasinghe, Marsh, and Bond (2003), who find only a negligible differences between social science and humanities compared to science. In a meta-analysis, Mutz, Bornmann, and Daniel (2012) use the results of 48 studies on reliability of the peer review process. On average there exists a substantial level of disagreement. Importantly, the field of the journal does not seem to affect the level of reliability in their meta study.

Overall, the existing literature is rather inconclusive regarding the agreement among expert reviewers in different fields. A serious threat to the external validity of such studies is the use of truncated samples. As mentioned by Pina, Hren, and Marusic (2015), the quality and composition of the applicant pool is likely to affect the observed level of agreement. Further, comparability is complicated by differences in the definition of research areas. Funding bodies and other institutions use various categories to group researches into fields. They are, however, not always perfectly compatible and hence results are difficult to compare.

Let me mention that this work does not relate to the strand of literature concerned with biases in evaluation. Bias can occur against certain minorities, topics etc. There

exists an exhaustive literature that I do not address here, mostly because in the absence of a true measure of quality an eventual bias is not detectable. In the absence of such superior information, I am assuming that every rating is unbiased but noisy – for the mentioned reasons of various biases an often times unrealistic assumption.

In the sample of VQR evaluations, I find the 16 fields to differ in the level of agreement among reviewers. Fields that can intuitively be classified as high-paradigm display higher levels of agreement among reviewers than low-paradigm fields.[1] In Mathematics/ Computer Science agreement among reviewers was more than 35% higher than expected by chance (Cohen's Kappa = 0.36), while for fields such as Law, Economics, and Political and Social Sciences this value is below 10%. Computing Intra-Class Correlations (ICC) across the different fields supports this conjecture. I also documents that evaluation regarding methodological rigor plays an important role for the level of agreement and is most often the most consensual dimension, in particular in fields with high agreement.

The paper is organized as follows. In section 2, I discuss the data used for the analysis of Inter-Rater Reliability in section 3. In Section 4, I discuss the results and provide some evidence on the role of methodological rigor compared to other aspects. Section 5 concludes.

2 Data

Institutional Framework

The Italian National Agency for University and Research Evaluation, ANVUR, is in charge of conducting a nationwide evaluation of the research conducted within

[1]The distinction of fields according to paradigm was famously introduced by Kuhn (1962). It has since then been used to distinguish disciplines according to the extent common assumptions and methodologies are established (see for example Lohdal and Gordon (1972))

all Italian university departments. The results are used to distribute up to 20% of public university funding and thereby create incentives for faculties to improve their research performance. The evaluation explicitly does not consider any other dimension of university quality such as teaching quality or job market success of students.

In order to evaluate university departments, ANVUR has established a large-scale peer review process. All scholars employed by a university are asked to submit two research articles published in the period under consideration (2011-2014).[2] The papers are then evaluated within 16 Groups of Experts (GEV). The respective group is chosen to match the author's active area of research. These groups play the role of the different scientific disciplines in my analysis. Table 1 lists the different disciplines together with their size, measured as the number of submitted papers, the number of peer reviewed papers and the size of the sample I was able to obtain ($\approx$ 10% of peer reviewed items).

[2]Researchers that are employed at a research institute rather than a university are asked to submit 3 items.

GEV	Field	Number of Outputs... submitted	peer reviewed	sampled
1	**Mathematics and Computer Science** (MAT)	6,062	2,356	228
2	**Physics** (PHY)	10,588	1,291	129
3	**Chemistry** (CHEM)	6,897	1,394	142
4	**Earth Science** (EARTH)	4,430	1,257	123
5	**Biology** (BIO)	10,986	2,183	212
6	**Medicine** (MED)	16,693	3,731	390
7	**Agricultural and Veterinary Sciences** (AGRI)	7,541	2,463	245
8.a	Architecture (ARCHI)	3,433	3,433	343
8.b	**Civil Engineering** (CIV ENG)	2,832	996	99
9	**Industrial and Information Engineering** (IND ENG)	11,564	3,346	332
10	Ancient History, Philology, Literature and Art History (LIT)	8,744	8,720	857
11.a	History, Philosophy, Pedagogy (HIST)	6,123	5,956	585
11.b	**Psychology** (PSY)	2,276	868	84
12	Law (LAW)	8,488	8,431	819
13	**Economics and Statistics** (ECON)	8,385	2,662	257
14	Political and Social Sciences (POL)	2,971	2,953	294
		118,036	52,040	5,140

Table 1: List of the Scientific Fields and their size in the VQR2 Evaluation; bibliometric fields in bold

Notes: Behind the name of each field I introduce a shorter abbreviation used throughout the rest of the paper for better readability.

These 16 fields follow the Italian National University Council classification of academic fields.[3]. The bold fields are bibliometric fields in which a majority (94%) of submitted items are journal articles found in citation databases, whereas in non-bibliometric fields less than half (43%) of the submitted items are published papers. ANVUR evaluates submitted journal articles in bibliometric fields mainly using an algorithm which combines citations and journal impact of the publication. In non-bibliometric fields each output is evaluated through peer review. Yet, there exists also a considerable number of research outputs in bibliometric fields that underwent peer review. This happens for several reasons; (i) a random sample of about 10% is subject to peer review due to an experiment conducted regarding the agreement of bibliometric indicators and peer review evaluation [4] (ii) a part of the submitted research items such as book chapters or monographs are not captured in bibliometric databases and (iii) some journal articles received an inconclusive bibliometric evaluation, e.g., high number of citations in a low ranked journal or vice versa. Column 3 lists the number of research outputs that were subject to peer review. I have obtained a random 10% sample of these research outputs in each field. The number of observations is rarely below 100 per field.[5]

Peer Review Protocol

Every paper that is evaluated through peer review is usually sent to two reviewers.[6] Each reviewer is then asked to evaluate the assigned output according to three criteria

[3]The detailed list can be accessed via `https://www.cun.it/documentazione/academic-fields-and-disciplines-list/`

[4]This is important due to the allocative role the evaluation plays. Ensuring equal chances and comparability across all fields is crucial for acceptance of such allocation rules.

[5]There exist several subfields within each field that undoubtedly would be interesting to distinguish. Unfortunately, the data do not contain information on the assignment to the various subfields.

[6]In a small number of cases (13.6%) the peer review is done directly within the GEV. Sometimes, due to external circumstances such as delay in delivery, a third reviewer is consulted. These deviations from the standard protocol are of little relevance for my analysis and my dataset contains the relevant two referee scores only ignoring who carried out the evaluation.

on a scale from 1-10 and to provide a written comment on the assigned scores. The three criteria are: *originality*, *methodological rigor*, and *attested or potential impact*. According to a pre-defined mapping, each score (between 3 and 30) is eventually converted into one of the five categories: Excellent, Good, Fair, Acceptable, Limited.

The reviewers are chosen according to their expertise and hence there exists a certain degree of matching quality among the papers under review and the reviewers. Having said that, my estimates on agreement can be considered lower bounds as agreement is expected to be lower if reviewers have to evaluate papers outside areas of their particular field, e.g. a macro-econometrician evaluating a game-theoretical paper in field 13 (Economics). Also reviewers are enabled to obtain a signal about the quality of the paper by observing the outcome of the publication process, mostly known at the time of evaluation. This would further reduce noise and increase agreement, emphasizing that my estimates play rather the role of a lower bound on consensus.

These five categories are identical to the ones used throughout the bibliometric evaluation process. The bibliometric algorithm combines number of citations and the ranking of the journal to assign one of these categories. There exist various other studies that investigate to what extent the peer review process delivers similar results when compared to the bibliometric evaluation (e.g., Bertocchi, Gambardella, Jappelli, Nappi, and Peracchi (2015) find high levels of consensus in area 13 (Economics and Statistics). At least on an aggregated institutional level, bibliometric and peer review assessment seem to suggest similar evaluation results in the VQR context (Traag, Malgarini, Cicero, Sarlo, and Waltman (2018)).

My analysis, in contrast, focuses on the reliability of the peer review process itself, ignoring possible discrepancies compared to bibliometric or other forms of evaluation. I also do not address the appropriateness of each rating that a large strand of related literature focuses on. For example Li and Agha (2015) find that

successful applications consequently achieve higher citations and better publications, pointing towards an important informational component in the peer review selection, based on merit.

Obtaining the scores that the two reviewers assign to an identical research item across three dimensions of evaluation together with the resulting final score allows me to study to what extent they agree on the merit of a particular research output. Further, it allows to distinguish which of the three criteria induces the largest disagreement and to what extend evaluation features different properties across fields. It is important to note that the key to this analysis is that two experts evaluate the same item. This allows to disentangle varying underlying quality (i.e. the papers differing in quality) from evaluation noise generated by disagreeing opinions.

Descriptive Analysis

In order to understand how expert opinion on quality is distributed within the different fields, let us begin by first documenting differences in the average evaluation across fields. For the moment, we ignore that each final evaluation of a paper consists of two individual raters' opinion. Naturally, we observe disciplines to differ in the distribution of scores. The drivers of this heterogeneity potentially are twofold. First, the distribution of actual quality may be different across fields.[7] Then, even under perfectly informative signals, the resulting evaluation is heterogenous across fields, mirroring the differences in quality distribution. It seems natural to expect that fields differ in the distribution of quality. For example, some fields might have very few highly influential papers of highest quality that are substantially better than an average paper, while in others the differences between papers are not as pronounced and hence a large body of rather average quality papers exists.

[7] With (actual) quality, I am referring to the (unobservable) true merit of a paper that would be observed if signals about quality were not noisy.

Second, evaluators in different fields are likely to differ in their assessment of quality. They could posses different biases towards others' work or be more generous in one discipline than in another. This would cause, even under the same distribution of quality (i.e. the above mentioned first source of differences being shut down), scores to differ across fields. Importantly, evaluators can also differ in the reliability of their evaluation. This is precisely the factor this paper tries to establish. This is however not necessarily visible in the distribution of average evaluations across two reviewers, but can only be detected by investigating the discrepancy between two reviewers of the same paper.

Having said this, the overall distribution is a result of multiple factors that most likely differ across fields. Therefore, the distribution of observed scores is merely a descriptive exercise to summarize some of the differences across fields and highlight that fields are not homogenous in the overall evaluation – even under a fairly standardized protocol of evaluation within a single institution. Such differences give rise to the question in consideration if and how the agreement between two raters differs across disciplines.

In Table 2, some moments of the respective distributions are depicted for each field. Most strikingly, the average scores vary in a range from 13 points (Economics) to 22 points (Chemistry) . This is at least surprising in the light of the comparable evaluation criteria and procedures. The differences suggest that on average fields such as Economics and Psychology award lower scores than fields like Chemistry or History. However, such differences should be interpreted carefully as they are a result of different habits and cultures (such as level of generosity or bias against peers), different evaluation precision and lastly different underlying qualities across disciplines.

Field	Mean	S.D.	p50	p90	kurtosis
1 MAT	20.51	6.25	22	27.50	2.93
2 PHY	20.80	5.52	22.50	26.50	3.68
3 CHEM	22.01	4.83	23.00	27.00	6.48
4 EARTH	18.59	6.31	20.00	26.00	2.36
5 BIO	19.85	4.86	20.75	25.50	2.82
6 MED	18.32	5.69	19.50	25.50	2.46
7 AGRI	17.68	6.30	19.00	25.00	2.22
8.a ARCHI	20.30	5.25	21.00	26.00	3.33
8.b CIV ENG	18.92	5.78	19.50	26.00	2.28
9 IND ENG	19.02	5.34	20.00	25.00	3.11
10 LIT	22.43	4.51	23.00	27.50	4.22
11.a HIST	21.78	4.68	22.50	27.00	3.98
11.b PSY	17.81	6.04	18.50	25.50	2.42
12 LAW	20.97	4.66	22.00	26.00	3.80
13 ECON	13.26	5.50	13.50	20.50	2.34
14 POL	19.18	5.58	20.50	25.50	2.83
Total	20.10	5.62	21.50	26.50	3.10

Table 2: Descriptive statistics of the score distribution by field

Differences in the distribution of scores are also present in the tails. The 90th percentile varies between 20.5 and 27.5. This is particularly important for funding and publication decisions where mainly researchers of high quality, those in the right tail, compete. Suppose a potential candidate from GEV 1 and a potential candidate from GEV 14 each consider applying for a contest where they need to end up in the top 10% of applicants. If they would use the distribution of VQR evaluations as an assessment of the hurdle they need to surpass, the applicant from GEV 14 would have in mind needing to receive a 25.5, while for GEV 1 this value is 27.5.

Lastly, the standard deviation of assigned scores also differs substantially across fields. Some fields display a standard deviation of above 6.3, while others are characterized by standard deviations of only around 4.5. This should, however, not be misinterpreted as reflecting precise or imprecise evaluation as it may as well be caused by differences in the underlying quality.

To sum up, this preliminary descriptive analysis of the score distribution across fields suggests that peer review assessment is characterized by cross field heterogeneity. Differences occur in the overall level of assigned scores as well as in the behavior regarding the tails of the distribution and its variance. In the next section, three different measures of agreement are applied to the VQR peer review data. The goal of the analysis is to construct a measure of the reliability of evaluation, isolating dispersion of expert opinion from other factors that affect the distribution of scores.

3 Inter-Rater Reliability

The problem of assessing the reliability of expert opinions or other measures is a well-known problem in many fields. In medicine, doctors evaluating the health condition of the same patient are required to reach similar conclusions, while in engineering and related fields certain measurement tools should ideally reach a similar and hence reliable result when measuring the same object. Hence there exists a variety of approaches to quantify the degree of agreement between two raters. I will investigate three commonly used approaches to conclude which disciplines are characterized by a relatively more reliable peer review procedure. In these fields experts agree to a larger extent.

Deviation between Reviewer scores

A first straightforward approach in order to quantify the extent of (dis)agreement, is to compute the absolute deviation of the two reviewers' evaluations from the mean or median of their ratings. This approach, suggested by Burke, Finkelstein, and Dusig (1999) and used in the related study by Pina, Hren, and Marusic (2015), is intuitively appealing. One can expect the deviation between the two scores to be low when evaluation is reliable. In such cases, every reviewer assigns a similar rating and consequently they only differ by a small amount.[8] In contrast, when evaluation is unreliable, the two raters may assign conflicting ratings leading to high deviations. Note that for the case of two raters, the deviation from the median score is the same as from the mean, which is why I will simply report the average absolute deviation of the two scores itself.

The results are depicted in in the first column of Table 3. It is worth noting that in disciplines intuitively characterized as being high-paradigm, this measure takes on lower values. In low-paradigm and more controversial fields, the opposite is true. For example, in area 14 (Political and Social Sciences) or area 13 (Economics and Statistics) a change in the rater comes with an expected adjustment of the score by around 6 points, while the same change in rater would only lead to an average change in score of around 4 points when taking place in fields like Chemistry or Physics.

While this measure is admittedly simple, it nevertheless captures one important aspect of reliability: by how much can a researcher expect an individual rater score to change if that rater is randomly replaced by another expert. Drawing conclusions on how reliable peer evaluation works across disciplines is, however, most likely

[8] In our setting, every rating is (on average) deemed appropriate, i.e. unbiased, as we have no further information, such as, for example, future impact of the paper, to evaluate if a rating actually reflects the true merit of the article under review.

premature as each field is characterized by a different distribution of scores (see above).

One immediate concern is related to the different variability of scores across the disciplines. Within chemistry (area 3) the standard deviation is relatively low, implying that papers receive more similar scores compared to areas with higher standard deviations (see Table 2). This of course contributes to observing what might look like a reliable evaluation of research articles. However, it might be misleading in the following sense: In disciplines with low variability of evaluation scores, i.e. where there is less dispersion of quality overall (as measured by the average scores), it is natural to observe raters awarding scores close to each other. This is maybe best understood by considering a (hypothetical) discipline where raters all award the same score to each paper and hence the standard deviation of scores is zero. Obviously, in such a discipline every rater submits the same score and hence the process looks reliable. The very concept of reliability, however, refers to the ability to recognize differences in quality reliably, i.e. agree on what is high and what is low quality – not agreeing on that everything is average. While being rather extreme, this example highlights the need to control for the variability of quality in the ratings. As an example, we can take the average absolute deviation of field 1 and 3 that amounts to roughly 4. However, we might want to conclude that raters in 1 displayed more agreement for the following reason. In field 1 scores in general vary more (see standard deviation from Table 2). Hence reviewers in field 1 face more uncertainty about the underlying quality, while reviewers in 3 must expect that the actual quality is somewhat close to the mean and submit ratings accordingly.

Field	Av. Dev.	ICC	Cohen's Kappa
1 MAT	3.99	0.66	0.36
2 PHY	4.25	0.60	0.19
3 CHEM	3.99	0.54	0.08
4 EARTH	4.46	0.64	0.26
5 BIO	5.00	0.36	0.15
6 MED	5.90	0.40	0.09
7 AGRI	5.29	0.55	0.21
8a ARCHI	4.91	0.47	0.11
8b CIV ENG	4.91	0.52	0.28
9 IND ENG	5.42	0.40	0.11
10 LIT	5.16	0.27	0.06
11a HIST	4.64	0.38	0.13
11b PSY	6.07	0.45	0.12
12 LAW	5.07	0.35	0.09
13 ECON	6.48	0.29	0.09
14 POL	5.83	0.34	0.13

Table 3: Measures of Agreement across Disciplines

Notes: All ICCs are statistically different from 0, and 95% confidents intervals do not overlap for a considerable amount of fields.

Intraclass Correlation

To address this concern, one of the most widely used approaches is to compute intraclass correlations. It is a correlation with respect to measures within the same class, here: the repeated (by different reviewers) measure of quality of a paper. The idea of this correlation measure is to quantify what fraction of variability in that repeated measurements is due to actual information and not due to noise or measurement error. Introduced by Shrout and Fleiss (1979), it is among standard

approaches to quantify the level of agreement by two (or more) experts and is widely used in the context of evaluating agreement of reviewers in different contexts.

The general setting is a R times repeated observation of a value y_{ir} in a sample of N objects i (often called targets). In this case $R = 2$ and refers to the different judgements by the two reviewers r.

For what follows I will use the One-Way Random Effects Model of the ICC, denoted by ICC(1,1). A discussion of the various different models can be found in Liljequist, Elfving, and Skavberg Roaldsen (2019). In the Appendix I describe in some detail how the observed data are used to compute the ICC. The ICC(1,1) model builds on a simple linear model where individual reviewer ratings are the sum of a true quality measure s_i and some error made by the reviewer that are added to the average quality μ

$$y_{ir} = \underbrace{\mu + s_i}_{\text{true quality}} + \underbrace{\varepsilon_{ir}}_{\text{error}}$$

with s_i being the effect of subject i (the quality of the research paper) and ε_{ir} being a rater and subject specific error term of rater r for subject i, both mean-zero normally and independently distributed with variances σ_s^2 and σ_e^2 respectively. Again, this assumption implies that each rater observes (and communicates) an unbiased estimate about the true quality of the paper.

The intraclass correlation coefficient ρ between the rating of one reviewer y_{ir} and that of another $y_{ir'}$ is given by

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}.$$ [9]

[9] This version of the ICC is referred to as One-Way Random-Effects Model. It does not control for effects of raters having different biases or lenience as we do not observe the identity of the raters. The sample variances are computed according to the ANOVA framework.

The two variances are computed using the matrix of $N \cdot R$ matrix of observed ratings. σ_e^2 is simply the Mean Squared Deviation Within targets, while σ_s^2 is computed using the Mean Squared Deviation Between Subjects. For a detailed description of the empirical estimation see the Appendix.

It is immediate to note that ICC is high, i.e. close to 1, whenever the variance across subjects σ_s^2 exceeds the error variance σ_e^2 by a large amount. This captures precisely the idea of high and low inter-rater reliability. Whenever differences across subjects (through σ_s^2) dominate the differences due to the assigned reviewer (through σ_e^2), the evaluation process is considered reliable. Subjects under review can then expect their rating to be determined largely by its merit and quality and less so by who evaluates their work. Contrarily, when the estimated variance of the error is high relative to the variance of true quality, the evaluation process is considered unreliable. Differences in true quality then play only an inferior role in the determination of the eventual rating of each reviewers rating y_{ir}. The subjective component is then the dominant factor.

The results in column 2 of Table 3 confirm the intuitive and previously found distinction. Fields, traditionally characterized as high-paradigm fields, are found to be more reliable when it comes to peer evaluation. The ICC, however, measures not only how disperse the scores are across the two raters of the same paper. It also takes into account to what extent quality of papers varies within the different fields. A field with relatively more similar papers (in terms of quality, measured through peer review) displays, everything else equal, lower reliability than a field with higher dispersion of paper quality. Yet, the hard scientific fields remain to appear relatively more reliable than the softer social sciences.

Cohen's Kappa

As a last descriptive measure I consider Cohen's Kappa. The main difference to the two previous statistics is that it can be computed for categorical variables. Often peer review evaluations are not made on a continuous scale but rather feature a classification into nominal variables (e.g., accept vs. reject, or very good, good, etc.). Since the precise evaluation of VQR raters is eventually translated into one of the five categories mentioned above (Excellent, Good, Fair, Acceptable, Limited), we can also construct this measure for the VQR dataset, providing results that are quantitatively comparable to a larger set of related studies often forced to use categorical judgements on quality. [10]

Technically, Cohen's Kappa measures the percentage of subjects under review on which reviewers agree, i.e. submit the same rating, in their evaluation. This percentage is then adjusted for the expected agreement due to chance.[11] Cohen's Kappa is computed as

$$\kappa = \frac{p_O - p_E}{1 - p_E}$$

where p_0 is the observed fraction of agreement and p_E the one to be expected due to chance. The computation is again described in some detail in the Appendix. With agreement we mean that both raters indicate a score that falls into the same category, hence award the same rating in terms of the nominal categories. Higher values indicate that observed agreement exceeded the random benchmark by a larger margin. Again, this measure takes into account how dispersed the evaluation is

[10] The mapping happens in the following way: Excellent: 30, 29, 28, 27; Good: 26, 25, 24, 23, 22; Fair: 21, 20, 19, 18, 17, 16; Acceptable: 15, 14, 13, 12, 11, 10, 9, 8; Limited: 7, 6, 5, 4, 3.

[11] If one were to re-shuffle all ratings to random items, there would (depending on the number of nominal categories and the frequency they were awarded) still be cases in which the reviewers agree despite no actual consensus.

across the five categories within a discipline. If a large fraction of papers receives one certain classification, agreement due to chance is higher and therefore Kappa is, everything else equal, lower. The interpretation is therefore similar to that of the ICC. It measures the agreement that raters are able to reach, controlling for how different or similar the objects under review are. Of course, if (almost) all papers would fall into one category, agreement would be mechanically high despite the process not being particularly reliable (see the hypothetical example from above where all raters award the same score).

The last column of Table 3 depicts the values of Cohen's Kappa across the 16 areas of research from the VQR. In Mathematics and Informatics, field 1, reviewers' classifications agree in more than 50% of the cases, while by chance only around 24% of agreement would be expected. This leads to the relatively high value of 0.36 for Kappa. Other fields display much less agreement, as is implied by the other measures as well. In, for example, Classical, Philological-Literary and Historical-Artistic sciences, field 10, agreement happens in less than 33% of cases. With expected agreement being 28%, Kappa amounts to only 0.06. Overall, agreement according to this measure is relatively low.

Only some fields can be classified to show more than slight agreement according to the threshold value of 0.2 of Landis and Koch (1977). These fields are Mathematics (1), Earth Science (4), Agricultural and Veterinary Sciences (7), and Civil Engineering (8b). Among the fields with the lowest Kappa values are Chemistry (3), Medicine (6), Ancient History, Philology, Literature and Art History (10), Law (12), and Economics and Statistics (13).

Overall, this confirms the intuitive idea that reviewer agreement is higher in natural sciences, as also documented in other studies (e.g., Pina, Hren, and Marusic (2015)). Among the highest agreement fields there are no fields from the are of the humanities. They are rather found to show low values of Kappa together with low

ICC. At the same time the natural sciences with low Kappa, display relatively higher ICCs. This may be a result of the in principle arbitrary thresholds used to define the five final categories. In Chemistry, where I observe 142 research articles, it might have simply happened more frequently that two rather similar but not equal ratings (out of 30) fell into two different categories. This could explain why Kappa is low despite the high level of ICC.

4 Discussion of the Results and Possible Causes

Section 3 makes it evident that concerns regarding varying degrees of agreement among experts across disciplines are well justified. Intraclass Correlations display striking differences and in fields such as Economics and Statistics the variance in true quality is estimated to only account for around one third in the observed differences in evaluations, while the remaining two third are due to judgement error by experts. In fields like Mathematics (1) with ICC above 0.6, the ratios are reversed and variance in true quality accounts for two third of the observed variance in observed ratings. This finding translates also in the agreement regarding the eventual classification of an object. In fields with high ICC agreement is often more than 20-25% higher than expected by chance, while the agreement does not strikingly outperform chance in all fields with low ICC, hence Kappa being close to zero.

In this section I exploit the fact that final ratings consist of three single ratings regarding three distinct dimensions of quality. Each paper receives a score from 1 to 10 on the three criteria

- Originality
- Methodoligal Rigor
- Attested or Potential Impact

Field	Originality	Meth. Rigor	Impact
1 MAT	0.596	**0.637**	0.590
2 PHY	0.473	**0.580**	0.527
3 CHEM	0.470	**0.544**	0.458
4 EARTH	0.614	0.541	0.539
5 BIO	0.351	0.266	0.343
6 MED	0.395	0.357	0.355
7 AGRI	0.475	**0.567**	0.474
8a ARCHI	0.428	0.423	0.440
8b CIV ENG	0.469	**0.561**	0.418
9 IND ENG	0.387	**0.400**	0.307
10 LIT	0.243	**0.268**	0.230
11a HIST	0.347	0.340	0.379
11b PSY	0.426	0.411	0.444
12 LAW	0.362	0.312	0.318
13 ECON	0.247	**0.307**	0.232
14 POL	0.277	**0.345**	0.333

Table 4: Intraclass Correlation for the Different Evaluation Criteria

Notes: Each value corresponds to the ICC(1,1), computed as in Table 3, but separately for the three categories; each with a score in the range of 1 to 10.

In Table 4, I compute ICC for each of the 3 criteria separately and by field. Naturally, the fields with overall higher agreement (see Table 3), display higher agreement also within the distinct categories.

Interestingly, fields that are low-consensus in general, are also characterized by similar low levels of agreement when it comes to methodological rigor. This is somewhat surprising given that methodological rigor is a relatively uncontroversial dimension, unlike originality and impact. Yet, experts in high-consensus do not only display higher levels of agreement on subjective dimensions, but in particular agree more on what concerns the methodological rigor of an object.

This suggests that these disciplines have better established methodological criteria that experts know and commonly use for evaluation. These established criteria then might also be useful for evaluating impact and originality, but still the disagreement in these categories is naturally lower.

To sum up, the differences in agreement prevail also when it comes to established methodological standards. Impact and Originality are naturally more controversial, even for fields that overall display high agreement. This makes the high- and low-paradigm distinction that has been used in many contexts even more convincing. High Paradigm disciplines are truly characterized by a high agreement on rigor compared to low paradigm disciplines, i.e. an established standard on the usage of different methodologies. It is therefore not only noise that comes from evaluation of subjective criteria that distinguishes these fields, but also (and even particularly so) the agreement methodological rigor.

Overall, this quantifies the intuitive wisdom that more scientific fields are characterized by experts that do agree more on the merits of an object under evaluation. The hypothesized role of established methods in these areas is confirmed. Experts in fields with high agreement, also agree more on the methodological rigor. This sheds some light on the heterogenous reliabilities of peer review across disciplines with potentially important implications for funding bodies etc.

5 Conclusion

I have discussed the issue of reliability in the context of peer evaluation in academia. The focus on differences across disciplines highlights that scholars from different thematic backgrounds evaluate their peers' work with different levels of agreement. More scientifically oriented fields are characterized by higher levels of reliability. This

finding seems to be in line with common sense. Softer disciplines are more subjective and lack well-established evaluation criteria compared to the harder natural sciences.

This insight can be used by institutions that run competitions among researchers from different backgrounds to ensure comparable evaluation practices for each researcher, irrespective of her field.

An interesting extension of this analysis would be to not only estimate field-specific differences but also rater-specific heterogeneity. Some raters may well exert more effort or be more talented in evaluating projects. Identifying those could be a useful endeavor and also help aggregating scores in a more efficient way. Certainly more future research can be done in this area, using richer datasets that allow for the identification of, for example, individual rater biases.

Lastly I want to draw attention to the fact that this pattern is not necessarily restricted to peer review in academia. Experts judging the quality has become a ubiquitous feature of the world and in particular the online world. Ratings on products, hosts or drivers are used by millions of users to guid their purchasing behavior. Taking into account that some products or services may inherently be more difficult to judge or some individuals are better in evaluating subjects may well offer promising improvements to existing rating schemes.

REFERENCES

ALFO, M., S. BENEDETTO, M. MALGARINI, AND S. SARLO (2017): "On the use of Bibliometric information for assessing articles quality: an analysis based on the Third Italian Research evaluation exercise," *Science, Technology & Innovation Indicators*.

ALONSO, R. (2018): "Recruiting and Selecting for Fit," *Working Paper*.

AZOULAY, P., J. S. GRAFF ZIVIN, AND G. MANSO (2011): "Incentives and creativity: evidence from the academic life sciences," *The RAND Journal of Economics*, 42(3), 527–554.

BERTOCCHI, G., A. GAMBARDELLA, T. JAPPELLI, C. A. NAPPI, AND F. PERACCHI (2015): "Bibliometric evaluation vs. informed peer review: Evidence from Italy," *Research Policy*, 44(2), 451–466.

BORNMANN, L., AND H.-D. DANIEL (2005): "Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions," *Scientometrics*, 63(2), 297–320.

BURKE, M. J., L. M. FINKELSTEIN, AND M. S. DUSIG (1999): "On average deviation indices for estimating interrater agreement," *Organizational Research Methods*, 2(1), 49–68.

CICCHETTI, D. V. (1991): "The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation," *Behavioral and Brain Sciences*, 14(1).

EUROPEAN COMISSION (2018): "Horizon 2020 Work Programme 2018-2020 3. Marie SkÅĆodowska-Curie actions," *available at* `http://ec.europa.eu/research/`

participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-msca_en.pdf.

GARFAGNINI, U., M. OTTAVIANI, AND P. N. SØRENSEN (2014): "Accept or reject? An organizational perspective," *International Journal of Industrial Organization*, 34, 66–74.

GEUNA, A., AND M. PIOLATTO (2016): "Research assessment in the UK and Italy: Costly and difficult, but probably worth it (at least for a while)," *Research Policy*, 45(1), 260–271.

GRAVES, N., A. G. BARNETT, AND P. CLARKE (2011): "Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel," *Bmj*, 343, d4797.

JAYASINGHE, U. W., H. W. MARSH, AND N. BOND (2003): "A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 166(3), 279–300.

JEFFERSON, T., M. RUDIN, S. B. FOLSE, AND F. DAVIDOFF (2006): "Editorial peer review for improving the quality of reports of biomedical studies," *Cochrane Database of Systematic Reviews*, (1).

JONES, B. F. (2011): "As science evolves, how can science policy?," *Innovation policy and the economy*, 11(1), 103–131.

KUHN, T. (1962): "The structure of scientific revolutions," .

LANDIS, J. R., AND G. G. KOCH (1977): "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374.

LI, D. (2017): "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 9(2), 60–92.

LI, D., AND L. AGHA (2015): "Big names or big ideas: Do peer-review panels select the best science proposals?," *Science*, 348(6233), 434–438.

LILJEQUIST, D., B. ELFVING, AND K. SKAVBERG ROALDSEN (2019): "Intraclass correlation–A discussion and demonstration of basic features," *PloS one*, 14(7), e0219854.

LINTON, J. D. (2016): "Improving the Peer review process: Capturing more information and enabling high-risk/high-return research," *Research Policy*, 45(9), 1936–1938.

LOHDAL, J. B., AND G. GORDON (1972): "The Structure of Scientific Fields and the Functioning of University Graduate Departments," *American Sociological Review*, 37(1), 57–72.

MUTZ, R., L. BORNMANN, AND H.-D. DANIEL (2012): "Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach," *PLoS ONE*, 7(10).

NICOLAI, A.-T., S. SCHMAL, AND C.-L. SCHUSTER (2015): "Interrater Reliability of the Peer Review Process in Management Journals," *Welpe I., Wollersheim J., Ringelhan S., Osterloh M. (eds) Incentives and Performance*.

OTTAVIANI, M. (2020): "Grantmaking," *IGIER Working Paper n. 672*.

PINA, D.-G., D. HREN, AND A. MARUSIC (2015): "Peer Review Evaluation Process of Marie Curie Actions under EU's Seventh Framework Programme for Research," *PLoS ONE*, 10(6).

SHROUT, P. E., AND J. L. FLEISS (1979): "Intraclass correlations: uses in assessing rater reliability.," *Psychological bulletin*, 86(2), 420.

SMITH, R. (2006): "Peer review: a flawed process at the heart of science and journals," *Journal of the royal society of medicine*, 99(4), 178–182.

STEPHAN, P. E. (2012): *How economics shapes science*, vol. 1. Harvard University Press Cambridge, MA.

TRAAG, V., M. MALGARINI, T. CICERO, S. SARLO, AND L. WALTMAN (2018): "Peer review uncertainty at the institutional level," .

ZUCKERMAN, H., AND R. K. MERTON (1971): "Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system," *Minerva*, 9(1), 66–100.

A Appendix

In this appendix I introduce in more detail the two key concepts to measure agreement used throughout the paper.

Intraclass Correlation (ICC)

I begin with the descriptive measure ICC that can be computed for numerical ratings, such as the score in the range of 3 to 30 in the paper. The name originates from the fact that this measure aims at quantifying a correlation within a class of two (or more) measures belonging to the same class, i.e. two ratings of the same paper. This is the basic feature distinguishing it from the ordinary Pearson correlation that can be computed for two distinct variables. Rather than establishing a (linear) relationship between two variables, the ICC describes how similar two measures from the same group are – in my case how much reviewers agree on the score of the same underlying paper.

Following the notation of Liljequist, Elfving, and Skavberg Roaldsen (2019), I start with the description of the data structure required to compute ICCs. The econometrician observes a random sample of n subjects from a general population. In the paper, the subjects are the papers undergoing peer review. On each of the n subjects a number k distinct measures of a variable of interest x is made. In my case two measurements (by the two reviewers) are made on the quality of the paper. In other contexts this could be several doctors judging the severity of a patients illness, or different tools measuring the length of an object etc.

Let me denote each such measurement as x_{ij}, i.e. the judgement of rater j on subject i. Then the data can be organized in a $n \times k$ matrix, which in my case would be a 228×2 matrix for the case of field 1 with 228 papers in the sample

$$\begin{pmatrix} \text{rating of rater 1 on paper 1} & \text{rating of rater 2 on paper 1} \\ \text{rating of rater 1 on paper 2} & \text{rating of rater 2 on paper 2} \\ \vdots & \vdots \\ \text{rating of rater 1 on paper 228} & \text{rating of rater 2 on paper 228} \end{pmatrix}$$

where it is important to note that rater 1 is not the same person along the rows of the matrix. Also the assignment which rater is labeled rater 1 and which is labeled rater 2 is arbitrary and will, in contrast to Pearson correlation coefficients, play no role in the computation of the ICC.

Having said this, let me proceed to the computation of the ICC. Due to the relatively little information contained in my sample, e.g. no identity of the reviewer, I restrict my analysis to the simple one-way ICC. This is based on a particular assumption about the underlying data generating process: each observed rating x_{ij} follows the simple law

$$x_{ij} = \mu + r_i + v_{ij}$$

where μ is the population mean, r_i is some individual component of true quality and ν_{ij} is the error of measurement j on subject i.[12] r_i and ν_{ij} are both assumed to follow mean-zero normal distributions with variances σ_r and σ_ν. Note that this implies reviewers to have no bias in evaluation. Deviations from $\nu + r_i$ are just due to idiosyncratic mean-zero errors. In richer specifications, one could allow different biases indexed by j to account for different attitudes of reviewers.

As far as agreement is concerned, the two variance terms are of interest. In particular, as the formula from above suggests the ratio of the two matters. Intuitively, whenever σ_ν is high relative to σ_r, we want to consider this as an unreliable measurement. Importantly, the size of the error variance should always be viewed in comparison to the variance of the actual variable.

For large enough n, the ICC ρ is given by

$$\rho = \frac{\sigma_r}{\sigma_{\nu+\sigma_r}}$$

which relates the variance of true quality to the overall variance including the variance of the error. It takes high values whenever σ_ν is small compared to σ_r. In our case that would translate into large differences between subjects (high σ_r), but at the same time little variation between the two raters of an identical object (low σ_ν).

The variance components can be computed from the observed matrix. Liljequist, Elfving, and Skavberg Roaldsen (2019) derive the relationship for a slightly more sophisticated model with a rater-specific bias from which I will abstract here. The rest follows their approach.

[12] This is the same equation as the one introduced in section 3. The only difference is that I follow the notation of Liljequist, Elfving, and Skavberg Roaldsen (2019) in the Appendix.

Let me introduce the following two concepts Mean Square Deviation Within Subjects (MSWS) and Mean Square Deviation Between Subjects (MSBS) computed as follows

$$MSWS = \frac{\sum_i \sum_j (x_{ij} - \bar{x_i})^2}{n(k-1)}$$

where $\bar{x_i}$ is the row average of the above matrix for each row i, i.e. the average rating of a paper. MSWS then measures how much the individual scores by two different raters deviate from the consensus score. To clarify the underlying idea, consider a paper with average rating $\bar{x_i} = 20$ that received scores $x_{i1} = 21$ and $x_{i2} = 19$. This paper contributed very little to MSWS. On the other hand a paper with the same average score $\bar{x_i} = 20$, but $x_{i1} = 30$ and $x_{i2} = 10$ leads to a sizable increase in MSWS.

On the other hand, MSBS is given by

$$MSBS = \frac{\sum_{i,j} (\bar{x_i} - \bar{x})^2}{n-1}$$

with $\bar{x}$ being the overall mean of all ratings, i.e. all entries in the matrix.

The idea of computing the ICC from the data is to link these two concepts to the parameters σ_r and σ_ν. In order to find the empirical counterparts of σ_r and σ_ν from the ANOVA, assume that one of them is equal to zero and compute the contribution of the non zero parameter on MSWS and MSBS respectively.

Begin with $\sigma_\nu = 0$, i.e. both raters make no error in measuring the quality of the paper. Then the entries in the matrix are the true qualities $r_i \forall i$ (assume $\mu = 0$ for simplicity) and

$$MSBS = \frac{\sum_{i,j} (r_i - \bar{x})^2}{n-1} = \frac{k \sum_i (r_i - \bar{x})^2}{n-1} = k \cdot \hat{\sigma_r}^2$$

where the last step exploits the fact that sample variance $\hat{\sigma_r}^2$ can be found by dividing the squared deviations of the true qualities from the sample average by (n-1). Since each component of the sum does not depend on j, we can in the second last step sum only over i and multiply by k instead.

Obviously, given that rater-specific effects are shut down ($\sigma_v = 0$), MSWS = 0 in that scenario.

Next, let $\sigma_r = 0$ and hence all subjects possess the same quality. Differences in ratings are then only due to individual errors of the raters. The entries in the matrix comprise all the individual error terms v_{ij}.

It follows that

$$MSWS = \frac{\sum_{i,j}(x_{ij} - \bar{x_i})^2}{n(k-1)} = \hat{\sigma_v}^2$$

and

$$MSBS = \frac{\sum_{i,j}(\bar{x_i} - \bar{x})^2}{n-1} = \frac{k\sum_i(\bar{x_i} - \bar{x})^2}{n-1} = k\frac{\hat{\sigma_v}^2}{k} = \hat{\sigma_v}^2$$

where the second last step uses the fact that

$$var(\bar{x_i}) = var(\sum_j x_{ij}\frac{1}{k}) = \frac{1}{k^2} \cdot k \cdot var(x_{ij}) = \frac{\hat{\sigma_v}^2}{k}$$

Summarizing the above contributions, the following connections between the variance parameters and the observed data hold

$$MSWS = \hat{\sigma_v}^2$$

and

$$MSBS = k \cdot \hat{\sigma_r}^2 + \hat{\sigma_v}^2$$

It follows that the population formula for the ICC

$$\rho = \frac{\sigma_r}{\sigma_{\nu+\sigma_r}}$$

can be estimated from the data as

$$\hat{\rho} = \frac{MSBS - MSWS}{MSBS + (k-1)MSWS}$$

which is the formula underlying the computations in section 3.

Cohen's Kappa

A related measure is Cohen's Kappa, widely used due to its ability to deal with categorical data. It is a relatively simple measure that is based on the idea that some agreement occurs due to chance and the interesting component of agreement is the one that goes beyond the level expected by chance.

Begin with its formula

$$\kappa = \frac{p_O - p_E}{1 - p_E}$$

where p_0 is the observed fraction of agreeing raters. If we think of the above matrix, it corresponds to the fraction of rows with identical entries (which can also be categorical variables).

p_E is computed based on the observed frequencies of each rating. In our case there are 5 different categories and 2 raters. Each of the five categories c occurs with a certain frequency among the first column $p_{c,1}$ and with a certain frequency among the second column $p_{c,2}$.[13]

[13]These can be interpreted as the unconditional probabilities that a certain rater attributes rating c to a subject. Naturally they have to sum to unity.

The expected agreement p_E is then the frequency of agreement if all ratings were to be allocated randomly but respecting the probabilities $p_{c,1}$ and $p_{c,2}$.

$$p_E = \sum_c (p_{c,1} \cdot p_{c,2})$$

which can be found empirically by simply counting the occurrences in the matrix from above. Denoting them by $n_{c,1}$ and $n_{c,2}$ respectively, we get that

$$p_E = \sum_c \frac{n_{c,1}}{n} \frac{n_{c,2}}{n} = \frac{1}{n^2} \sum_c n_{c,1} n_{c,2}$$

which can easily be computed from the observed matrix of ratings.