The Foerder Institute for Economic Research
at Tel Aviv University

The Eitan Berglas School of Economics

מכון למחקר כלכלי על שם
ד"ר ישעיהו פורדר
על יד אוניברסיטת תל אביב

בית-הספר לכלכלה ע"ש איתן ברגלס

Judging under Public Pressure

Florian Auferoth, Alma Cohen, Zvika Neeman

Working Paper No. 6-2021

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

Judging under Public Pressure[1]

Florian Auferoth[2] Alma Cohen[3] Zvika Neeman[4]

April 25, 2021

[1]Acknowledgments to be added.

[2]Friedrich-Alexander-Universität, Erlangen-Nürnberg

[3]Harvard Law School, Tel Aviv University, CEPR, ECGI, and NBER

[4]Tel Aviv University

Abstract

Individuals who engage in "judging" – that is, rendering a determination in a dispute or contest between two parties – might be influenced by public pressure to favor one of the parties. Many rules and arrangements seek to insulate such individuals from public pressure or to address the effects of such pressure. We study this subject empirically, investigating the circumstances in which public pressure is more and less likely to affect judging.

Using detailed data from the Bundesliga, Germany's top soccer league, our analysis of how crowd pressure affects the decisions of referees yields two key insights. First, we show that crowd pressure biases referee's decisions in favor of the home team for those decisions that cannot be unambiguously identified as erroneous but not for those decisions that can. In particular, a referee exhibits a bias in favor of the home team with respect to more subjective decisions such as the showing of yellow cards (cautions), which is based on the referee's judgment, but not with respect to more objective decisions such as validating goals and awarding penalty kicks, where live TV coverage often allows for objective identification of errors.

Second, we show that the effect of crowd pressure on referee decisions depends on the extent to which such pressure is viewed by the referee as understandable or reasonable (or even justified). Specifically, a referee's bias in favor of the home team in yellow card issuance is strengthened after the referee makes an objectively identifiable error against the home team and thus might view crowd heckling as understandable. This effect is stronger when the referee's error is costlier to the home team because the game is more important or the error is more consequential due to the closeness of the game at the time of the error.

The introduction of VAR (Video Assistant Referee) technology in 2017 and the restrictions imposed due to Covid-19 pandemic, which caused games to be played without crowds for the second half of the 2019–20 season, allow us to test our results under three different regimes (pre-VAR, VAR, and VAR/no-crowd). Inspection of the results under these three different regimes serves to reinforce them. As expected, VAR reduces the number of referee errors, but the pattern of no bias with respect to errors is preserved. VAR has no effect on the number of yellow cards, or on the number of goals.

Once the crowd disappears, so does the home advantage in goals. Referee errors are unaffected, but the home bias with respect to yellow cards disappears as well. This confirms the effect that the crowd has on referees' more subjective decisions.

JEL CODES: K40, Z20

KEYWORDS: Judging, judicial decisions, public pressure, sub judice, make-up call.

1 Introduction

In many settings, an independent decision maker is required to make one or more determinations in a dispute or contest between parties with competing interests with respect to these determinations. We refer to any such determinations as "judging." Defined as such, judging is very broad: it includes decisions by judges and jurors in court cases, decisions by arbitrators in arbitration proceedings, determinations by independent fact-finding commissions on issues that are publicly or politically contested, and decisions by referees in sports events. Indeed, in his nomination hearings, Chief Justice John Roberts famously compared the job of a Supreme Court justice to that of a baseball referee; the analogy drawn between these two types of activities reflects a view that both of them involve what we refer to as "judging."[1]

Even when the individual who engages in judging is independent of the relevant parties, her determinations may be subject to and affected by public pressure. In the case of judges, this is especially relevant for elected state court judges who expect to run for reelection,[2] as well as for other judges who have career concerns and expect the conformity of their decisions with those favored by the general public to affect their career prospects. Even judges who are appointed for life and thus presumably are free of such career concerns may be influenced by public demands in various direct and indirect ways. As for jurors, there is a long-standing recognition and concern that pretrial publicity, and the resulting public opinion about the "correct" outcome, may have a significant effect on jurors' decisions.

In recognition of the distorting effects of public pressure on judging, some rules and arrangements have been devised in order to insulate judging decisions from public pressures. In

[1] Opening Statement of Judge John Roberts before the Senate Judiciary Committee, Monday, September 12, 2005.

[2] An earlier study by two of us (Cohen et al., 2018) shows how the prospect of elections affects the decisions of federal judges. For a wider perspective on this issue, see Shepherd (2011) and the references therein.

many common law jurisdictions, there have been long-standing limitations on public commentary for cases that are *sub judice* (Latin for "under a judge," that is, under trial or otherwise under consideration by a judge or court). In the US, the First Amendment guarantee of free speech has prevented tight restrictions on comments regarding matters that are *sub judice*, but State Rules of Professional Conduct governing attorneys often place restrictions on the out-of-court attorney statements regarding ongoing cases.

Furthermore, in jury cases, judges often seek to minimize as much as possible the exposure of jurors to outside public pressure. It is common for judges to instruct jurors not to read newspaper articles about the trial. In addition, the law enables criminal defendants to move the trial venue to a different state when pretrial publicity would make it difficult to find jurors that would not have been exposed to the publicity and to the public pressures produced by it. And some criminal convictions have been overturned due to the exposure of jurors to media coverage or to the atmosphere of a "media circus." (See Phillipson, 2008, for an extensive discussion of these issues.)

In this paper, we investigate empirically whether and when public pressure affects judging. We study the subject in a setting where it is possible to obtain rich and detailed data that facilitates such an investigation. In particular, we use comprehensive and detailed data from the Bundesliga, the premier soccer league in Germany. We investigate whether and when crowd pressure influences the decisions of referees and the extent to which those decisions are biased in favor of the home team, which is favored by the (great) majority of the crowd.

What distinguishes our approach are two things. First, we consider referees' errors in addition to referees' decisions. This is important because *correct* decisions do not imply a bias, even if they are made disproportionately more often in favor or against one of the teams. Errors, on the other hand, do indicate bias if they are disproportionately made in favor or against one

of the teams. Second, we are fortunate to be able to compare our results across three different regimes:

- The first regime is our "pre-VAR regime." It covers the eight seasons prior to and including the 2016-17 season, before the introduction of VAR (Video Assistant Referee) technology.

- The second regime covers the next two and a half seasons, up to March 2020. This second regime is identical to the first, except that referees had access to VAR technology.[3] We refer to this regime as the "VAR regime."

- The third regime covers the period from March 2020 until the middle of the 2020-21 season. This third regime is identical to the second regime in that referees had access to VAR technology, but due to the Covid-19 pandemic, all games included in this regime were played behind mostly closed doors, with small or no crowds.[4] We refer to this regime as the "VAR/no-crowd regime."

As explained below, the comparison between these three different regimes allows us to draw two main insights into the way that referees respond to public pressure and to the effect of more precise refereeing technology.

Our first main insight is that crowd pressure should be expected to influence some referee decisions significantly more than others, and we find empirical evidence that is consistent with our hypothesis. In particular, we conjecture that crowd pressure has less influence on referees'

[3]VAR technology was introduced into the second Bundesliga only in the 2019/20 season, and so games played in the second Bundesliga during the 2017/18 and 2018/2019 seasons are included together with the pre-VAR regime.

[4]Starting in autumn 2020, fans were permitted into stadium for some games under certain restrictions that varied by region. For example, some stadiums imposed a partial ban on singing, and a limit of approximately 20 % was imposed on occupancy.

decisions when errors are indisputable because in such situations the countervailing force of referees' concern for their reputation is strong. By contrast, we expect crowd pressure to have more influence on referee decisions when errors are not indisputably observable, because such situations make it easier for referees to rationalize and defend their decisions without incurring large costs to their professional reputation.

Our data enables us to test this conjecture because it includes two types of decisions: (1) referees' decisions whether to validate goals and whether to award penalty kicks (which are converted to goals with a high probability), and (2) referees' decisions whether to sanction players with yellow cards (cautions). The correctness of referees' decisions with respect to goals and penalty kicks is (relatively) indisputably observable: live TV coverage provides an immediate replay of all player movements in the seconds before the actions that are the subject of the referee decisions, often from several different angles, and this generally enables an objective assessment of whether the referee made a correct or erroneous call. As a result, information on whether the referee did or did not make an error spreads quickly among all spectators of the game.

By contrast, referee decisions on whether to issue a yellow card are (relatively) more subjective. Such decisions typically involve a judgment call over which reasonable people may disagree. As a result, a referee can more easily defend a decision to issue (or refrain from issuing) a yellow card that benefits the home team without incurring a significant reputational, if any.

We provide evidence that consistent with our hypothesis that errors made in referee decisions whether to validate goals or award penalty kicks do not exhibit a bias in favor of either team. In particular, these decisions do not exhibit a bias in favor of the home team. By contrast, we find evidence that decisions to issue yellow cards are biased in favor of the home team. In

particular, we find that referees issue more yellow cards to the away team relative to the home team, and that this effect is economically and statistically significant. Interestingly, we observe this effect under both the pre-VAR and the VAR regimes, but not under the VAR/no-crowd regime. That is, this bias in favor of the home team disappears in the absence of crowds.

Our second key insight concerns the circumstances that enhance the magnitude of the effect of crowd pressure on those decisions that are amenable to such pressure (in our setting, the issuance of yellow cards). We hypothesize that when referees make an error on an indisputable decision, they are more inclined to view subsequent crowd pressure and heckling as understandable or reasonable, at least to some extent, and try, possibly subconsciously, to make up for their error in some way. The fact that, on average, 90% of the crowd in our sample is composed of fans of the home team suggests that referees are more likely to make up for their errors against the home team than for their errors against the away team.

We find that referees *do* make up for their errors against the home team by issuing more yellow cards to the away team, but not when the game is played behind closed doors. This pattern is asymmetric. Referees do not make up for their errors against the away team by giving more yellow cards to the home team (or fewer yellow cards to the away team). This is consistent with the view that these decisions are motivated, possibly subconsciously, by a desire to appease the crowd. Moreover, the increase in bias in favor of the home team is magnified when the referees' erroneous decision occurred (i) during a game that is more important, or (ii) at a point in time in the game when the score is close and so the error is likely to be more consequential. However, we find that referees *do not* make up for their errors with respect to the validation or invalidation of goals and the awarding of penalty kicks, by making additional such errors.

The introduction of VAR (Video Assistant Referee) technology in 2018 and the restrictions

imposed by the Covid-19 pandemic, which caused games to be played without crowds for the second half of the 2019–20 season, reinforce our results. As expected, the availability of VAR reduces the number of referee errors. VAR does not completely eliminate referee errors because a referee can still err if no consultation takes place between the referee and the VAR technology when it should. We find that VAR has no effect on the number of yellow cards issued. This is to be expected because the rules for engaging the VAR do not allow for yellow cards to be checked by it. The pattern of no bias with respect to referee errors is preserved under VAR.

Notably, under the VAR regime, the compensation to the home team through yellow cards to the away team after an error against the home team increases in size. Presumably, this is due to the fact that the availability of VAR technology makes a referee's errors more glaring (because not only did the referee make an incorrect call, but the incorrect call could have been corrected). Thus, with VAR, referee errors are likely both to elicit a stronger reaction from the crowd and to strengthen the referee's subconscious desire to compensate the team he has wronged. Interestingly, with VAR, referees also make up for errors against the away team by giving more yellow cards to the home team, but the number of games played under the VAR regime is not large enough to make this effect statistically significant. This is not surprising, because the fact that referees' errors become more glaring as explained above, implies that the crowd's reaction to these errors is likely to be stronger, and the referees may feel a stronger compulsion to make up for their errors, for *both* teams.[5]

Under the third regime (VAR/no-crowd), the data shows that once the crowd disappears, so does the home advantage in goals.[6] This suggests that the crowd has a strong effect on the

[5]The tendency to compensate the away team need not be manifested under the pre-VAR regime because that regime errors are less glaring and so elicit a weaker response from the crowd, and a correspondingly weaker personal motivation to compensate the away team.

[6]This has been observed in a number of recent papers, including Bryson et al. (2021), Endrich and Gesche (2020), and Scoppa (2021).

players. Referee errors are unaffected, but the referees' tendency to make up for their errors against one team by giving more yellow cards to the other team, which was manifested under the second regime with VAR, disappears as well.

Literature Review

Social influence, of which public pressure is a part, is a formidable force. Much of the research in the field of sociology is devoted to the study of its determinants and many manifestations. However, the study of social influence in the context of judging is more limited.

There is a substantial literature on the desirability of *sub judice* restrictions on public commentary on pending court cases. In particular, there is a widely shared concern that the ubiquity of media coverage makes fair trials impossible because of the difficulty of preventing jurors from being influenced by media coverage (Phillipson, 2008; Marder, 2014). Relatedly, media coverage of the work of commissions of inquiry in Canada has been called "concerning" by the head of several such committees (Gomery, 2006).

However, the empirical literature on the effect of public pressure on judging has been much more limited. Some work has investigated the subject empirically by examining how sentencing decisions by state court judges are influenced by their proximity to reelection (Huber and Gordon, 2004; Berdejó and Yuchtman, 2013; Cohen et al., 2018). There is also experimental evidence and surveys that show that pretrial publicity affects jurors (Steblay et al., 1999). We seek to contribute to this limited body of empirical work on the subject.

Due to the widely shared interest in sports and the availability of rich datasets of sport events, there exists a large literature on the decisions of sports referees. For a recent survey of this literature, see Dohmen and Sauermann (2016). This literature has uncovered several types of biased referee decisions.

There is ample evidence for bias in referee decisions about stoppage time. At the end of the game, after the regular playing time of 90 minutes has passed, the referee is allowed to make an allowance for "time lost." Garicano et al. (2005) have shown that, in the Spanish Primera Division, the referee is likely to add more time when the home team is behind, presumably to give it another opportunity to even the score. Others have found similar effects in the Bundesliga (Germany), the Premier League (England), Series A and B (Italy), Major League Soccer (USA), the Brazilian Championship Series, and the Colombian Professional League.

Boyko et al. (2007) and Page and Page (2010) find evidence for referee fixed effects on goal difference, that is the number of goals scored by the home team vs. by the away team in the Premier League that are moderated by factors that relate to social pressure.[7] Dohmen (2008) and Sutter and Kocher (2004) evaluate expert judgements and journalist reports on referee decisions to award goals and penalty kicks, and find that both goals and penalty kicks awarded to the home team are significantly less likely to be awarded correctly. Pettersson-Lidbom and Priks (2010) rely on the exclusion of spectators from games played in the Italian league in the 2006–7 season to argue that referees who are exposed to crowd noise issue significantly fewer yellow cards and fouls to the home team. Their finding has been confirmed in a laboratory study by Nevill, Balmer, and Williams (2002). More recently, Bryson et al. (2021), Endrich and Gesche (2020), and Scoppa (2021), have exploited the fact that because of the Covid-19 pandemic, all major European leagues held all their games behind closed doors to show that the home advantage in points, goals, shots, etc., as well as in referees' sanctions, is significantly reduced when games are played behind closed doors.

Various studies (see, e.g., Dawson et al., 2007, and Buraimo et al., 2010) present evidence for bias in favor of the home team in the issue of yellow and red cards. These papers, and

[7]However, Johnston (2008) failed to replicate this result on a smaller sample.

others, also show that the bias in favor of the home team is larger when the stakes are higher (such as when the reward for winning the game is increased from 2 to 3 points). Attendance, the ratio of attendance to stadium capacity, and the composition of the crowd, were also shown to contribute to the bias.

Another interesting factor that affects home bias is the physical proximity of the crowd to the field. Dohmen (2008) finds evidence for weaker bias in Bundesliga stadiums in which the crowd is separated from the field by an athletics track compared to stadiums in which it is not. His findings have been replicated by others in other leagues. Boeri and Severgnini (2011) document the effect of bribes on referee decisions. Similar findings have also been reported for other sports (Dohmen and Sauermann, 2016).

Recent literature has also examined the effect of VAR on referees' decisions. Carlos et al. (2019) find a reduction in yellow cards in the Italian Series A and the German Bundesliga after the introduction of VAR. Lago-Peñas et al. (2020) find no impact of VAR on yellow cards in the Spanish Primera Division. Dawson et al. (2020) investigating the introduction of the off-field referee in rugby, find that the addition of an off-field referee did not decrease – and perhaps even slightly contributed to – the advantage typically enjoyed by home teams in rugby. The authors' explanation for this surprising finding is that "[the on-field] referees may have been consciously or unconsciously seeking to avoid contributing to home bias before the introduction of a further official who is remote from the effects of the crowd," so that the additional remote referee my have allowed the referees in the field to perhaps lower their guard somewhat.

However, this literature has not examined the two key issues regarding the effect of public pressure on judging that are the focus of our analysis: the difference between the type of decisions that are more and less affected by crowd pressure (and in this connection the importance of the extent to which decisions can be objectively assessed by outside observers) and how the

impact of crowd pressure is moderated by the extent to which this pressure is viewed as understandable or reasonable by the decision-maker (and in this connection the power of crowd pressure to generate make-up calls).

The rest of the paper proceeds as follows. In the next section, we describe the institutional setting and the categorization of errors. The data is described in Section 3. In Section 4, we discuss our empirical methodology and present our results with respect to referee errors. In Section 5, we examine the subject of yellow cards, and show that referees' errors are stochastically independent, which indicates no bias with respect to verifiable errors. Finally, Section 6 offers a brief conclusion.

2 Institutional Background

The data we use is from the German premier soccer league, the Bundesliga. Soccer is the most popular sport in Germany. A Bundesliga game has an average number of 30,000 spectators present in the stadium.

The Bundesliga consists of two divisions. The first Bundesliga is the highest soccer division in Germany and consists of 18 teams, which face each other twice every season. Each team is the home team in one such game and the visiting or away team in the other game. This generates 34 match days per season with nine matches per match day. The second Bundesliga has the same number of teams and the same number of match days per season. At the end of each season, the two lowest-ranked teams in the first Bundesliga are demoted to the second Bundesliga, and the two highest-ranked teams of the second Bundesliga are promoted to the first Bundesliga. The team that is ranked third from last in the first Bundesliga (16th place) plays two matches against the third-ranked team of the second Bundesliga to determine the team that

will play in the first Bundesliga during the next season (a similar arrangement applies between the second and third Bundesliga).

In each game, the objective of each team is to win the game by scoring more goals than the other team. The interactions between the two teams are regulated by the laws of soccer, which specify the playing time and permitted actions. Each team is permitted to use all body parts except for arms and hands to move the ball in order to score goals. Physical tackles between players are regulated.

Each game is refereed by an official who is responsible for ensuring that both teams adhere to the rules of the game. Bundesliga referees are experienced and are selected through a system of sequential promotion tournaments. After passing a written and physical test, referees typically start in the lowest Bundesliga division. Once they have been promoted to the sixth division, they can be promoted at most one division each year if judged to be qualified by official observers. The performance of referees is monitored and judged by an official observer of the German Football Association (DFB) who attends each game, and evaluates the game's referee for being "decided, secure, with the courage to take unpopular decisions, and unimpressed by complaining players" as well as for how well the referee interprets and implements the laws of soccer. Referees who are found to be biased are dismissed (Dohmen, 2008).[8]

Two assistants support the referee. The assistants' task is to indicate whether the players were offside and whether the ball was out of bounds. The referee's task is to detect (and sanction) violations of the rules, to stop play if the rules are violated, and to ensure that play then continues according to the rules of the game. Failure of the referee to detect violations or to recognize non-violations as such can have a critical effect on the outcome of the game. The referee

[8]Referees are paid 3,800 and 2,000 Euros per match in the first and second Bundesliga, respectively, on top of an annual base salary of 35,000 Euros in the second Bundesliga, and twice as much in the first Bundesliga. They are also compensated for their travel expenses, including hotel accomodations and transportation. In 2019, the average wage in Germany was a little over 48,000 Euros.

has the final authority to decide whether the rules of the game have been violated. Specifically, the referee determines the followings: whether tackles between players are illegal (a "foul"); whether a player illegally touched the ball with his hand or arm (a "hand ball"); whether a player was in an illegal position ("offside"); whether the ball crossed the perimeter line ("out of bounds"); and whether players or officials violated the rules of the game in other ways.

Violations of the rules lead to stoppages of play. In the case of fouls or hand balls, play then restarts with a penalty kick or free kick for the team that did not commit the violation. A penalty kick is given if the violation was committed inside the penalty box and provides an excellent opportunity to score a goal by giving the team an uninterrupted shot from a distance of 11 meters to the goal, which may only be blocked by the goalkeeper (close to 80% of the penalty kicks in our sample were converted into goals). A free kick can also be a good goal-scoring opportunity, but any player may block it. The free kick's location is where the offense occurred. Following players' violations of the rules, the referee may also sanction players with a warning (yellow card) or dismissal (red card), depending on the severity of the violation. Offenses justifying a yellow card include unsportsmanlike behavior, persistent infringement of the rules, delaying the restart of the game, or dissent by word or action (FIFA, 2018). Red cards are awarded for seriously foul play, illegally denying goal-scoring opportunities, violent conduct, insulting behavior, or receiving a second yellow card (FIFA, 2018).

The referee may consult with two assistants. Video replay (VAR) was only introduced in August 2017, and was not available to the referees in the games included in our pre-VAR regime dataset. However, as mentioned above, it was generally available for the spectators watching the game, possibly with delay. At any point during the game, the referee needs to make immediate decisions with respect to whether the rules were violated and, if so, to determine the appropriate sanction. Thus, a referee faces a quick succession of situations that demand his

attention and consideration, and errors in his assessment of the situation or in his sanctions, including serious errors that have a large effect on the outcome of the game, are not uncommon. Players, coaches, and sports' fans alike all hotly debate referees' errors both during and after each game.

For the purpose of our analysis we consider referee errors that were recorded on the website www.wahretabelle.de. This website was established in 2006 with the goal of recording the correct result for all Bundesliga games. Accordingly, the website establishes what would have been the final score of the game if the referee had not made any errors. The website allows users to submit photos and video-recorded scenes from Bundesliga games for review if they believe that the referee's decision was wrong and potentially affected the result of the game.

A panel of experts assembled from the regular contributors to the website who have distinguished themselves for impartial contributions to the discussions and rulings on the website decides whether the referee's decision was correct or not. The members of the panel represent supporters of different clubs. Their number fluctuates over time; as of March 2021, it included seventeen members. Neither the panel nor the website plays any official role in the running of the German leagues. In particular, referees are not held accountable to this panel or to the website.[9] The purpose of the panel is just to accurately record events on the website, based on the panel's best judgement. On the day after each game, the panel convenes to vote on referees' errors. For each ruling by the panel, the individual votes are publicly displayed on the website and each panel member issues a brief statement explaining its decision. To protect the panel members' anonymity, no data on the panelists' identities is provided.

[9]The Bundesliga also has an internal referee evaluation system. However, the data used in this system and its deliberations are not publicly available. This system is frequently criticized for its lack of transparency, and for its overreliance on the goodwill of the head referee who is in charge of the system (see https://www.n-tv.de/sport/fussball/collinas_erben/Manipulation-Machtmissbrauch-Mobbing-article20099310.html) for an example report.)

Wrong referees' decisions are recorded as such only if they are judged by the panel to have had a direct impact on the final score of the game. If the team that was advantaged by the error did not profit from the error, the error is not recorded in this dataset. In particular, this is the case if:

- A referee incorrectly approves a goal. For example, if the referee fails to notice that an attacking player fouled a defending player just prior to scoring the goal or touched the ball with his hand.
- A referee sanctions the defending team incorrectly, and this leads to a goal. For example, by the referee incorrectly approves a penalty kick, which is then converted into a goal.
- A referee incorrectly denies a goal. For example, the referee incorrectly calls an offside offense against the attacking team.
- A referee fails to call an offense by the defending team that has a sanction that would have provided the attacking team with an excellent opportunity to score a goal. For example, if the referee fails to notice a foul inside the penalty box, which would have resulted in a penalty kick for the attacking team.

3 Data and Summary Statistics

Our dataset consists of all the matches in the first German Bundesliga from the start of the 2009-10 season, and in the second Bundesliga from the start of the 2013-14 season, until approximately the middle of the 2020-2021 season. Video assistant referee (VAR) technology was introduced into the first Bundesliga in the 2017-18 season and into the second Bundesliga in the 2019-20 season. As expected, the availability of VAR greatly reduced the number of referee

errors. The presence of VAR also changed the nature of referees' errors. Due to the Covid-19 pandemic, all the games played after March 13, 2020, were played behind closed doors, with no crowd present.

The data we use is extracted from publicly available match summaries at www.wahretabelle.de using a computer algorithm. The data is supplemented by data from forum discussions on www.wahretabelle.de in all cases where match summaries are contradictory or incomplete. All reports about possible errors are submitted for consideration through this forum. The forum also exhibits the vote and the panel's ruling on submitted possible errors, as well as all the evidence that was considered. We use majority ruling by the panel to classify errors. We use the match summaries available at www.wahretabelle.de to extract other relevant information about the game.

We supplement our dataset with match-, team-, and referee-related data extracted from the German Football Association's website and the Kicker website www.kicker.de. Kicker is Germany's largest soccer magazine. It records extensive statistics and offers live descriptions of matches. From these descriptions we extract the timing of yellow and red cards issued to players at all games in the dataset.[10] Team rankings in the league table are also extracted from Kicker.

Our data lists 76 unique referees, with about 28 unique referees per division, per season. The average age of referees in our sample is 37 with a standard deviation of 5.5.[11] For each referee in each game, we have information about the referee's experience at the time of the game. In our sample, on average, the referee of each game already refereed 67 previous games

[10]The number of red cards in our sample is too small to use as a separate dependent variable; therefore, we count red cards as two yellow cards.

[11]This data is taken from the website of the German Football Association.

with a standard deviation of 53.[12]

The data contains 4,274 games played in the pre-VAR period under the pre-VAR regime, 1,062 games played under the VAR regime, and 395 games played under the VAR/no-crowd regime. Table 1 below shows the following summary statistics for each one of the regimes. All numbers represent the average per game in the regime considered.

On average, under the pre-VAR regime 2.79 goals are scored in each game: the away team scores 1.24 goals, and the home team scores 1.55 goals. A t-test indicates that the rather large difference between these two means is statistically significant at the 1% level. The difference in the number of goals scored indicates that the home team generally possesses an advantage over the away team.

Out of the 4,274 games under the pre-VAR regime, in 1,260 games there was at least one referee error. In 609 games the first error was against the away team and in 651 games the first error was against the home team. Table 1, Panel A shows that referees make 0.37 errors per game on average, 0.19 against the Away Team, and 0.18 against the home team. This difference is small and not statistically significant. The number of errors has to be considered in relation to the average number of goals in the game, which is 2.79. Hence, when an error occurs, it has a large effect on the final score of the game.

Table 1, Panel A shows that the mean time to the first error is 47.20 minutes, with 48.29 minutes if the first error is against the away team, and 46.04 minutes if the first error is against the home team. The p-value for the difference between these values is marginally statistically significant and small in magnitude. We rely on this number to control for the difference between games with and without errors, as explained below.

[12]Referees in the second Bundesliga division are four years younger, on average, than those in the first. They are also less experienced, on average, than those in the first. Referees in the first Bundesliga division have refereed 76 Bundesliga previous games in our sample, while those in the second Bundesliga division have only refereed 53 previous games in our sample.

Table 1, Panel A also shows that referees issue an average of 3.92 yellow cards per game, with 2.15 yellow cards issued to the away team, and 1.77 to the home team. This difference is large and statistically significant at the 1% level. It is indicative of the presence of a bias in favor of the home team, as we argue below.

Table 1, Panels B and C provide the same summary statistics for the VAR and VAR/no-crowd regimes. The mean time to the first error under the VAR and VAR/no-crowd regime is similar to that in the pre-VAR regime.

Table 1, Panel B shows that in games played under the VAR regime the percentage of games with a referee error decreased to 0.22, compared to 0.37 in the pre-VAR regime. As in the pre-VAR regime, referees do not seem to err less against the home compared to the away team. The average number of errors against the away and home teams is 0.12 and 0.10, respectively. A t-test indicates that these two means are not different.

Table 1, Panel B exhibits a similar home bias in goals compared to the pre-VAR regime. On average there are 3.02 goals per game, with the away and home teams scoring 1.34 and 1.68 goals, respectively. A t-test indicates a large and statistically significant difference between these two means (-0.34 with a p-value smaller than 1%).

The table shows that referees issue more yellow cards to the away compared to the home team also under the VAR regime. Referees issue on average 3.78 yellow cards per game, with 2.06 and 1.73 yellow cards issued to the away and home teams, respectively. Again, this difference is large in magnitude and statistically significant at the 1% level.

Table 1, Panel C shows that under the VAR/no-crowd regime, the well-known persistent home bias in goals disappears. On average there are 3.03 goals per game, with the away and home teams scoring 1.45 and 1.58 goals each, respectively. A t-test confirms that this rather small difference is also statistically insignificant.

Panel C also shows that referees stop favoring the home team with respect to yellow cards. Referees issue on average 4.39 yellow cards per game, with 2.23 and 2.17 yellow cards issued to the away and home teams, respectively. As with the difference in goals, this small difference is also statistically insignificant.

Finally, stadiums in our sample have an average capacity of 35,000 spectators. Attendance in games is usually high, with an average attendance to capacity ratio of 86%. There are 64 unique referees in our sample; in each season there are about 40 unique referees. On average they each referee 17 games per season. The average age of referees is 37.

4 "Disputable" vs. "Indisputable" Errors

On average, the crowd in each game consists of more than 30,000 spectators, of which about 92% support the home team (with a standard deviation of 6.2%).[13] This implies that public pressure on the referee is likely to be very strong, especially in favor of the home team.

Nevertheless, as explained in the introduction, we hypothesize that: (1) referees' integrity is sufficiently strong to withstand crowd pressure on those decisions where an error would be "indisputable," such as decisions on the validation or invalidation of goals and penalty kicks. (2) Referees' integrity is not sufficiently strong to withstand crowd pressure on those decisions

[13]Data on the number of spectators is obtained from kicker.de and included in our dataset. The number of away fans is obtained from the website www.fussballmafia.de, which collects estimates on the number of fans directly from the away team. This number is available from the start of the 2017-18 season. Our data includes information on the number of spectators in each game up to the 26th week in the 2019-20 season, which, because of the Covid-19 pandemic, was the last game played in front of a live crowd. Our data also includes information on the number of away team fans in each game from the start of the 2017-18 season until the 26th week in the 2019-20 season obtained from the website www.fussballmafia.de, which collects estimates on the number of fans directly from the away team. We use this information to impute the number of away team fans for the games played in the seasons prior to when this information became available. We performed the imputation by regressing the number of away team fans, for those seasons for which we do have data, on the following covariates: division, round, distance (for the away team), fixed effects for each of the two teams, teams' rankings, various measures of the record of the two teams' previously played games, and the day of the week. The correlation obtained between the actual number of away fans and the predicted number of away fans is about 0.9.

where an error would not be "indisputable," such as decisions on the issuance or nonissuance of a yellow card.

In order to test this hypothesis, we run the following regression:

$$\begin{aligned} Y_{g,i} = & \\ & \beta_1 Home\ Team_{g,i} + \beta_2 Home\ Team_{g,i} \times Mild_g \\ & + \beta_3 Home\ Team_{g,i} \times High_g + \alpha_g + \epsilon_{g,r,i}, \end{aligned}$$

where g represents indexed games. For each game we have two rows, one for each team. The index $i \in \{1, 2\}$ represents the first or second playing team. The dependent variable $Y_{g,i}$ is the number of errors against/yellow cards given to team i in game g. Our main variable of interest is *Home Team*, which is a dummy variable that is equal to 1 if the game is a home game for the team, and 0 otherwise. This variable is supposed to capture the bias, if any, created by crowd pressure. The variables *Mild* and *High* are dummies that are equal to one if the game is "mildly" or "highly" important, and zero otherwise. Remaining games are considered to be of "low" importance.

We define a game to be important for a team if the team ranks among those in the top or bottom third in the Bundesliga table at the time of the game, and the absolute difference in score between the team and those adjacent to it in the Bundesliga table (from above and below) is less than or equal to 2. This definition of importance is motivated by the idea that teams in the top or bottom third of the Bundesliga table have a stronger incentive to do well. Those in the top want to secure their top position, which allows them to compete in European leagues and promises other rewards, and those in the bottom third want to avoid being demoted to a lower division at the end of the season. We define a game to be "of low importance" if it is not important to

either the home or away team according to our definition; a game is "mildly" important if it is important to the away but not to the home team; and a game is "highly" important if it is important to the home team (regardless of its importance to the away team). The reason we assign a bigger weight to the home team is that most of the crowd in the stadium supports the home team (as mentioned above, on average, only about 8–10% of the crowd support the away team), so that if a game is important to the home team, it is also important to a much larger fraction of the crowd. Hence, our measure of importance captures whether a game is expected to be important, ex-ante, before the game is played.

The regression, as well as all other regressions below, controls for unique game dummies.[14] Standard errors are clustered by the referee and unique game.[15] Table 2 below presents a basic analysis of home bias.

Table 2 shows the results of running the regression on errors and yellow cards while controlling for the importance of the game and its interaction with the variable *Home Team*. Columns (1), (3), and (5), exhibit the results when the dependent variable is the number of errors under the three regimes, respectively. The estimated measure of crowd pressure (measured through the coefficients of *Home Team* and its two interactions) is not statistically significant under all three regimes. The results indicate that there are more errors in mildly and highly important games, but these effect are borderline statistically significant. This implies that, even in games that are more important to the home team, where we expect crowd pressure from the home team fans to be stronger, there is no extra bias in favor of the home team. We also find that the ratio of attendance to capacity, which measures how packed the stadium is, which intuitively is positively correlated with the pressure that the crowd exerts on the referee, and the

[14]Running the regressions controlling for division, season, week, teams, team rankings and referee dummies instead of for unique game dummies does not change our results.

[15]Running the regressions using different clusterings does not change our results.

size of the stadium, have no effect of the number of errors.

Table 2, Columns (2), (4), and (6) exhibit a similar analysis for the number of yellow cards. The results show a large and statistically significant home bias at the 1% level under both the pre-VAR and VAR regimes. On average the *Home Team* receives approximately 20% fewer yellow cards in a game. This effect disappears under the VAR/no-crowd regime. Under all three regimes, the importance of the game has a positive effect on the number of yellow cards issued, which is to be expected, but this effect is not statistically significant.

The results in Table 2 are consistent with our hypothesis that errors made in referee decisions whether to validate goals or award penalty kicks do not exhibit a bias in favor of the home team. However, we find evidence that decisions whether to issue yellow cards are biased in favor of the home team when the game is played in front of a live crowd. In particular, we find that referees issue more yellow cards to the away team, relative to the home team, and that this effect is both economically and statistically significant. The bias in favor of the home team in the issue of yellow cards disappears when the game is played behind closed doors, with no crowd.

5 The Effect of an Indisputable Error

Our second hypothesis concerns referees' responses to their previous indisputable errors. An indisputable error against the home team is very likely to elicit a strong response from the crowd, and so, as explained in the introduction, we hypothesize that a referee who errs against the home team in an indisputable way tries, possibly subconsciously, to make up for his error in some way. On average, less than 10% of the crowd supports the away team. This implies that an indisputable error against the away team is less likely to elicit a strong reaction from the crowd, and so is also less likely to induce the referee to try to make up for it.

Indeed, we show that referees make up for their errors against the home team by giving more yellow cards to the away team, and not vice versa, but not when the game is played behind closed doors, which is consistent with our hypothesis. Importantly, we also find that referees *do not* make up for their errors with respect to the validation or invalidation of goals and the awarding of penalty kicks, by making additional such errors.

5.1 Yellow Cards

In this section, we show that referees exhibit a home bias in the following sense: they issue more yellow cards to the away team after an error against the home team, but not vice versa, whenever the game is played in front of a crowd.

To test the idea that referees issue more yellow cards to the away team after an error against the home team, but not vice versa, we compare the number of yellow cards given to the two teams before and after the first error in games with at least one error with the number of yellow cards given to the two teams before and after halftime in games with no errors.

To make this comparison, from each game in our sample we derive four observations: two observations for the home team and two observations for the away team. The first observation for each team is of the number of yellow cards *before* the first error, or halftime in games with no errors; and the second observation is of the number of yellow cards *after* the first error, or halftime in games with no errors.

However, if the first error occurred, say, 30 minutes after the start of the game, then it is likely that the number of yellow cards given in such a game before the error occurred, is smaller than the number of yellow cards given in the first half (45 minutes) of a game without errors.[16] Likewise, the number of yellow cards given in such a game after the error is likely to be larger

[16] A game is played for 90 minutes. Halftime begins in the 45th minute.

than the number of yellow cards given in the second half of a game without errors, because in the former game, there are 60 minutes in which yellow cards can be given, whereas in the latter game, there are only 45 minutes.

Therefore, in order for the comparison between the number of yellow cards in games with and without errors to be meaningful, we multiply the number of yellow cards given before and after an error by the ratio of the number of minutes to the first error and 45. For example, if the first error occurred in the 30th minute, then we split the game into two parts: the first part includes all yellow cards up to the 30th minute, and the second part includes all yellow cards in the remaining 60 minutes after the error. In order to account for the different length of time in which a yellow card can be given in these two parts of the game, we adjust the dependent variable by multiplying the number of yellow cards in the first and second parts of the game by $\frac{45}{30} = \frac{3}{2}$ and by $\frac{45}{60} = \frac{3}{4}$, respectively. More generally, if the first error occurred in the τ-th minute, then we multiply the number of yellow cards given before and after the first error by $\frac{45}{\tau}$ and by $\frac{45}{90-\tau}$, respectively. This transformation accounts for the fact that the time before and after the first error occurred can be longer or shorter than halftime, or 45 minutes.[17] We also winsorize the number of yellow cards before and after the first error to avoid extreme values in case the first error occurred in the first or last few minutes of the game.

In order for the comparison between the number of yellow cards given before and after the first error to be valid, we need to verify that both the occurrence and, importantly, the *direction* of the first error is not itself biased in favor of the home (or away) team. We therefore run similar regressions to those reported in Table 2, except that the dependent variable is *First error*, which is a dummy variable equal to one if the first error was against team i in game g, and zero otherwise.

[17] Recall that the mean time to a first error in games in which one occurs is 47.49 minutes.

The results of these regressions are presented in Table 3. Columns (1), (3), and (5) exhibit the results for all games. Columns (2), (4), and (6) repeat the analysis on the subset of games with errors. The results show that regardless of the regime and of whether we run the analysis on all games or just on the subset of games with referee errors, the estimated measure of a home bias (the coefficient of *Home Team*) is not statistically significant. The results also show that the interaction of the importance of the game with *Home Team* is not statistically significant. This implies that there is no extra home bias that depends on the importance of the game. These results are consistent with our hypothesis that pressure from the crowd does not bias the direction of the referee's first error in either way, and justifies treating the first error as an unbiased variable.

As explained above, we hypothesize that in games with errors, the number of yellow cards given to a team following a first error against the other team is larger than if no error was made. As explained above, we find that, in games that are played in front of a crowd, this is indeed the case when the first error was committed against the home team, but not in games where the first error was committed against the away team.

To test this hypothesis we run the following regression:

$$
\begin{aligned}
Yellow\ Cards_{g,i,t} = & \\
& \beta_1 Home\ Team_{g,i} + \beta_2 After_{g,t} + \beta_3 Home\ Team_{g,i} \times After_{g,t} \\
& + \beta_4 Home\ Error_g \times Home\ Team_{g,i} \\
& + \beta_5 Away\ Error_g \times HomeTeam_{g,i} + \beta_6 Home\ Error_g \times After_{g,t} \\
& + \beta_7 Away\ Error_g \times After_{g,t} + \beta_8 Home\ Error_g \times Home\ Team_{g,i} \times After_{g,t} \\
& + \beta_9 Away\ Error_g \times Home\ Team_{g,i} \times After_{g,t} + \alpha_g + \epsilon_{g,i,t}.
\end{aligned}
$$

The dependent variable $Yellow\ Cards_{g,i,t}$ is the number of yellow cards given to team $i \in (1,2)$, that played in game g, in part $t \in \{first, second\}$, where $t = second$ if it is the second half of the game in games with no errors or if it is after the first error in games with errors, and $t = first$ otherwise (if it is the first half of the game in games with no errors or if it is before the first error in games with errors). The dependent variable $Home\ Team_{g,i}$ is a dummy variable that is equal to 1 if team i is the home team. $After_{g,t}$ is a dummy variable that is equal to 1 if $t = second$. Namely, it is equal to 1 if the yellow card was issued after the first error or in the second half of the game in games with no errors. $Home\ Error_{g,i}$ and $Away\ Error_{g,i}$ are dummy variables that are equal to 1 if the first error in the game was against the home and away teams, respectively.

The coefficient of the variable *Home Team* describes whether the home team receives more/less yellow cards. The coefficient of the variable $After$ describes whether more/less yellow cards are given to the team after a referee error, or in the second half in games with no errors. The coefficient of the variable $Home\ Team \times After$ describes whether the home team receives even more/less yellow cards after a referee error, or in the second half in games with no errors. The coefficient of the variable $Home\ Error \times Home\ Team$ measures whether in games with a first error against the home team, more/less yellow cards are given to the home team, and the coefficient of the variable $Away\ Error \times HomeTeam$ measures whether in games with a first error against the away team, more/less yellow cards are given to the home team. The coefficient of the variable $Home\ Error \times After$ measures whether in games with a first error against the home team, more/less yellow cards are given, and the coefficient of the variable $Away\ Error \times After$ measures whether in games with a first error against the away team, more/less yellow cards are given. Finally, the 3-way interaction terms $Home\ Error \times Home\ Team \times After$ and $Away\ Error \times Home\ Team \times After$ are our main variables of interest. The coefficient of the former measures

whether in games with a first error against the home team, more/less yellow cards are given to the home team after the error, or in the second half of the game in games with no errors, and the latter coefficient measures whether in games with a first error against the away team, more/less yellow cards are given to the home team after the error, or in the second half of the game in games with no errors. Finally, the coefficients $alpha_g$ are unique game dummy variables that measure the effect of all the variables that are fixed for the game, such as location, division, season, week, teams, team rankings, referee, etc. The variable $\varepsilon_{g,i,t}$ describes random noise.

Table 4 below shows the results. Columns (1), (3), and (5), include all games, and Columns (2), (4), and (6), exclude games with a first error in the last five minutes of the game. In such games, the referee's chance of finding himself in a situation where he can compensate a team for an error he made against it is quite limited.

In the pre-VAR regime, the results reported in Columns (1) and (2) show that, on average, home teams receive less yellow cards. This effect is large and statistically significant. We also find that the number of yellow cards in the second part of the game is larger and statistically significant, which suggests that games become more intense as they progress over time. That is, players play more aggressively and are cautioned more often for aggressive play. However, the fact that the $Home\ Team_{g,i} \times After_{g,t}$ interaction variable is small and insignificant implies that there is no additional home bias in the second half of the game. In games where the first error was against the home team, the home team receives more yellow cards than in games with no errors. However, there is no statistically significant difference in the number of yellow cards given to the home team in games where the first error was against the away team. The coefficients of the two terms $HomeError_{g,i} \times After_{g,t}$ and $AwayError_{g,i} \times After_{g,t}$ are both positive and statistically significant. These variables capture the intensity of the game in the second half,

in games with referees' errors. Such games are probably more intense, which implies both more errors and more yellow cards.

Notably, the coefficient of the 3-way interaction term $HomeError_{g,i} \times HomeTeam_{g,i} \times After_{g,t}$ is negative and statistically significant. This suggests that referees favor the home team by giving it fewer yellow cards after a first error against the home team. The fact that the coefficient of the term $AwayError_{g,i} \times HomeTeam_{g,i} \times After_{g,t}$ is very small and not statistically different from zero implies that referees do not exhibit a symmetric attitude towards the away team after a first error against it. Column (2) shows that the results described in Column (1) are not affected by the last minutes of the game.

Columns (3) and (4) show the results for the VAR regime. As with the pre-VAR regime, on average, away teams receive more yellow cards, and the number of yellow cards in the second part of the game is larger (both coefficients are statistically significant at the 1% level). The table also shows that the $HomeTeam_{g,i} \times After_{g,t}$ interaction variable is small and insignificant, which implies that the difference in the number of yellow cards between home and away teams is not larger in the second half of the game. The results for all the interaction terms between *HomeTeam* and *Error* are small and statistically insignificant. The coefficient of the $HomeError \times After$ term is statistically significant and larger in magnitude compared to the pre-VAR regime. This suggests that in games with VAR, the crowd responds more angrily to referees' errors against the home team, which contributes to the intensity of the game and generates more yellow cards to both teams. This effect is not found after a first error against the away team. The regression results that we get for the 3-way interactions in Columns (3) and (4) are somewhat different than those in Columns (1) and (2). The 3-way interaction $HomeError_{g,i} \times HomeTeam_{g,i} \times After_{g,t}$, when all games are included (Column (3)), is negative, with similar magnitude as in the pre-VAR regime, however, it is not statistically significant. But

when games in which the first error occurred in the last five minutes of the game are excluded (Column (4); there are only 19 such games), then the 3-way interaction coefficient increases in magnitude and becomes statistically significant at the 5% level (-0.309 under the pre-VAR regime and -0.505 under the VAR regime). In comparison to the pre-VAR regime, the coefficient of 3-way interaction $Away\ Error_{g,i} \times Home\ Team_{g,i} \times After_{g,t}$ increases substantially in magnitude (from -0.008 and 0.016 under the pre-VAR regime to 0.211 and 0.304 under the VAR regime, respectively). While this coefficient is not statistically significant when all games are included (Column (3)), it is statistically significant at a level of 10% when games in which the first error occurred in the last five minutes of the game are excluded (Column (4)).

Columns (5) and (6) show the results for the VAR/no-crowd regime. In contrast to the other two regimes we do not find that away teams receive more yellow cards than home teams. The magnitude of the coefficient of *Home Team* is substantially smaller (-0.031 compared to -0.162 and -0.166 under the pre-VAR and VAR regimes, respectively) and is not statistically significant. All the 2-way interaction terms are not statistically different from zero. We also find that the 3-way interactions are small in magnitude and are not statistically significant. However, this finding is tempered by the fact that the number of games included in the VAR/no-crowd regime is much smaller compared to the other two regimes (about 4,000, 1000, and 400 games were played under the pre-VAR, VAR, and VAR/no-crowd regimes, respectively).

Inspection of the results presented in Table 4 is consistent with our hypotheses as described in the introduction:

1. Referees make up for their errors against the home team by giving more yellow cards to the away team, but not when the game is played behind closed doors. Referees do not make up for their errors against the away team by giving more yellow cards to the home team (or less yellow cards to the away team).

2. Under the VAR regime, the compensation to the home team through yellow cards to the away team after an error against the home team increases in size. Notably, with VAR, referees also make up for errors against the away team by giving more yellow cards to the home team. As explained above, this is not surprising because under VAR referees' errors become more glaring, which implies that the crowd's reaction to these errors is likely to be stronger.

3. Under VAR/no-crowd regime the referees' tendency to make up for their errors against one team by giving more yellow cards to the other team disappears.

An alternative explanation that could account for the finding that referees issue more yellow cards to the away team after an error against the home team may be that when the home team is behind, it changes its game strategy by playing more offensively (Bartling et al., 2005, provide evidence that supports this view).[18] Presumably, this implies that the away team is pushed to play more defensively, which in turn implies that it also receives more yellow cards. This could indeed be an alternative explanation for our finding that the referee issues more yellow cards to the away team after an error against the home team (but not vice versa). However, the results described in Columns (5) and (6) show that this effect disappears when the game is played behind closed doors. Thus, this alternative explanation requires that home teams change their strategies when behind *only* when the game is played in front of a crowd, but not when the game is played behind closed doors. While such an explanation is not inconceivable, we find it to be less plausible than our explanation that the increased number of yellow cards is due to increased pressure from the crowd following an error of the referee against the home team.

In Table 5, we exhibit the results of a regression that is similar to the one described in Table 4, except that we run the regression separately for games that are more or less "important" to

[18]This explanation was suggested to us by a referee.

the teams (Columns (1)–(3)) and "close" or "not-close" (Columns (4)–(5)). A game is defined as "close" if the score of the game was even, at the moment of the first error in games with at least one error, or at halftime in games with no errors. While "importance" measures whether a game is considered important, ex-ante, before the start of the game, "closeness" is an endogenous measure of importance, because it depends on how the game develops. We perform this analysis only for the pre-VAR regime, because splitting the VAR and VAR/no-crowd regimes into these subgroups reduces the number of observations per subgroup to an extent that does not allow us to obtain accurate results. Yet, despite the fact that the results are noisier, they are qualitatively similar to those observed under the pre-VAR regime.[19]

Inspection of the sign and statistical significance of the 3-way interaction term, which is our main variable of interest, indicates that referees tend to deferentially compensate the home team only in games that are mildly and highly important for the home team. Specifically, Column (1) of Table 5 shows that in games that are of low importance for the home team the coefficients of our main variables of interest are small and insignificant. However, in Columns (2) and (3) of Table 5, which report the results of more important games, these coefficients are larger in magnitude and also statistically significant at the 5% and 1% levels.

Column (5) of Table 5 includes games in which the score of the game was tied at the moment of the first error or halftime. Column (4) includes all the remaining games. Inspection of the results presented in Table 5 reveals that our main variable of interest, namely, the coefficient of the 3-way interaction term, is negative and statistically significant only in "close" games. We hypothesise that this is due to the fact that the crowd exerts a stronger pressure on referees in close games. The effect is of greater magnitude than in Table 4, and is statistically significant at the 1% level.

[19] We have not included the results of these regressions in the paper.

The results reported in Table 5 are consistent with our hypothesis that the increase in bias in favor of the home team is magnified when the referee's erroneous decision occurred (i) during a game that is more important, or (ii) at the point of time in the game in which the score was close and so the error was likely to be more consequential.

5.2 Referee Errors

In this subsection, we show that referees do not make up for their errors against one team by making an error against the other team. To test this, we examine whether a referee who made an error against the home (away) team is more likely to make a second error against the away (home) than against the home (away) team.

The distribution of the order of referee errors is depicted in the tree-diagram below. The numbers are listed in the order of their relevant regime, so that, for example, the entry "4274, 1062, 395 Games" means 4274, 1062, and 395 games played under the pre-VAR, VAR, and VAR/no-crowd regime, respectively.

Out of the 4274 games played under the pre-VAR regime, in 3014 games (71%) there were no errors, in 609 games (14%) the first error was against the home team, and in 651 games (15%) the first error was against the away team.

Out of the 609 games played under the pre-VAR regime in which the first error was against the home team, in 485 games (80%) there were no additional errors, in 57 games (9%) the second error was against the home team, and in 67 games (11%) the second error was against the away team. Out of the 651 games played under the pre-VAR regime in which the first error was against the home team, in 527 games (81%) there were no additional errors, in 60 games (9%) the second error was against the Home team, and in 64 games (10%) the second error was against the away team. The fact that the proportion of games with no additional errors beyond

the first one in these two (sub-)conditional distributions, 80% and 81%, respectively, is larger than the proportion of games with no errors at all, 71%, is due to the fact that less time remains for a second error to occur compared to a first error.[20]

As can be seen from the tree-diagram above, among games played under the pre-VAR regime, errors against the home and away teams are distributed quite symmetrically. This pattern holds true both for the first and second errors. This suggests that crowd pressure has little, if any, influence on referee's decisions that are indisputably observable. In particular, following a first error against the home team, out of the 124 games with at least two errors, in 54% of the games the second error was against the away team, and following a first error against the away team, out of the the 124 games with at least two errors, in 48% of the games the second error was against the home team. A simple t-test fails to reject the null hypothesis that these numbers are equal to one-half, which implies that there is no home bias.

The number of games played under the VAR and VAR/no-crowd regimes is too small for a similar comparison to be meaningful because a difference of even one error has a large effect on the conditional distribution of the second error following a first error against one of the teams.

However, because a simple t-test cannot control for additional variables, we run the following regression:

$$2^{nd}Error_{g,i} =$$

$$\beta_1 Home\ Team_{g,i} + \beta_2 Home\ Team_{g,i} \times 1^{st}Error\ Against\ Home_g + \alpha_g + \epsilon_{g,r,i}$$

[20] Among the games played under the pre-VAR regime, there are only 45 (24 and 7 in the VAR and VAR/no-crowd regimes respectively) games with three or more errors. There are four different conditional distributions of the third error (after two errors against the home team, after an error against the home team that is followed by an error against the away team, etc.). The number of games in each one of the relevant categories is small and therefore we do not present these games here. However, these games are included in the regression analysis that is performed below.

The dependent variable, 2^{nd} *Error*, is a dummy variable equal to one if there was a second error, and zero otherwise. *1st Error Against Home* is a dummy variable equal to one if the first error was against the home team, and zero otherwise.

We expect stronger pressure from the crowd on the referee after an error against the home team compared to after an error against the away team because, on average, most of the crowd (90%) consists of fans of the home team. We expect this pressure to be even stronger in games that are deemed "important," and/or "close," as defined above. As explained above, we expect the referee to exhibit a stronger bias in favor of the home team when he is subject to stronger crowd pressure. The fact that we do not find any indication of such a bias even in close and important games as shown in the regression tables below indicates that no such bias exists.

Table 6, Column (1) shows the results for the pre-VAR regime when all games are included. The results show that our coefficient of interest, namely, the coefficient of the interaction term, is small in magnitude and insignificant. Columns (2)–(4) reveal similar results when only "low", "mild," and "high" importance games are included. In Columns (5) and (6) we rerun the regression separately for games that we define as "close" and "not close." Columns (5) and (6) show that regardless of whether the games are close or not, our coefficient of interest is small in magnitude and insignificant.

The coefficients of our main independent variable of interest, which is the interaction of *Home Team* and *1st Error Against Home*, is small and insignificant in all the above regressions, suggesting that referees do not compensate for their errors by making additional errors against the other team. Our results are consistent with our hypothesis that crowd pressure has less influence on referee decisions that are indisputable.

Table 7 reproduces the results reported in Table 6 for the other two regimes. Column (1) shows the results for the VAR regime and Column (2) shows the results for the VAR/no-

crowd regime, when all games are included. The results show that our main coefficient of interest, namely, the coefficient of the interaction term, is small in magnitude and statistically insignificant. We do not run the regressions for the VAR and VAR/no-crowd regimes separately by the importance and closeness of the game, because splitting the set of games reduces the number of observations per subgroup to an extent that does not allow us to obtain accurate results.

6 Conclusion

There is a long-standing concern about the effects of public pressure on judging. Our objective in this paper has been to contribute to the empirical investigation of this subject. In particular, we have focused on investigating the circumstances under which public pressure is more and less likely to affect judging. To this end, we have used detailed and rich data from Germany's top soccer league.

A key insight of our analysis is that the extent to which public pressure affects judging depends on the extent to which judging decisions seeking to placate public pressure can be indisputably identified as erroneous by outside observers and thereby impose a reputational cost on the decision maker. Another key insight is that with respect to those decisions where public pressure affects judging, the strength of the effect depends on the extent to which such pressure is viewed by the decision maker as understandable or reasonable. We hope that future empirical work will further study the circumstances in which public pressure is more and less likely to affect judging.

7 Tables

Table 1: Summary Statistics

Panel A: Pre-VAR regime

	Average	Away Team	Home Team	diff	P-value
Number of Goals	2.79	1.24	1.55	-0.31	0.00
	(1.70)	(1.17)	(1.3)		
Number of Referee Errors	0.37	0.19	0.18	0.01	0.12
	(0.64)	(0.45)	(0.43)		
Minutes to First Error	47.20	48.29	46.04	2.24	0.11
	(25.16)	(25.7)	(24.53)		
Number of Games with $\geq$ 1 Errors	1260	722	675		
Number of Yellow Cards	3.92	2.15	1.77	0.38	0.00
	(2.12)	(1.41)	(1.33)		
Number of Games	4274				

Panel B: VAR regime

	Average	Away Team	Home Team	diff	P-value
Number of Goals	3.02	1.34	1.68	-0.34	0.00
	(1.70)	(1.24)	(1.36)		
Number of Referee Errors	0.22	0.12	0.10	0.02	0.21
	(0.47)	(0.34)	(0.32)		
Minutes to First Error	47.80	48.10	47.44	0.66	0.86
	(26.51)	(25.94)	(27.32)		
Number of Games with $\geq$ 1 Errors	207	114	93		
Number of Yellow Cards	3.78	2.06	1.73	0.33	0.00
	(2.02)	(1.31)	(1.31)		
Number of Games	1,062				

Panel C: VAR/no-crowd regime

	Average	Away Team	Home Team	diff	P-value
Number of Goals	3.03	1.45	1.58	-0.13	0.18
	(1.69)	(1.27)	(1.32)		
Number of Referee Errors	0.18	0.08	0.10	-0.01	0.56
	(0.43)	(0.29)	(0.31)		
Minutes to First Error	49.31	49.0	49.69	-0.69	0.92
	(26.43)	(26.60)	(26.70)		
Number of Games with $\geq$ 1 Errors	64	35	29		
Number of Yellow Cards	4.40	2.23	2.17	0.05	0.57
	(2.11)	(1.32)	(1.49)		
Number of Games	395				

Note: Standard deviations are in parentheses.

Table 2: Number of Errors/Yellow Cards by Importance

	Pre-VAR		VAR		VAR/no-crowd	
	Errors	Yellow Cards	Errors	Yellow Cards	Errors	Yellow Cards
	(1)	(2)	(3)	(4)	(5)	(6)
Home Team	0.0013	-0.3953***	0.0064	-0.3718**	0.0220	-0.0659
	(0.0235)	(0.0680)	(0.0310)	(0.1543)	(0.0515)	(0.1847)
Home Team x Mild	-0.0201	0.0136	-0.0252	0.0936	0.0225	-0.0563
	(0.0327)	(0.0903)	(0.0321)	(0.1756)	(0.0782)	(0.2320)
Home Team x High	-0.0231	0.0123	-0.0220	-0.0071	-0.0267	0.0519
	(0.0274)	(0.0771)	(0.0327)	(0.1720)	(0.0623)	(0.2274)
N	8,324	8,324	2,000	2,000	790	790
R^2	0.06	0.21	0.02	0.22	-0.01	0.12
Mean	0.18	1.96	0.11	1.89	0.09	2.19

Note: Standard errors are clustered by the referee and unique game and are in parentheses.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

Table 3: First Error

	Pre-VAR		VAR		VAR/no-crowd	
Games	All	With Errors	All	With Errors	All	With Errors
	(1)	(2)	(3)	(4)	(5)	(6)
Home Team	0.0066	0.0239	0.0000	0.0000	0.0110	0.0588
	(0.0214)	(0.0781)	(0.0355)	(0.1993)	(0.0457)	(0.2506)
Home Team x Mild	-0.0263	-0.0868	-0.0188	-0.0769	0.0335	0.2269
	(0.0323)	(0.1099)	(0.0346)	(0.1850)	(0.0743)	(0.4492)
Home Team x High	-0.0179	-0.0623	-0.0156	-0.0874	-0.0063	-0.0285
	(0.0240)	(0.0861)	(0.0369)	(0.2084)	(0.0570)	(0.3314)
N	8,324	2,460	2,000	392	790	128
R^2	0.414	0.002	0.446	0.006	0.458	0.019
Mean	0.148	0.500	0.098	0.500	0.081	0.500

Note: Standard errors are clustered by the referee and unique game and are in parentheses.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

Table 4: Yellow Cards Home/Away Team, Before/After

	pre-VAR		VAR		VAR / no-crowd	
	All Games	Last 5 min. excl.	All Games	Last 5 min. excl.	All Games	Last 5 min. excl.
	(1)	(2)	(3)	(4)	(5)	(6)
Home Team	-0.166***	-0.166***	-0.162***	-0.162***	-0.031	-0.031
	(0.018)	(0.018)	(0.031)	(0.031)	(0.061)	(0.061)
After	0.640***	0.640***	0.608***	0.608***	0.656***	0.656***
	(0.036)	(0.036)	(0.043)	(0.043)	(0.096)	(0.096)
Home Team x After	-0.036	-0.036	-0.012	-0.012	0.027	0.027
	(0.040)	(0.040)	(0.064)	(0.064)	(0.115)	(0.115)
Home Error x Home Team	0.149***	0.155***	0.046	0.055	-0.079	-0.056
	(0.050)	(0.053)	(0.131)	(0.137)	(0.233)	(0.258)
Away Error x Home Team	-0.034	-0.039	-0.008	-0.044	-0.011	-0.034
	(0.054)	(0.057)	(0.101)	(0.101)	(0.130)	(0.130)
Home Error x After	0.313***	0.341***	0.552***	0.615***	-0.015	0.086
	(0.068)	(0.063)	(0.153)	(0.158)	(0.295)	(0.305)
Away Error x After	0.224***	0.217***	0.030	0.047	0.308	0.297
	(0.066)	(0.063)	(0.175)	(0.160)	(0.350)	(0.350)
Home Error x Home Team x After	-0.304***	-0.309***	-0.342	-0.505**	0.223	0.043
	(0.077)	(0.082)	(0.245)	(0.235)	(0.465)	(0.467)
Away Error x Home Team x After	-0.008	0.016	0.211	0.304*	-0.071	-0.049
	(0.093)	(0.096)	(0.187)	(0.179)	(0.286)	(0.283)
N	16,530	16,188	3,980	3,904	1,404	1,392
R^2	0.3706	0.3718	0.3817	0.3831	0.3474	0.3506
Mean	0.9716	0.9676	0.9291	0.9284	1.0807	1.0817

Note: Standard errors are clustered by the referee and unique game and are in parenthesis.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

Table 5: Yellow Cards Before/After an Error by Importance/Closeness

	Importance			Closeness	
	Low	Mild	High	No	Yes
	(1)	(2)	(3)	(4)	(5)
Home Team	-0.130***	-0.166***	-0.178***	-0.189***	-0.160***
	(0.035)	(0.043)	(0.026)	(0.051)	(0.020)
After	0.582***	0.630***	0.664***	0.408***	0.687***
	(0.074)	(0.066)	(0.038)	(0.062)	(0.038)
Home x After	-0.128^t	0.020	-0.030	0.010	-0.053
	(0.084)	(0.083)	(0.049)	(0.091)	(0.045)
Home Mistake x Home Team	0.167*	0.267**	0.102^t	-0.089	0.201***
	(0.097)	(0.133)	(0.066)	(0.127)	(0.063)
Away Mistake x Away Team	-0.228*	0.041	-0.018	-0.093	-0.020
	(0.135)	(0.102)	(0.076)	(0.117)	(0.056)
Home Mistake x After	0.225*	0.481***	0.316***	0.090	0.358***
	(0.113)	(0.162)	(0.090)	(0.231)	(0.058)
Away Mistake x After	0.319^t	0.356**	0.122	0.346*	0.201***
	(0.204)	(0.136)	(0.086)	(0.183)	(0.070)
Home Mistake x Home Team x After	-0.032	-0.488**	-0.320***	0.118	-0.357***
	(0.193)	(0.240)	(0.117)	(0.264)	(0.094)
Away Mistake x Home Team x After	0.295	-0.118	-0.010	-0.088	0.033
	(0.262)	(0.201)	(0.100)	(0.220)	(0.095)
N	2,984	4,084	9,120	2,644	14,020
R^2	0.3789	0.3727	0.3711	0.3323	0.3807
Mean	0.9288	0.9723	0.9782	0.9438	0.9759

Note: Standard errors are clustered by the referee and unique game and are in parentheses.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

Distribution of Errors under the pre-VAR, VAR, and VAR/No-Crowd Regimes

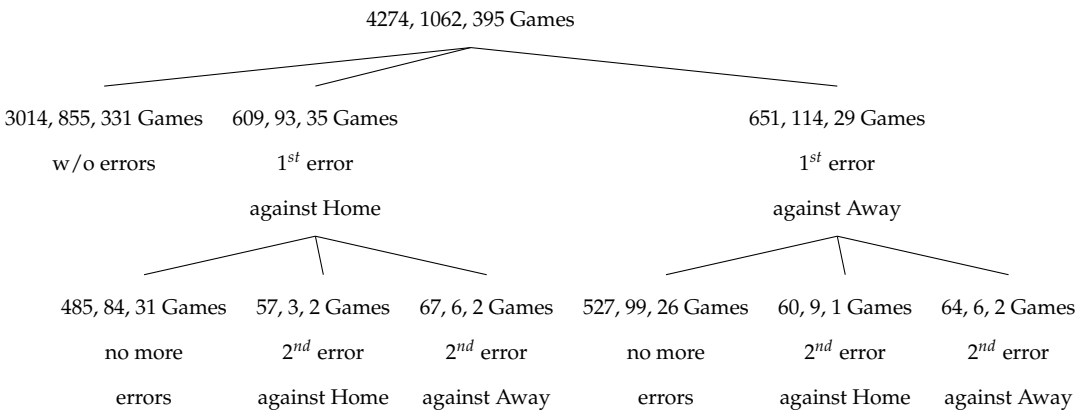

Table 6: Second Error by Importance/Closeness under the Pre-VAR Regime

		Importance			Closeness	
	(1)	(2)	(3)	(4)	(5)	(6)
	All	Low	Mild	High	No	Yes
Home Team	-0.0094	-0.0294	0.0296	-0.0219	-0.0300	-0.0056
	(0.0160)	(0.0368)	(0.0326)	(0.0220)	(0.0356)	(0.0187)
Home Team x 1st Mistake to Home	-0.0141	-0.0173	-0.0497	0.0042	0.0184	-0.0200
	(0.0254)	(0.0570)	(0.0545)	(0.0285)	(0.0444)	(0.0301)
N	2,460	418	636	1,406	372	2,088
R^2	0.445	0.442	0.442	0.449	0.478	0.440
Mean	0.100	0.110	0.107	0.094	0.048	0.109

Note: Standard errors are clustered by the referee and unique game and are in parentheses.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

Table 7: Second Error under the VAR and VAR/No-Crowd Regimes

	(1) VAR	(2) VAR / no-crowd
Home Team	0.0190	-0.0345
	(0.0308)	(0.0603)
Home Team x 1st Error to Home	-0.0520	0.0345
	(0.0361)	(0.0840)
N	392	128
R^2	0.4720	0.4737
Mean	0.0587	0.0547

Note: Standard errors are clustered by the referee and unique game and are in parentheses.

Stars denote the level of statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We control for unique game dummies.

References

Balmer, N., A. M. Nevill, and A. M. Williams (2002) "The Influence of Crowd Noise and Experience upon Refereeing Decisions in Football," *Psychology of Sport and Exercise* 3, 261-272.

Bartling, B., L. Brandes, and D. Schunk (2015) "Expectations as reference points: field evidence from professional soccer," *Management Science* 61, 2646-2661.

Berdejó, Carlos, and Noam Yuchtman (2013), "Crime, Punishment, and Politics: An Analysis of Political Cycles in Criminal Sentencing," *Review of Economics and Statistics*, 95(3), 741–756.

Boeri, T., and B. Severgnini (2011), "Match rigging and the career concerns of referees," *Labour Economics*, 18(3), 349-359.

Bryson, A., P. Dolton, J. J. Reade, D. Schreyer, and C. Singleton (2021) "Causal effects of an absent crowd on performances and refereeing decisions during Covid-19," *Economics Letters* 198, 109664.

Buraimo, B., Forrest, D., and R. Simmons (2010),"The 12th man?: refereeing bias in English and German soccer," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2), 431-449.

Carlos, C., Ezequiel, R., and Anton, K. (2019) "How does Video Assistant Referee (VAR) modify the game in elite soccer?", *International Journal of Performance Analysis in Sport*,19, 646-653

Cohen, A., Klement, A., and Neeman, Z. (2014), "Judicial Decision Making: A Dynamic Reputation Approach," *Journal of Legal Studies* 44, 133-159.

Dawson, P., Dobson, S., Goddard, J., and Wilson, J. (2007), "Are football referees really biased and inconsistent?: evidence on the incidence of disciplinary sanction in the English Premier League," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1), 231-250.

Dawson, P., Massey, and Downward, P. (2020) "Television match officials, referees, and home advantage: Evidence from the European Rugby Cup," *Sport Management Review*, 23, 443-454.

Dohmen, T. J. (2008), "The influence of social forces: Evidence from the behavior of football referees," *Economic Inquiry*, 46(3), 411-424.

Dohmen, T. J., and J. Sauermann (2016), "Referee Bias," *Journal of Economic Surveys*, 30(4), 679-695.

Endrich, M., and T. Gesche (2020) "Home-bias in referee decisions: Evidence from "Ghost Matches" during the Covid19-Pandemic," *Economics Letters* 197, 109621.

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

Gomery, John H. (2006), "The Pros and Cons of Commissions of Inquiry, " *McGill Law Journal*, 51(4), 783-798.

Huber, Gregory A., and Sanford C. Gordon (2004), "Accountability and Coercion: Is Justice Blind When It Runs for Office?" *American Journal of Political Science*, 48(2), 247-263.

Lago-Peñas, C., Gómez, M. A., and Pollard, R. (2020) "The effect of the Video Assistant Referee on referee's decisions in the Spanish LaLiga." *International Journal of Sports Science Coaching*

Marder, Nancy S. (2014), "Jurors and Social Media: Is a Fair Trial Still Possible," *Southern Methodist University Law Review* , 67(3), 617-668.

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

Pettersson-Lidbom, P. and M/ Priks (2010) "Behavior under social pressure: empty Italian stadiums and referee bias," *Economics Letters* 108, 212-214.

Phillipson, Gavin (2008), "Trial by the Media: The Betrayal of the First Amendment's Purpose," *Law & Contemporary Problems*, 71(15), 15-29.

Scoppa, V. (2021) "Social pressure in the stadiums: Do agents change behavior without crowd support?" *Journal of Economic Psychology* 82, 102344.

Shepherd, Joanna M. (2011), "Measuring Maximizing Judges: Empirical Legal Studies, Public Choice Theory, and Judicial Behavior," *University of Illinois Law Review*, 2011, 1753-1766.

Steblay, Nancy Mehrkens, Jasmina Besirevic, Solomon M. Fulero, and Belia Jimenez-Lorente (1999), "The Effects of Pretrial Publicity on Juror Verdicts: A Meta-Analytic Review," *Law and Human Behavior*, 23(2), 219–235.

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