Cultural Homophily and Collaboration in Superstar Teams

Gábor Békés[a,*], Gianmarco I.P. Ottaviano[b]

[a] *Central European University, KRTK and CEPR*
[b] *Bocconi University, Baffi-CAREFIN, IGIER, CEP and CEPR*

Abstract

One may reasonably think that cultural homophily, defined as the tendency to associate with others of similar culture, affects collaboration in multinational teams in general, but not in superstar teams of professionals at the top of their industry. The analysis of an exhaustive dataset on the passes made by professional European football players in the top-5 men's leagues reveals that, on the contrary, cultural homophily is persistent, pervasive and consequential even in superstar multinational teams of very high skill individuals with clear common objectives and aligned incentives, and involved in interactive tasks that are well-defined and not particularly culture intensive.

Keywords: Organizations, teams, culture, homophily, diversity, language, globalization, big data, panel data, sport

Revised version: December 2023.

1. Introduction

To compete in the global economy, companies are increasingly calling on a multinational workforce. As discussed, for instance, by Neeley (2015), this has pros and cons (Lazear, 1999*b*,*a*). On the one hand, a multinational workforce allows companies to build teams that feature the best talent from around the world, and draw on the benefits of international diversity by bringing together people from many cultures with varied work experiences and perspectives. On the other hand, teams like these also

[*] Corresponding author, mail: bekesg@ceu.edu, Central European University, Quellenstrasse 51, Vienna, Austria.

[**] Békés thanks the support of the 'Firms, Strategy and Performance' Lendület grant of the Hungarian Academy of Sciences. We thank Paola Conconi, Gábor Kézdi, Miklós Koren, Balázs Kovács, Marc Kaufmann, Alice Kugler, Mats Koster, Glenn Magerman, Marco Molitor, Balázs Muraközy, Balázs Lengyel, Ádám Szeidl and seminar participants at CERS-HAS, CEU, IMT, and ULB for useful comments and suggestions. We are grateful to Endre Borza, Bence Szabó for outstanding and extensive research assistance.

face several hurdles. When team members have different cultural background, communication can rapidly deteriorate, misunderstanding can ensue and cooperation can degenerate into distrust. Collaboration may end up being reconfigured along cultural lines leading to homophily, defined as the tendency to associate with similar others. Such reconfiguration can limit the gains from diversity, but it is not by itself necessarily inefficient as long as it may allow the team to minimize objective barriers to collaboration. It becomes inefficient when it is the result of team members' overestimation of the abilities of similar others, or simply a fundamental irrational indication of in-group partiality.

While the downsides of multiculturality can be mitigated and possibly eliminated by careful team selection, training and tasking, we show that cultural homophily is persistent and pervasive even in superstar multinational teams of very high skill individuals with clear common objectives and aligned incentives, and involved in interactive tasks that are well-defined and not particularly culture intensive. This is a context where most of the understood antecedents of homophily are not present, or at least are at a minimum (McPherson et al. (2001); Lawrence and Shah (2020)). We also show that the observed homophily has consequences for team performance. We then argue that, in such context, homophily can be more consistently interpreted as a way to circumvent objective barriers to collaboration rather than as an expression of subjective favoritism. Finally, we conjecture that the observed cultural homophily in the workplace could be due to the fact that individuals with similar cultural traits have more opportunities to become friends outside the workplace.

We base our investigation on a newly assembled dataset recording all passes made by professional European football players in the top five men leagues (Premier League in England, Ligue 1 in France, Bundesliga in Germany, Serie A in Italy, La Liga in Spain) over eight sporting seasons (2011-12 to 2018-19) together with information on key players' and teams' characteristics.[1] Collaboration is the situation of two or more people working together to create or achieve the same thing (dictionary.cambridge.org). In our context, we measure collaboration as the 'pass rate' defined as the count of passes from a passer to a receiver relative the passer's total passes when both players are fielded together during a half-season. Passes are the essential building blocks of football. They represent how players work together for the common objective of scoring a goal. Importantly, as we also show, passes are positively correlated with winning more league points and achieving higher league standings. Hence, because passes are the main way by which a team moves the ball towards a position from which players can score, they are an adequate indicator of collaboration between teammates. Our dataset includes 10.7 million passes made in 14,608 games by 6,998 players from 138 countries fielded by 154 teams. Pass data are aggregated to obtain the sum of passes by passer-receiver player pairs in a half-season, a time period composed of 16-20 games (depending on

[1] With 'European football', or simply 'football' henceforth, we refer to 'association football', which is commonly known as 'football' in Europe and 'soccer' in the United States (Tovar, 2020) .

the country) in which football clubs have stable squads. Aggregated pass data are then matched with detailed information on player characteristics. The analysis is carried out on the resulting three-dimensional (unbalanced) panel of 669 thousand passer-receiver pairs over 16 half-seasons.

This dataset has several unique advantages in terms of investigating cultural homophily and collaboration in superstar teams. First, it allows us to study a host of multinational workplaces where teams are not geographically dispersed. This is seldom the case in multinational companies where geographical dispersion may *per se* inhibit collaboration (Joshi et al., 2002). Second, the European football (soccer) industry is very globalized: it is the world's most popular sport, fans are spread around the globe, and multinational teams are the rule in the top five leagues.[2] Third, these leagues represent the pinnacle of the industry with superstar companies that can be expected to act as such with regard to team selection, training and tasking. Fourth, football players are very mobile internationally, and their mobility decisions are typically made for work-related reasons, with pay being the most prominent of them. Fifth, in the top five leagues players are very diverse in terms of origin as they come from over a hundred countries. At the same time, they are all very high skilled (and well-paid) workers hardly facing obstacles with integration outside the workplace. Moreover, while language may matter for collaboration, the role of language as a sheer means of communication rather than a broader cultural trait is unlikely to dominate as football tasks are not particularly language intensive (Nüesch and Haas, 2013). Sixth, all sorts of player as well as team characteristics and performance indicators are precisely measured, and fastidiously recorded. Moreover, extensive media coverage can be readily used to shed light on any odd data patterns. Seventh, while team composition is exogenous to players' decisions, collaboration with other team members is mostly up to their individual choices. Eighth, the 'rules of the game' are codified, and crystal clear to all players and teams, ruling out the possibility that players of any specific culture may collaborate more with one another only because they happen to have a better grasp of those rules than other players. Last, unlike most other team ball games, the rules always leave a player with a wide range of options on which teammate to pass to.[3] We are not aware of any other empirical setup in the existing literature that ticks all these boxes.

All the foregoing features of the dataset allow us to directly investigate collaboration in competitive global teams of high skilled workers with precise common objectives, leveraging a big dataset on interactions in an actual workplace rather than in an artificial experimental lab (Jackson et al., 2003), while also exploiting an extremely rich set of team and worker controls. Moreover, the fact that all players are men allows us to analyze how cultural barriers affect collaboration in multinational teams separately

[2] On average teams in our sample have a squad of players from 13 countries and field a starting eleven with players from 6 countries.

[3] Yet another advantage of our football dataset is that the network of links (passes) between nodes (players) has maximum density, which has important implications for our empirical strategy as Section 2 will discuss in detail.

from issues of gender diversity.[4]

We characterize the cultural background (henceforth, simply ‘culture’) of team members in terms of a set of cultural traits (Spolaore and Wacziarg, 2016; Desmet and Ortuño-Ortín, 2017). These include norms, values and attitudes that are transmitted intergenerationally, which we operationalize through nationality. We also test alternative operationalizations based on colonial legacy (past membership of a colonial empire), federal legacy (past membership of a political union), native language, linguistic and geographical proximity, and shared values.

To illustrate what we are after, it is useful to go through a simple concrete example involving 24,299 Spanish midfielders of the Spanish league and their pooled passes over the period of observation. We pick midfielders as they have the greatest freedom of choice in passing and are approximately equally distanced from teammates during a game. We look at the relation between a midfielder's pass rate to a given receiver and the latter's market value as a consensus measure of player quality. We split the sample of receivers in two groups, depending on whether or not they have the same culture as the passer.

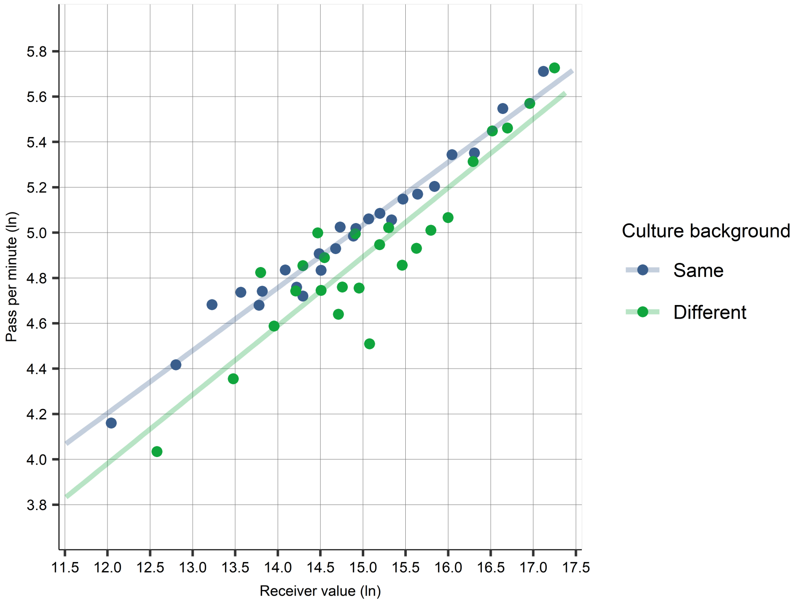
The result is Figure 1, which depicts a binscatter of (log) pass rate against (log) receiver value with fitted linear regression lines for the two groups of receivers. The bins are constructed to simplify the scatter plot. There are 25 of them for each receiver group with a bin containing about 500 players. The figure reveals a clear positive relation between pass rate and passer quality for both groups. However, the higher fitted line for same culture receivers shows that, for given receiver quality, the pass rate is higher to same than to different culture receivers. The gap between the two fitted lines visualizes homophily.[5]

Despite the visual appeal of Figure 1, pushing the analysis beyond its expressiveness is not straightforward. The key methodological challenge has been highlighted in several studies on homophily (Lawrence and Shah, 2020). That team members of same culture collaborate more than team members of different culture is a statement about homophily. It highlights common cultural traits as the antecedents of homophily, that is, the specific attributes that serve as its basis, while singling out collaboration as the targeted consequence of homophily (Ertug et al., 2021). In this respect, in studying homophilous behavior an important distinction has been made between two underly-

[4] There is a growing literature investigating the influence of gender composition on group performance and decision making. See, e.g., Adams and Ferreira (2009) and Apesteguia et al. (2012) on how boardroom gender diversity relates to measures of corporate performance; Ahern and Dittmar (2012) Ahern and Dittmar (2012) and Matsa and Miller (2013) on how the introduction of gender quotas for directors affect firm value; Adams and Funk (2012) on how those findings may be explained not only by the different behavior of diverse boards but also by inputs into board behavior that vary with boardroom gender diversity.

[5] In Figure 1 there seems to be a difference in slope between the two fitted lines. This is, however, not statistically significant. The same applies not only to this subset of Spanish midfielders but also to the full dataset.

Figure 1: Pass rate, receiver value and homophily



Note: Sub-sample: Spanish midfielders in La Liga, 2011-2019, all half-seasons, N=24,299. Same cultural background is equality of language, colonial legacy, federal legacy or nationality. Player values at the start of the half-season. Bin scatter and lines by cultural background groups, 25 bins each group. Linear regression lines fitted.

ing mechanisms: opportunities and preferences (McPherson and Smith-Lovin, 1987; McPherson et al., 2001). According to the former mechanism, individuals' distributions across categories within a social context define the probability they choose similar others (Lawrence and Shah, 2020). This may mechanically 'induce' homophily, irrespective of whether players have any actual preference for similar others, and thus it may not tell much about their real tendency to associate with similar others. Lawrence and Shah (2020) offer the following simple example. Consider a group of 100 geoscientists who associate with one another during a conference workshop. If 40 percent are geochemists and 60 percent are hydrologists, the expected rate for geochemists associating with other geochemists is 0.40. Only when the proportion of geochemists' associations with other geochemists exceeds this baseline, it demonstrates a preference for geochemists to associate with other geochemists. It is this preference that distinguishes 'choice homophily' from 'induced homophily'.[6] Hence, to be of any interest at

[6] In the sociological literature the tendency of people of different types to associate with similar others in excess of the baseline of their types' relative population sizes is also called 'inbreeding homophily' (e.g. Coleman (1958); Marsden (1987); McPherson et al. (2001). See also Currarini et al. (2009).

all, the statement that team members of same culture collaborate more has to be based on choice homophily after controlling for induced homophily, that is, for any constraint on the choice set.

Defining the baseline is quite straightforward in the previous example. It is much less so when individuals may or may not differ along several potential attributes that could confound the roles of the targeted antecedents of homophily, making it harder to ascertain whether individuals are mechanically induced to choose similar others. We address these identification issues by designing the baseline in terms of a discrete choice model of players' passing behavior. The model determines how the pass rate for a pair of players is pinned down by their characteristics and opportunities when fielded together in a given time period. It is implemented empirically by a Poisson regression with player characteristics as well as player by half-season fixed effects as controls. As we discuss, deriving the regression from the discrete choice model is important in order to figure out that player by half-season fixed effects also control for players' choice sets when passes are made. Results are then corroborated by a rich set of robustness checks.

We find strong evidence of choice homophily: players have a preference to pass more to players of their same culture than to other players. In a regression with passer by half-season fixed effects as well as receiver by half-season fixed effects, conditioning on pass features shows that player pairs of same nationality have a pass rate 2.42% higher than player pairs of different nationality. Among the alternative operationalizations of culture, only same colonial legacy is as consequential as same nationality, while all others have little relevance. The estimated homophily is important for team outcomes. If in the typical team all players had different culture, in a counterfactual half-season in which all its players shared the same culture it would make between zero and one points more than its usual points. For a team usually in the top or the bottom parts of the standings, one point more could still make a big difference in terms of sporting outcomes by allowing the team to participate to lucrative and prestigious international competitions or to avoid costly and humbling relegation to a lower league. Computing a monetary equivalent value of the estimated homophily, we find that passing to a receiver with same culture is as likely as passing to a receiver with different culture but valued a remarkable 10.5% more, which corresponds to 367'500 and 809'550 euros for the median and average players as these are valued 3.50 and 7.71 million euros respectively.

We then discuss whether the estimated homophily premium is more likely due to an objective cost for the team as passes are easier to coordinate within a group ('cost saving') or to a subjective cost for the passer who prefers to keep the ball within his own group ('favoritism'). We rely on three identifying assumptions. Favoritism should subside when the stakes are high for the passer as pondered decisions are more likely in this case. Favoritism should be more visible when the passer belongs to a minority group, in which case we should observe more homophily if the passer belongs to a small rather than to a large group. Focusing on passes between players of different culture, favoritism should promote passes to smaller groups as these are less likely to keep the ball within them, whereas cost saving should promote passes to larger groups as these are more likely to keep the ball within them and thus reduce a team's passing frictions. We find that homophily is not less (and in some case is actually more) pronounced when stakes are high and is higher for players in large groups than for those in small groups. Moreover, passes are more likely from small to large groups than from large to large groups, and passes from large to large groups are more likely than those to small groups, no matter whether these originate from large or small groups. Accordingly, homophily is unlikely due to favoritism and the higher frequency of passes to large groups is consistent with the idea that keeping the ball in those groups minimizes a team's passing friction.

Finally, we discuss possible mechanisms behind the observed cultural homophily that have been highlighted in the literature. First, people who share a national identity may want same nationals to do well. Our results showing how post-colonial links matter do not support the national identity mechanism. While it may still support a mechanism working through broader cultural identity, our finding that inter-group passes favor receivers in large groups is inconsistent with in-group favoritism as large homophilous groups are more likely than small homophilous groups to keep the ball within them. Second, people may have a false overly confident belief in the ability of same culture peers and underestimate the ability of different culture ones. While the prejudice mechanism can be detected if homophily declines with shared experience, we find the opposite to hold. Third, being in a small group leads to salience of cultural affiliation as people are more likely to be aware of belonging to the same group when the group is small, which is not supported by our findings on group size and inter-group passes. Fourth, people may have no personal preference when they start collaborating, but will build friendships over time through interaction inside and outside the workplace. Friendship may be easier to build with people from the same culture, as they typically speak the same language, have the same social cues, like the same cuisine, listen to the same music, watch the same TV series or sports, and so on. While we do not know how much time players pass together outside the workplace, we still observe that homophily increases with the time they spend in the same clubs.

Overall, we conjecture that our results can be consistently interpreted as hinting at cost saving as the source of homophily and off-pitch familiarity as a possible mechanism through which same culture leads to homophily.

The rest of the paper is organized as follows. Section 2 offers a selective overview of the related literature beyond works already referenced in this introduction. Section 3 describes data collection and our dataset. Section 4 introduces the discrete choice model of passing behavior and its estimation, whose results are then discussed in Section 5. Section 6 tackles the distinction between the cost saving and in-group favoritism nature of homophily, and possible mechanisms behind it. Section 7 concludes.

2. Related literature

Collaboration is the essence of teamwork and one may reasonably think that cultural preferences affect collaboration in multinational teams in general, but not in superstar teams of professionals at the top of their industry. We reject this hypothesis by studying cultural homophily and collaboration in multinational teams of high-skilled individuals leveraging data on professional players in the top European football leagues. While in the Introduction we have listed several reasons why such data has unique advantages for our purposes, in this section we place our contributions in the context of the existing literature on homophily and social interactions.

The literature on homophily is vast and interdisciplinary. Recent influential surveys emphasize different aspects and methods depending of their specific discipline of interest. For example, McPherson et al. (2001) look at homophily through a sociological perspective Lawrence and Shah (2020), as well as Ertug et al. (2021) emphasize the management viewpoint, whereas Jackson et al. (2017) discuss homophily from the perspective of economics with a focus on social network.[7] The aim of this section is to build on those surveys to assess the paper's main contributions.

Lazarsfeld and Merton (1954) define homophily as the tendency for friendships to form between those who are alike in some designated aspect. The concept has then been extended beyond friendship as the principle that contact between similar people occurs at a higher rate than among dissimilar people (McPherson et al., 2001) or the tendency of individuals to associate with similar others (Lawrence and Shah, 2020). Following Jackson et al. (2017) and Ertug et al. (2021), we take these consistent and widely agreed definitions as collectively providing a clear sense of what homophily is: "birds of a feather flock together ", as the proverbial expression originally quoted by Lazarsfeld and Merton (1954) has it.

Homophily is an extensively studied and well-documented feature of social interaction (Currarini and Mengel, 2016). It has been found to manifests itself along many dimensions of similarity and typologies of social ties with important consequences on individual and societal outcomes, raising wide-ranging policy issues such as the discussion about "parallel societies" (sex-)segregated education, the costs and benefits of cultural diversity, the management of ethnic conflicts, and the design of fair and efficient organizations among many others.

Against this backdrop, three important distinctions made in the literature are particularly relevant for our purposes. The first is the distinction between "induced" and "choice" homophily already discussed in the Introduction. McPherson et al. (2001) contrast the homophily effects created by the demography of the potential tie pool, which they call "baseline" homophily, with homophily measured as explicitly over and above the opportunity set, which they call "inbreeding" homophily. While the former

[7] Homophily has also been thoroughly studied in education studies (Terenzini et al. (2001)) and social psychology (Dovidio and Gaertner, 2010).

is determined by the constraints limiting the actors' choice set, the latter results from their preferences within the constrained choice set available to them, as clarified in the geoscience workshop example by Lawrence and Shah (2020) reported in the Introduction.

Currarini et al. (2009) and Currarini et al. (2010), for instance, study friendship formation in US schools when students have types (ethnicities) and may see type-dependent benefits from friendships. They show that any baseline matching process such that types are matched in frequencies in proportion to their relative stocks cannot replicate the homophily they observe in their data. On the contrary, a static model with both type-sensitive preferences ("choosing friends") and a matching bias ("meeting friends through friends") generates the observed patterns of homophily. In this respect, an actor's observed behavior can be understood as deriving from preference-driven decisions ("choice") over constrained alternatives ("opportunities") In the same vein, Jackson et al. (2017) highlight the role of homophily in determining the observed deviations of social network structures from the patterns determined by degree distributions invariant to relabeling the network nodes.[8] In the presence of homophily, social networks can exhibit strong segregation patterns due to the fact that there are relatively fewer links connecting nodes of different types when most of the links connect similar nodes.

When it comes to telling choice from induced homophily, the most important challenge stressed in the literature concerns the assessment of the constraints on individual choice sets. As this assessment typically represents a daunting task in observational studies, carefully designed experiments have been increasingly used across disciplines (Jackson et al., 2017; Lawrence and Shah, 2020; Ertug et al., 2021). The first contribution of our analysis is to show how choice can be purged from opportunities also in observational studies as long as adequate network information is available and the constraints on an actor's choice set in the network are fixed. The basic idea is that such fixity makes the actor's choice set an unobserved characteristic that can be absorbed through an individual fixed effect, clearing the ground for the identification of choice homophily. We develop this idea in a theory-based way through a discrete choice model of bilateral collaborative ties in a network that allows for the structural interpretation of an actor's individual fixed effect as including also the average attractiveness of all available alternatives against which the actor's choice can be benchmarked.[9]

The second important distinction made in the literature concerns the antecedents of homophily, that is, the dimensions of similarity on which homophily is based. Following Lazarsfeld and Merton (1954), McPherson et al. (2001) distinguish between "status"

[8] In a network, the degree of a node is the number of its connections to the other nodes, and the degree distribution is the probability distribution of those degrees over the entire network.

[9] This interpretation of individual fixed effects is analogous to that of country fixed effects in the gravity equations used to model bilateral flows within international trade networks where each country has several potential trade partners (Head and Mayer, 2014).

and “value” homophily. Whereas in the former case the basis for similarity is informal, formal, or ascribed status (such as age, behavior patterns, education, ethnicity, gender, language, occupation, race or religion), in the latter it is values, attitudes, beliefs and other internal states orientating future behavior. With respect to many antecedents, physical proximity is a key mediator, given that it takes more effort to connect with someone who is further away (Zipf, 1949). The role of proximity has been studied at different levels of aggregation showing that distance-related factors play an important role in creating and reinforcing social ties. McPherson et al. (2001) highlight the importance of various “organizational foci”, that is, places where focused activity puts people into contact with one another to foster the formation of personal relationships (Feld, 1981, 1982, 1984). The most prominent of such places in the literature are school and work, which are both relevant for our analysis and are the subjects of imposing bodies of studies with a focus on friendship, education and organization Currarini et al. (2009); Jackson et al. (2017).

Differently the much of the existing literature, in our multinational teams of superstar footballers most antecedents highlighted in previous studies are expected to be of little relevance. Hence, a second contribution of our analysis is to show that this indeed applies to value, but not to status homophily: there is more collaboration within than between individuals of more similar status even in superstar multinational teams of very high skill workers with clear common objectives and aligned incentives, and involved in interactive tasks that are well-defined and essentially status blind. We then argue that persistent and pervasive status homophily is likely due to friendship ties created outside the workplace. This in line with existing evidence that, when employees have more social interactions with their managers, they are promoted at a higher rate (Cullen and Perez-Truglia, 2023). Our findings can thus be interpreted as revealing that also among peers privileged treatment in the workplace can stem from social interactions outside the workplace (“old boys’ club”).

The third important distinction in the literature constrasts the antecedents with the consequences of homophily. Taking stock of the works surveyed by McPherson et al. (2001) and Lawrence and Shah (2020), Ertug et al. (2021) state that, while the antecedents of homophily are well documented and understood, the same cannot be argued about its consequences, defined as outcomes that go beyond the formation of ties and relationships. In this respect, a crucial trade-off has been highlighted and investigated especially in management and economics. On the one hand, homophily facilitates coordination, communication and trust (Opper et al., 2015; Castilla, 2011), with most studies of this aspect presuming the formation of a relationship between actors based on similarity and investigating its consequences. On the other hand, homophily reduces diversity in knowledge, perspectives and network reach (Burt, 1992; Cross and Cummings, 2004; Horwitz and Horwitz, 2007). Which aspect is more relevant determines whether the impact of homophily on outcomes is positive, negative or neutral, and this in turn depends not only on the specific outcomes studied, but also on the specific antecedents of homophily identified, the specific context considered by the analysis, and other specific contingency factors that can act as moderators (Jackson et al., 2003). For example, in social networks the distribution of connectivity, as determined by the formation of ties and relationships, has been shown to be critical in fostering or hampering all sorts of diffusion processes with important economic consequences. According to Jackson et al. (2017), a general conclusion of the existing literature is that homophily and segregation patterns can both help and hinder diffusion. High levels of homophily can lead groups to be insular and maintain different customs and behaviors from other groups, which can enable poverty traps and underinvestment due to complementarities in behaviors, as well as slow the spread of information across groups. At the same time, high homophily can also help incubate behaviors within one group that might not take hold otherwise, and these can then spread outside the group (Currarini and Mengel (2016)).

The consequence of homophily we focus on is collaboration that matters for team performance. In this respect, two research streams are particularly relevant to us. The first is concerned with 'diversity spillovers', which improve economic performance in a diverse environment (Ottaviano and Peri, 2006, 2005). Mostly focusing on urban cultural diversity related to country of birth, this stream highlights four main channels through which a diverse environment affects economic performance (Buchholz, 2021). Diversity increases productivity: (i) when people from different countries work on problems together, in turn identifying better solutions by combining their knowledge ("interactive problem-solving"); (ii) when a diverse environment promotes the specialization, the variety of skills and the approaches to tasks within an occupation, though without necessarily requiring interaction between people from different countries of birth ("complementary task specialization"); (iii) when people from the same country of birth cluster in particular occupations and this clustering facilitates stronger knowledge exchanges ("niching effects"); and (iv) when simply through exposure to a diverse range of knowledge and approaches to problems workers learn and become more productive ("exposure effects"). Within this taxonomy, homophily is ambivalent: it hampers channels (i) and (iv), but fosters channels (ii) and (iii). The evidence on US Metropolitan Statistical Area reported in Buchholz (2021) supports exposure effects as the main channel, but also interactive problem-solving and complementary task specialization seem to play an important role.

This first stream does not leverage information on diversity and collaboration within teams, which is what we do. In this respect, our investigation is more closely related to a second research stream that studies how individuals of different ethnicities may complement each other in production, but workers of the same ethnic background may collaborate more effectively (Lazear, 1999*b*,*a*; Lang, 1986).[10] Specifically related to our investigation are works highlighting how distortions due to ethnic diversity and discriminatory worker attitudes affect firms and their organization of production. These

[10] See (Jackson et al., 2003) for a review of studies on the effects of workplace diversity on teams and organizations.

studies face stiff data challenges. To systematically examine the effects of culture and language within a firm, one needs a host of detailed data: the nationalities of all workers must be identifiable, each worker's skills and output as well as the collective output of the firm must be measurable, and all other factors of production should be held constant (Kahane et al., 2013). That is why works on firms typically rely on experiments that, differently from observational analyses, allow for controlled environments (Bertrand and Duflo, 2017).

Homophily and discrimination due to "in-group bias" have been found to be particularly salient by experiments in the context of developing countries.[11] For instance, Hjort (2014) studies team production at a plant in Kenya, where an upstream worker supplies and distributes flowers to two downstream workers, who assemble them into bunches. He finds that upstream workers undersupply non-coethnic downstream workers ("vertical discrimination") and shift flowers from non-coethnic to coethnic downstream workers ("horizontal discrimination"), at the cost of lower own pay and total output. Team pay, whereby the two downstream workers are remunerated for their combined output, is shown to mitigate discrimination and its allocative distortions, whereas conflicts exacerbate discrimination.[12] In Hjort (2014), the upstream worker's decision on distributing flowers to the downstream workers resembles the choice a football player faces on passing the ball to his teammates. The context is, however, quite different. Whereas a Kenyan plant is a low skilled, highly charged context in a developing country with ethnic conflicts, a European football team is a high skilled, lowly charged context in a developed area with no real conflicts during our period of observation. Again, one would expect the ethnic antecedents of homophily to be quite negligible in the latter context.

Closer to our high-skill low-charge set-up are experiments with students. For example, Calder-Wang et al. (2021) exploit a dataset of MBA students who participated in a required course to propose and start a real micro-business that allows them to examine horizontal diversity (i.e., within the team) as well as vertical diversity (i.e., team to faculty advisor) and their effect on performance. The course was run in multiple cohorts in otherwise identical formats except for the team formation mechanism used. In several cohorts, students were allowed to choose their teams among students in their section.

[11] Whereas homophily refers to the tendency to interact with similar others, in-group bias refers to the tendency to treat others of shared social identity more favorably. Homophily and in-group bias are often intertwined. On the one hand, the former may be caused by the latter, if for example it originates from a rational reaction to the expectation of preferential treatment from the opponent (i.e. to the anticipation of in-group bias). On the other hand, by influencing the patterns of interaction in favour of homogenous contacts, homophily may itself lead to in-group bias (Currarini and Mengel, 2016).

[12] Hjort (2014) finds that a period of ethnic conflict following Kenya's 2007 election led to a sharp increase in discrimination at the flower plant. Using data from GitHub on collaborative efforts in coding, the world's largest hosting platform for software projects, Laurentsyeva (2019) finds that political conflict that burst out between Russia and Ukraine reduced online cooperation between Russian and Ukrainian programmers.

In other cohorts, students were randomly assigned to teams based on a computer algorithm. In the cohorts that were allowed to choose, Calder-Wang et al. (2021) find strong selection based on shared attributes.[13]. Among the randomly-assigned teams, greater diversity along the intersection of gender and race/ethnicity significantly reduced performance. However, the negative effect of this diversity is alleviated in cohorts in which teams are endogenously formed.[14]

A high-skill low-charge context also characterizes observational studies on homophily in scientific publications. For example, looking into scientific papers written by US-based authors from 1985 to 2008, Freeman and Huang (2015) find evidence of choice homophily as individuals of similar ethnicity co-author together more frequently than predicted by their proportion among authors, and greater homophily is associated with publication in lower impact journals and with fewer citations, even holding fixed the authors' previous publishing performance. By contrast, diversity in inputs by author ethnicity, location, and references leads to greater contributions to science as measured by impact factors and citations. In the same vein, AlShebli et al. (2018) study the relationship between research impact and five classes of diversity: ethnicity, discipline, gender, affiliation, and academic age. Using randomized baseline models, they establish the presence of homophily in ethnicity, gender and affiliation. However, ethnic diversity has the strongest correlation with scientific impact.[15]

A third contribution of our analysis is to improve on observational studies like these (with potential implication also for the econometrics of experimental studies) by leveraging information on the network structure of our dataset not only to design a sharper baseline as discussed above, but also to achieve better identification and estimation of the parameters regulating the existence and the intensity of homophily. This is another bonus of our theory-based empirical approach clarifying how individual fixed effects also control for the actor's unobserved choice sets. Proper estimation of individual fixed effects requires a dense network, that is, a network in which the number of links of each node is close to the maximal number of nodes. If this does not happen, fixed effect estimation is plagued by an identification bias and an incidental parameter bias that arises whenever many parameters are estimated with relatively few observations (Andrews et al. (2012)). While this is clearly relevant in the case of scientific publications, it is also a recurrent feature in several other contexts, from friendship and marriage to business collaborations and everyday interactions. In the

[13] See McFarland et al. (2014) for a discussion of homophily, self-selection and segregation in the context of adolescent social structures

[14] This result echoes the findings on voluntary organizations (McPherson and Smith-Lovin, 1987). In this respect, as long as the manager of a football team acts as mediator allowing the team to internalize the effects of diversity, the negative impact of diversity on collaboration we find can be seen as a lower bound estimate with respect to what would be found in randomly composed teams.

[15] See Lix et al. (2022) for a study of the effects of cognitive diversity on team performance as reflected by the degree to which the meanings conveyed by group members in a given set of interactions diverge from one another.

case of publications, among other unobserved individual characteristics, the estimated individual fixed effect captures an author's choice set of co-authors in an unbiased way only if the author writes with ideally all potential co-authors (Bonhomme et al. (2019)), which is not what one sees in the data where the co-authorship network is typically very sparse. As a sparse network has much fewer links than the possible maximum number of links, the numbers of individual fixed effects tends to be dangerously close to the number of observations used to estimate them, which gives rise to the incidental parameter bias.[16] Our data on football teams, within which all players interact with one another on a bilateral basis through passes, allows us to work with a large number of networks of maximum density.

While, to the best of our knowledge, we are the first to highlight the virtues of fixed effect estimation in the context of homophily studies on social networks and the related advantages of using team sports data, we are not the first to exploit such data to analyze the potential gains and losses from employing culturally diverse work teams. In the case of the top North American ice hockey league (National Hockey League), (Kahane et al., 2013) find that the presence of European players (with Europe being the typical origin of foreign players) does increase firm-level performance: teams that employ a higher proportion of European players perform better. However, their results also indicate that teams perform better when their European players come from the same country rather than being spread across many countries. When teams have players from a wide array of European countries, integration costs associated with language and cultural differences may start to override any gains from diversity. Parallel evidence based on European football leads to mixed conclusions. In the top German league multinational teams have been found to perform worse than teams with less national diversity (Nüesch and Haas, 2013), whereas the opposite has been found in top continental tournament (Ingersoll et al., 2017). Studying the top leagues of England and Spain, Tovar (2020) suggests that conflicting results may derive from a hump-shaped relation between team performance and predominant nationality, which is consistent with the findings of Earley and Mosakowski (2000) on trasnational management teams. This echoes (Kahane et al., 2013) in that an optimal degree of diversity may exist.

What distinguishes our analysis from these and related works is that we zoom in on collaboration (which we can measure accurately through the pass data) and study the relations between homophily, collaboration and team performance for given diversity, also investigating the possible relevance of various antecedents beyond nationality.[17]

[16] See Eeckhout and Kircher (2011) for a discussion of identification in assortative matching environments, and Andrews et al. (2012) for a test showing the sensitivity of individual fixed effects estimates to different degrees of network sparsity.

[17] Beyond diversity, team sports data are increasingly used to study different issues in labor and public economics. For example, close in spirit to the experimental design of Hjort (2014), Price and Wolfers (2010) use data on the National Baseball Association (NBA) to study racial discrimination by referees when awarding personal fouls against players. In the same vein, Parsons et al. (2011) exploit data from the top North American baseball league (Major League Baseball) on the way umpires judge

In addition, possibly of more general relevance are our investigation of whether the observed choice homophily in the workplace is due to an actor's efficient response to objective barriers to collaboration or inefficient in-group bias, and the discussion of the most likely mechanism behind such homophily. This is the fourth contribution to the existing literature we wish to highlight.

3. Data

To estimate how homophily may affect collaboration we use a novel dataset created by matching data from different sources. In this section we describe the scope of the unique data we use and briefly summarize the players' data, the data on passing events, and the combined dataset used for the model estimation.

The raw data used in the research had been web-scraped from different sources. The passing dataset as well as the game-level information are based on www.whoscored.com, a sports website. The player value and career data comes from www.transfermarkt.com, a football player information and valuation website.[18]

3.1. Background

Before discussing the data, an important caveat is in order. American readers should be aware that the analysis is not about what they call "football", it is about what they call "soccer". The term "soccer" appeared for the first time in England in the 1880s as an Oxford "-er" abbreviation of the word "association (football)". It was contrasted with "rugger" for "rugby (football)". Here we follow the European convention and call our subject matter simply "football".[19]

Football in this sense is a game played between two teams of 11 players. who move a ball around a rectangular field, called "pitch", during a match consisting of two 45-minute halves. The match is won by the team that scores more goals by moving the ball inside a rectangular-framed "goal" defended by the opposing team. The borders of the pitch are identified by two shorter "goal lines" and two longer "touchlines". The application of all rules during the game is overseen by a "referee"with the help of two "linesmen".

Passing the ball is a crucial component of football as it allows a team to control and maneuver the ball with the objective of putting one of its players in the best condition

throws by pitchers of different race/ethnicity to study discrimination and its impact on discriminated groups' behavior. Kleven et al. (2013) rely on data for professional football players in Europe to shed light on the international mobility responses of workers to tax rates and their impact on local labor markets. Arcidiacono et al. (2017) use data on the top North American basketball league (National Basketball Association) to assess whether worker compensation is influenced by productivity spillovers to coworkers. Gauriot and Page (2019) use European football data for the five top leagues to understand how workers' valuations may be affected by luck.

[18] Replication codes and data for the empirical analysis are available on the author's website.

[19] Szymanski (2010), D'Orsogna and Ottaviano (2011).

to score a goal. In modern football passing has become increasingly prominent with respect to dribbling, whereby a player alone takes the ball forwards past opponents.

Players are allowed to move the ball and defend the goal with their feet, and occasionally with any other parts of their body except their hands and arms. Only one player per team, called "goalkeeper", is allowed to use also his hands and arms, but only within a rectangular area in front of the goal. The other ten players in the team are arranged across the pitch in positions that belong to three broad categories: "defenders" are closer to their own team's goal, "forwards" are closer to the opponent team's goal, and "midfielders" are between them. Within these categories, there are more detailed assignments with role names that vary with the specific layout of the team as chosen by the coach.

When a player unlawfully touches the ball with his hands or arms, he commits a "foul" and the ball is consigned to the opposing team for a "free kick". The same happens when a player contrasts an opponent with excessive energy, or receives the ball with only one opponent separating him from the team's goal of the latter when the ball is initially passed to him ("offside").[20] Whenever the ball exits the pitch, it is given to an opponent of the last player who was in contact with the ball before it exited. The opponent returns the ball to the pitch by a "throw-in" if, in exiting, it passed over one of the pitch's lateral touchlines. He does so by a "corner kick" or a "goal kick" if the ball passed over the goal line of the defending or the attacking teams respectively. For the purposes of the present analysis "passes" include also free kicks, throw-ins, corner kicks and goal kicks as long as the ball stays with the same team. All these together represent, however, only about 5% of all passes.

3.2. Scope

The dataset covers professional European football in the top five men leagues: Premier League in England, Ligue 1 in France, Bundesliga in Germany, Serie A in Italy, La Liga in Spain over eight sporting seasons. These leagues were selected because of their undisputed reputation as the pinnacle of national football competitions. Moreover, data availability is the most comprehensive for these leagues.

The dataset covers all games played in the sporting seasons from 2011-12 to 2018-19, which offer the highest data quality and predate the COVID-19 pandemic. A season is the time period between mid-August to mid-May, during which each team plays twice (home and away) with every other team in its league.

A season is composed of two halves: the Fall half-season runs from mid-August till the end of December, the Winter-Spring runs till mid-May. The Premier League, La Liga, Serie A and Ligue 1 are all composed of 20 teams (playing $20\times19=380$ games),

[20] When a player commits a foul inside the area where his goalkeeper can touch the ball also with his hands and arms, a "penalty kick" is awarded, which allows the opponent team a free kick from a central spot inside the area with no other players allowed to be there except the kicker and the defending goalkeeper. For this reason, that area is called the "penalty area".

while there are 18 teams (18×17=306 games) in the Bundesliga. In any given season, there are 98 teams in our sample, and we have 98x16=1568 team by half-season units in our dataset. Due to relegation and promotion, we have a total of 154 teams in the sample. Overall, our dataset covers a total of 8×(380×4+306)=14,608 games.[21]

3.3. Player dataset

We have 6,998 players in our sample, for whom we can fully map their entire career, with a typical team relying on a squad of about 30 players. For every player, data include his country of birth, single or multiple citizenship information, country of birth, date of birth, height, and participation in a national team. These are all time-invariant in our dataset.

European football is truly globalized as there are players from 138 countries of citizenship in our sample. French, Spanish and Italian players make up the largest citizenship groups, followed by Germans, English, Brazilians and Argentinians. Other countries of citizenship with several players include the Netherlands, Serbia, Senegal, and Uruguay. Table 1 reports the share of countries in terms of first citizenship of players, for countries with at least a 1% share.

To characterize the cultural background (or, simply, 'culture') of team members we consider a set of cultural traits (Spolaore and Wacziarg, 2016; Desmet and Ortuño-Ortín, 2017). These include norms, values and attitudes that are transmitted intergenerationally, which we proxy through nationality, native language, colonial legacy (past membership of a colonial empire) and federal legacy (past membership of a political union). Nationals of the same country are more likely to share a common heritage, covering not only inherited traditions, monuments, memories and objects, but also contemporary institutions, activities, meanings, and behaviors drawn from them. Common language and shared history due to bygone political ties may allow individuals of different nationality to share at least some aspects of such a common heritage. This is a central theme of the literature on colonial past and intercultural relations (Bobowik et al., 2018), according to which the way both formerly colonizing and colonized peoples remember the colonial past determines their collective identities and intergroup behavior in the present. Research in this area indeed shows that collective memories may hinder or improve present-day intergroup relations, depending on the way the past is remembered or framed, and stresses the importance of considering colonial legacies in the study of contemporary intercultural relations.[22]

[21] Data quality and coverage are both very high in our datasets. Nevertheless, a few small data cleaning steps were needed and we discuss them in Appendix D

[22] In Section 5.2.3 we provide a critical assessment of our operationalization of culture, discussing in finer detail not only the impact of bygone political ties, but also of language and geographical proximity. There we also explore alternatives based on people's values and beliefs as reported in the World Values Survey (www.worldvaluessurvey.org). As we will see, colonial legacy tends to improve rather than hinder collaboration, whereas the opposite holds for federal legacy. Alternatives to our operationalization do not appear to be as consequential.

Table 1: Most frequent nationalities

Country	share (%, all players)
Spain	13.5
France	12.1
Italy	9.8
Germany	8.4
England	6.9
Brazil	4.3
Argentina	3.4
Portugal	1.8
Netherlands	1.6
Senegal	1.5
Belgium	1.3
Serbia	1.2
Uruguay	1.2
Switzerland	1.2
Cote d'Ivoire	1.1
Croatia	1.1
Morocco	1.0
Denmark	1.0

Player level dataset, frequency of first citizenships. List of countries with at least a 1% share.

To determine whether two players have the same culture, we consider the chosen cultural traits as follows. First, nationality is defined based on citizenship of a country. As some players have multiple citizenships, we define two players as same nationality if they share at least one of them, or have the same country of birth. Second, to ascertain common colonial legacy, we use colonial links data from CEPII as Head and Mayer (2014). We define two players as sharing the same colonial legacy if their nationalities include a former colonial ruler and its subject (e.g. Spain and Argentina) or two subjects of the same colonial ruler (e.g. Argentina and Uruguay).

Third, by common federal legacy we refer to countries that formed political unions some time in the 20th or 21st centuries. These include: (i) countries of the former Soviet Union (USSR) including Russia and Ukraine, (ii) countries of the former Yugoslavia including Croatia and Serbia, (iii) Czech Republic and Slovakia, and (iv) Ireland, Northern Ireland, and Great Britain (itself including three constituent footballing countries: England, Wales, and Scotland). Though possible due to multiple citizenship, it is extremely rare that players share both colonial and federal legacies. For these players, same colonial legacy subsumes same federal legacy.

Fourth, for language we rely on CEPII data as in Head and Mayer (2014) to ascertain whether or not two countries share one or more common languages. We assume that a player speaks (as mother tongues) the official and widely spoken languages of his country of citizenship at the beginning of his career. We consider some languages that are very close, even if not identical as one language (See Appendix C.1 for details). The fact that our language variable refers essentially to mother tongue implies that it should indeed be seen more as a cultural marker than as a means of communication. For many players (such as Argentinean and Spanish, Brazilian and Portuguese, or French and Senegalese players) same colonial legacy subsumes same language. For other players (such as Croatian and Serbian, Czech Slovakian or Irish and British players) same federal legacy subsumes same language. As a result, there is a small residual group of players that share the same language but neither colonial nor federal legacies.

Based on nationality, language, colonial legacy, and federal legacy, we define the following categories: 'same nationality' if two players share nationality (e.g. two Argentinian players); 'same colonial legacy' if they have different nationality, but same colonial legacy (e.g. Argentinian and Spanish); 'same federal legacy' if they have different nationality, but same federal legacy (e.g. Croatian and Serbian); 'same language' if they have different nationality, different colonial legacy and different federal legacy, but same language (e.g. Belgian and French); 'no shared culture' of they have different nationality, different colonial legacy, different federal legacy and different language (e.g. Argentinian and French).

More than a quarter of players have multiple citizenship. In such cases, if two players are citizens of the same country or of at least two different countries with common language, they are considered as speaking the same language. Analogously, if two players are citizens of the same country or of at least two different countries with common colonial (federal) legacy, they are considered as having the same colonial (federal) legacy. See Appendix C.1 for additional details.

In our dataset, 37.9% of the players have the same nationality, 9.5% have the same colonial legacy, 1.5% have the same federal legacy and 2.8% have the same language but different colonial legacy, different federal legacy and different nationality. We consider all these players as having the same culture. According to this definition, 50.6% of the players in our sample have the same culture, whereas 49.4% of them do not.

3.4. Pass dataset

The pass dataset contains aggregate information about passes between any two players at the half-season level. A pass includes any movement of the ball from one player to another (including free kicks, throw-ins, corner kicks and goal kicks). There are about 365 successful passes on average per game, which for two teams implies 730 passes per game, or about 8.1 passes per minute. We have 10.73 million passes in total.

In a game most players pass to each other, but with different frequency depending on their positions. On average the pass frequencies of outfield players between broad positions (defender, midfielder and forwards) are fairly balanced. The three highest frequencies are observed from defenders to midfielders (11.58%), from defenders to defenders (10.9%), and from midfielders to midfielders (8.86%). The lowest frequency is observed from goalkeepers to forwards (1.89%).

Passes are aggregated to average out match contingencies as prescribed by the model.[23] Aggregation is at the level of half-seasons. The partition in half-seasons is determined by the timing of the transfer windows, which are located between seasons (summer transfer window) and at the beginning of the calendar year (winter transfer window). It also splits the number of games during a season into two approximately equal parts: the number of games per team in a half-season ranges between 16 and 20 compared with the exact equal split of 17 for the German Bundesliga and 19 for the other top five leagues.

The dataset does not include pairs with zero pass count by design. We made a key assumption: if two players never pass to each other during a half-season, it must be that it is impossible for them to do so due to fielding or positioning reasons (e.g. the two players are only fielded to substitute each other as forwards), we drop the corresponding player pairs from the dataset. However, if we observe that a player passes to a given teammate but is never reciprocated, we keep the player pair. This implies that we have some zeros in the dataset recording the lack of passes from a player to a teammate from whom he nonetheless receives passes. Only 7.8% of the observations give rise to such zeros.

An alternative to half-seasons would be considering full seasons. However, half-seasons have advantages compared to seasons. The presence of the winter transfer window implies that during a season a team's squad may change composition. Our

[23] See Section 4, equation (2).

Table 2: Variable types - based on level of aggregation

player specific	player-pair specific	half-season specific	Example variables	N
yes	-	-	player height, year of birth, nationality	6,998
yes	-	yes	players age, value, team id, half-season id, experience with the team	37,026
-	yes	-	player-pair's shared nationality indicator	310,501
-	yes	yes	player-pair's number of passes in half-season, shared experience with club	669,022

Estimation dataset. N refers to the number of different values, ie there are 7 thousand different players and 669 thousand different passer-receiver pair observed in a half-seasons.

assumption of unchanged player quality makes more sense in a half-season than in a season, especially as younger players may evolve. The fact that half-seasons are separated by transfer windows allows us to cleanly map players' careers as they change teams, thereby combining player and pass information in a consistent way. Finally, considering a half-season allows us to investigate the role of common experience as players who spend more time together on the pitch may learn to pass more to each other.

3.5. Combined dataset

The final task to prepare our estimation dataset is to combine player information with pass information and obtain a relational dataset linked via player names as well as additional information.

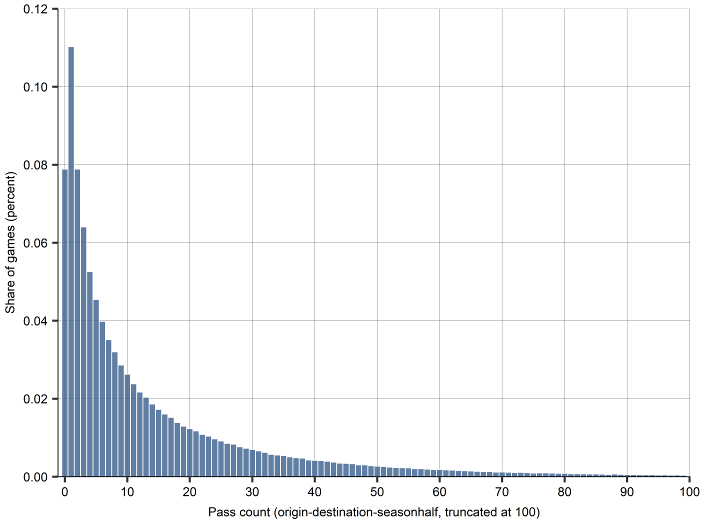
To match player and pass data, we had to identify players in both datasets and create a unique identifier for players. This process has proved to be a difficult task. First, there are players who are recorded differently across datasets - especially when their names have diacritical marks (such as 'é'), are translated from a non-Latin alphabet, or include many middle names. Second, different players may have the same name, especially in the case of frequent family names. To solve this issues, we developed a matching algorithm based on player names and additional information.[24] Variables are aggregated at different levels, as shown in Table 2.

The resulting estimation dataset is a directional pass dataset that, keeping track of who is the passer and who is the receiver, consists of 669,022 observations at the passer, receiver and half-season level. In a half-season, a player makes a total of 294 passes on average (ranging between 2 and 2166, with median equal to 228). On average, he

[24] The procedure is detailed in Appendix D where we also discuss a few decisions regarding data cleaning, such as dropping players who only have a single passing partner or those we could not identify. All results are robust to these decisions.

passes to 18.07 receivers (ranging from 2 to 35 with median equal to 19). The average pass count from passer to receiver is 15.92 (ranging from 0 to 488 with median equal to 8). The distribution looks highly skewed to the right as shown in Figure 2, where its support is truncated at 100 (98.63% of observations) for better visibility.

Figure 2: Distribution of passes



4. Separating Choice from Opportunities

A crucial challenge in assessing how common culture affects collaboration through passes arises from the conflation of choice and opportunity. As discussed in the introduction, individuals may collaborate more with similar others because they choose to do so ('choice homophily') or because collaboration with similar others is forced on them by circumstances ('induced homophily'). In this section we develop a discrete choice model to help us disentangle choice from opportunity in an internally consistent way by controlling for observable player characteristics (such as team, position, value, citizenship), hardly observable player characteristics (such as choice sets) or pass features (such as average distance).[25]

For now we are interested in assessing whether choice homophily actually exists while remaining agnostic about whether its existence represents an efficient outcome

[25] The existing literature on discrete choice models is extremely rich and covers several research fields. Two classic papers are McFadden and Train (2000) and Petrin and Train (2010). Two useful surveys are Anderson et al. (1992) and McFadden (2001). See Keane and Wolpin (2009), Todd and Wolpin (2010) and Keane et al. (2011) for surveys of applications of dynamic programming models of discrete choice in labor economics and other applied microeconomic fields.

promoting team performance, or rather a manifestation of inefficient in-group favoritism detrimental to the team. On the one hand, cultural similarity may enhance collaboration as long as culturally similar players may find it easier to anticipate each other's moves, may be more likely to trust each other to reciprocate, or may have better chances to understand each other by socializing off-pitch. On the other hand, choice homophily may be the result of players' overestimation of the abilities of similar others, or simply a fundamental irrational indication of in-group partiality. We will come back to this issue in Section 6, where we will discuss the mechanisms that could result in homophilous passing.

4.1. A Discrete Choice Model of Passing Behavior

A naive model of passing behavior would explain the number of passes between two players intuitively in terms of the time they are fielded together, the number of passes made by the passer to all teammates, the number of passes received by the receiver from all teammates, and the average bilateral distance between the two players on the pitch over a half season. The first three variables would be expected to positively affect the number of passes between the player pair, whereas the fourth would be expected to have a negative influence. Homophily would then act as a shifter leading to more passes between player pairs of similar culture after controlling for all the above variables.

While intuitively attractive, the naive approach may lead to biased estimates of homophily as it neglects the potential role of other player characteristics, in particular those that are hardly observable and can not be taken off the shelf to enrich the model's specification. Among these, the most important source of concern is arguably that any reasonable model of passing behavior should account for the alternatives the passer faces in terms of receivers, as the decision of passing to one of them is made after assessing the benefit of giving the ball to each of them. Analogously, one should also account for the alternatives the receiver faces in terms of passers as the passes he receives depends on which teammates are actually fielded.

A more structured approach can offer better guidance in this respect.[26] Consider a football team of $N = 11$ players, indexed from 1 to N, engaged in a half-season consisting of P team passes.[27] During the half season each player is assigned to a particular position on the pitch, which implies that a player's index identifies both his name and his position.[28] Focusing on two players, labeled o (a mnemonic for 'origin') and d (a mnemonic for 'destination'), a 'pass' from o to d is defined as a movement of the ball determined by a decision made by player o ('passer') to kick or throw the ball to teammate d ('receiver'). For $d = o$ the passer keeps possession of the ball.

[26] Here we offer a streamlined presentation. Additional details can be found in Appendix A.

[27] The model could be extended to allow for a squad of $N > 11$ players and different selections of players fielded during a half-season. Such extension, however, would not alter the model's insights informing our empirical analysis.

[28] In some matches a player may be fielded in a different position than the one he is typically assigned to, but such idiosyncratic events are averaged out over a half-season.

In the half season, let P^o be the number of passes made by player o to his teammates, P^d be the number of passes received by player d from his teammates, and $P^{o,d}$ be the number of passes made by player o to player d, such that we have $P^o = \sum_{d=1}^{N} P^{o,d}$, $P^d = \sum_{o=1}^{N} P^{o,d}$ and $P = \sum_{o=1}^{N} \sum_{d=1}^{N} P^{o,d}$. Moreover, let $T^{o,d}$ be the number of passes made by player o when player d is on the pitch (i.e. player d is in the passer's choice set) and $\tau^{o,d} \in [0,1]$ be the share of those passes made to player d, such that $T^{o,d} = P^o \tau^{o,d}$ holds. Finally, let $\pi^{o,d} \in [0,1]$ be the share of passes to player d in the total number of passes player o makes when teammate d is on the pitch, such that $P^{o,d} = T^{o,d} \pi^{o,d}$ holds. Based on these definition, we can express the number of passes made by o to d as:

$$P^{o,d} = P^o \tau^{o,d} \pi^{o,d} \tag{1}$$

We are interested in characterizing $\pi^{o,d}$ in terms of the probability that player o passes to player d rather than to any of the other nine teammates when player d is a viable option. We assume that player o wants to maximize team payoff and understands that the benefit for the team of one of its players controlling the ball is determined by the characteristics of that player, and by some randomness due to the vagaries of the game. A player's characteristics may include, for example, quality and experience as these affect what he can do with the ball. A game's vagaries may include, for instance, the performance of the opposing team, the referee's decisions or the weather conditions. We use U^d to denote the deterministic part of the team's benefit as determined by player d's characteristics, and z^d to denote the realization of its random part ('shock') due to match contingencies.

Player o also understands the challenges he faces in passing the ball to receiver d and we use $\widetilde{c}^{o,d}$ to denote the 'passing friction' capturing the cost associated with tackling those challenges. This cost may be objectively associated with the physical effort of passing or the mental effort of anticipating the receiver's moves. In this case, if the mental effort of passing to a teammate with similar cultural traits were lower, any choice homophily would be efficient for both the passer and the team as it would be due to objective constraints. Alternatively, the cost may subjectively derive from the passer's in-group favoritism. In this case, the passer's choice homophily would be efficient for him but inefficient for the team as his passes would deviate from what is objectively good for the latter.

Lastly, player o is aware of the difficulty receiver d may face in taking control of the ball, which depends on the receiver's circumstances. We use $\varphi^d \in [0,1]$ to denote the probability that receiver d takes control of the ball, and we call it the probability of a successful pass.

Apart from passing or keeping the ball, player o may decide to do something else with the ball generating team benefit U^o. For example, he may try to score a goal, or decide to kick the ball out of play to allow his team to reorganize.

We use $\beta \in [0,1]$ to denote the relative importance the team attaches to passing in general, independently of the specific passing episode. This is an important characteristic of the team's style of play. For example, low β would be associated with teams that try to score goals by quickly moving the ball into scoring range by long passes, through long balls or long air balls, whereas high β would refer to teams that prefer to play less quickly, using many short passes (also sideways or backwards) to find a weakness in the opposing team's tactics. Clearly, the weight given to U^o affects the passer's decision between passing or keeping the ball and doing something else with the ball, but it is immaterial for his choice among alternative receivers.

As a result, player o's passing decision is determined by the comparison of team utilities $U^o + \beta\varphi^d U^d - \widetilde{c}^{o,d} + z^d$ across all potential receivers $d = 1, ..., N$. However, given that z^d is a random shock, the outcome of this decision is an array of probabilities $\pi^{o,d}$ of passing to each potential receiver (including the passer himself). These probabilities can be readily characterized under appropriate assumptions on the probability distribution of the shocks. In particular, we make a customary assumption in the discrete choice literature that z^d is the realization of a random variable Z following a Gumbel distribution with zero mode and concentration around the mode positively related to $\kappa > 0$.[29] Zero mode implies that there is no systematic deviation from the deterministic part of the team's benefit across players' assessments of match contingencies. As all players share the same κ, this is a team characteristic: players are trained to assess match contingencies in a common way. Larger κ can then interpreted as resulting from more intense training to reduce variation in their individual assessments.

Under the chosen distributional assumption, the model predicts that the probability that player o passes to player d when the latter is on the pitch evaluates to

$$\pi^{o,d} = \frac{\left(c^{o,d}\right)^{-\kappa} P^d}{\left(\Lambda^o\right)^{\kappa} \left(\Lambda^d\right)^{\kappa}} \tag{2}$$

so that, by (1), the number of passes made by player o to player d in a half-season is

$$P^{o,d} = \frac{P^o}{\left(\Lambda^o\right)^{\kappa}} \left(c^{o,d}\right)^{-\kappa} \tau^{o,d} \frac{P^d}{\left(\Lambda^d\right)^{\kappa}} \tag{3}$$

with definitions

$$\Lambda^o = \left[\sum_{d=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^d}{\left(\Lambda^d\right)^{\kappa}}\right]^{\frac{1}{\kappa}} \text{ and } \Lambda^d = \left[\sum_{o=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}\right]^{\frac{1}{\kappa}}$$

for $c^{o,d} = \exp \widetilde{c}^{o,d}$.

Expression (3) has clear implications once Λ^o and Λ^d are given intuitive interpretations. The former is a weighted geometric average of the passes received by all players

[29] The cumulative density function of this Gumbel (or Type-I Extreme Value) distribution is $G(z) = \exp\left(-\exp(-\kappa z)\right)$ for $z \in (-\infty, +\infty)$.

in the team with weights determined by their bilateral passing frictions with respect to passer o. It is, therefore, a compact index that measures the passer's multilateral access to teammates considering both the overall number of passes they attract and the difficulty to reach them. Given that, in this optimizing framework, the number of passes a player attracts depends on the benefit he generates for the team also through his own future passes, Λ^o captures the average team benefit of passer o's forward-looking options. Analogously, Λ^d is a weighted geometric average of the passes made by all players in the team with weights determined by their bilateral passing frictions with respect to receiver d. It is thus a compact index measuring the receiver's multilateral accessibility by teammates considering both their characteristics and the difficulty to be reached by them. As the number of passes a player makes depends on the benefit he generates for the team, Λ^d captures the forward-looking average team benefit of all passes player d may receive.

The first implication from (3) is that the number of passes $P^{o,d}$ from player o to teammate d increases with the number of passes $T^{o,d} = P^o \tau^{o,d}$ made by player o when player d is on the pitch. This is outside players' control as it reflects their coach's selection decisions. The second implication is that the number of passes $P^{o,d}$ from player o to teammate d increases with the overall number of passes P^o made by the former and the overall number of passes P^d received by the latter, where the number of passes a player makes or receives depends on the benefit he generates for the team. Its third implication is that the number of passes from player o to teammate d decreases with the passing friction $c^{o,d}$ between them. While these three implications are in common with the naive approach, a fourth crucial implication is unique to the structural approach: the number of passes from player o to teammate d is a decreasing function of the average team benefit of passer o's options (Λ^o) and the average team benefit of passes player d receives (Λ^d). In other words, there are more passes between any two players when these have less attractive alternatives to pass or receive, which depends on the characteristics (including the positions) of all teammates on the pitch. Neglecting these multilateral variables, as the naive approach does, would lead to biased homophily estimation.[30]

The distinction between distance and culture related challenges in passing from player o to player d embedded in $c^{o,d}$ can be made explicit by specifying the bilateral passing friction multiplicatively as

$$c^{o,d} = e^{\varepsilon^{o,d}} \left(g^{o,d}\right)^{\gamma} \left(l^{o,d}\right)^{\lambda} \tag{4}$$

where $\varepsilon^{o,d}$ is an error term allowing for imprecise measurement of the bilateral friction. In this expression $g^{o,d}$ is the physical distance between the two players' positions so

[30] Another unique implication of the structural approach is that the rate at which the number of passes from player o to teammate d decreases with the bilateral passing friction is higher for larger κ: when the variation in players' assessments of match contingencies is smaller, differences in bilateral frictions become more salient. We do not explore this interesting implication as it goes beyond the scope of the present paper.

that $\left(g^{o,d}\right)^{\gamma}$ captures all distance-related frictions that make it hard to pass the ball from passer o to receiver d independently of their identities. The term $\left(l^{o,d}\right)^{\lambda}$ captures, instead, all non-distance-related frictions that make it hard to pass the ball from passer o to receiver d independently of their positions. These may include, for instance, limited experience in playing together but, crucially, also different cultural traits.

We are now ready to translate (3) into an estimating equation. Specifically, we define the half-season 'pass rate' $p^{o.d}$ as the ratio $P^{o,d}/P$ of the number of passes from player o to teammate d over the total number of team passes. Then, substituting (4) into (3) and taking logs gives

$$\log p^{o,d} = \log \tau^{o,d} + \log P^{o} \left(\Lambda^{o}\right)^{-\kappa} + \ln P^{d} \left(\Lambda^{d}\right)^{-\kappa} - \kappa\gamma \log g^{o,d} - \kappa\lambda \log l^{o,d} - \log P + \varepsilon^{o,d} \quad (5)$$

which we will use in the next sections as the theoretical basis to empirically investigate the relation between the pass rate $p^{o,d}$ and the cultural dimensions of $l^{o,d}$. Before proceeding, three remarks are in order. First, equation (5) distinguishes the role of homophilous preferences ('passing to teammates'), which work through $\log l^{o,d}$, from the implications of homophilous meeting rates ('passing to teammates through teammates'), which work through the forward-looking terms Λ^{o} and Λ^{d}. Second, such distinction allows us to argue that in (5) the cultural dimensions of $l^{o,d}$ determine choice homophily, while induced homophily is determined by all other terms on the right hand side of (5). Third, the pass rate in equation (5) is conditional on players being together on the pitch, and homophily may play a role in the selection of fielded players. As long as this affects induced homophily, accounting for $\tau^{o,d}$ allows us to net it out.

4.2. Empirical Implementation of the Passing Model

The empirical implementation of our theoretical model requires a generalized linear estimator with a log link function. In such setup there is a rich literature on the benefits of using a Poisson model rather than a log(count) model with a large number of fixed effects. The former approach is in line with best practices in the estimation of gravity equations through fixed effects Poisson Pseudo Maximum Likelihood (FE-PPML) in international trade (see e.g. Fally (2015) and Santos-Silva and Tenreyro (2022)).[31] Nonetheless, we will also provide robustness checks with a log(count) model.

We map our theoretical model into a Poisson model as follows. We use the number of passes from player o to player d in period t as dependent variable. We call it $pass\_count_{o,d,t}$, which corresponds to $P^{o,d}$ in the theoretical model. The total number of

[31] Specifically, we follow the procedure described in Berge (2018). As discussed by Hinz et al. (2021), a drawback of fixed effect models in general is the incidental parameter bias: having several nuisance parameters to estimate, the estimated coefficient of the variable of interest may be biased. FE-PPML estimates can deal with this type of bias better than non-linear OLS (Santos-Silva and Tenreyro, 2022). While Weidner and Zylkin (2021) show that the Poisson model still leaves some room for potential bias, with no double player fixed effects and a large number of observations the bias should be small in our case.

passes by their team in a half-season (P in the theoretical model) is absorbed through team by half-season fixed effects.

We capture distance-related frictions that make it difficult to pass the ball from o to d (i.e. $\left(g^{o,d}\right)^{\gamma}$ in the theoretical model) by constructing the following measure:

$$PassFric_{o,d,t} = \gamma_1 PassDist_{o,d,t} + \gamma_2 Forwardness_{o,d,t} + \eta Position_o Position_d$$

where, on average in a half-season, $PassDist_{o,d,t}$ is the distance of passes between the two players, $Forwardness_{o,d,t}$ is the share of passes between the two players with a forward direction, and $Position_o Position_d$ is a dummy variable capturing the two players' broad positions (such as defender, midfielder and forward). As we acknowledge that $PassDist_{o,d,t}$ and $Forwardness_{o,d,t}$ may actually be a mechanism rather than a confounder, we will show results with and without them.

As for the cultural aspect of non-distance-related frictions that make it difficult to pass the ball from passer o to receiver d independently of their positions, we measure cultural similarity through the time-invariant variable $SameCult_{o,d}$, which combines the categories described in Section 3.3: it takes value 1 if two players share either nationality, colonial legacy, federal legacy or just language, and value 0 otherwise.

In light of the previous discussion, we estimate two different passing models, all by Poisson regressions with standard errors clustered at the passer by half-season and receiver by half-season level. The first is a naive model explaining the number of passes between two players in terms of the number of passes made by the passer to all teammates ($P_{o,t}$), the number of passes received by the receiver from all teammates ($R_{d,t}$), bilateral distance-related frictions ($PassFric_{o,d,t}$), and cultural similarity ($SameCult_{o,d}$):

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + PassFric_{o,d,t} + \ln P_{o,d,t} + \ln P_{o,t} + \ln R_{d,t} + \phi_{c,t} \tag{6}$$

where $\phi_{c,t}$ is a team by half-season dummy and the exposure variable $P_{o,t}$ is handled as an offset variable by forcing its coefficient to be equal to 1. A positive estimate for the coefficient δ of the same culture indicator $SameCult_{o,d}$ would reveal the presence of choice homophily as it would imply that, after partialling out distance-related frictions, relative to overall passing patterns players with same culture pass more to each other than to players of different culture. Accordingly, we call the estimated δ the 'homophily premium'. Comparing estimates obtained with or without the inclusion of the number of passes made by player o when player d is on the pitch will allow us to tell how much of the premium is mediated by coach decisions.

The second passing model we estimate implements equation (5) of the discrete choice model. For the reasons detailed above, this is our preferred specification:

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + PassFric_{o,d,t} + \ln \tau_{o,d,t} + \upsilon_{o,t} + \upsilon_{d,t}) \tag{7}$$

where $\upsilon_{o,t}$ and $\upsilon_{d,t}$ are player by half-season fixed effects while $\ln \tau_{o,d,t}$ is an exposure variable corresponding to $\tau^{o,d} = T^{o,d}/P^o$ in (5) and is defined as the (log) ratio of the number of passes made by player o when player d is on the pitch to player o's total passes. The exposure variable is handled as an offset variable, forcing its coefficient to be equal to 1.[32] Again, comparing estimates obtained with or without the inclusion of the share of passes made by player o when player d is on the pitch will allow us to tell how much of the premium is mediated by coach decisions.

It is important to highlight that the player fixed effects absorb also the additional constraints imposed by team composition and all teammates' positions through Λ^o and Λ^d. As we have seen, these terms account for the alternatives the passer faces in terms of receivers and the receiver faces in terms of passers respectively: neglecting them may lead to biased estimation due to omitted variables.[33]

Note that, whereas the discrete choice model model is written for a single team and a single half-season, its empirical implementation is estimated on the pooled dataset. However, as we have passer by half-season fixed effects together with receiver by half-season fixed effects, the estimated impact of homophily is very close to the mean estimate over the (1484) team by half-season regressions we could alternatively run. Without additional controls, using OLS and weighting by pass count, the two estimates would be exactly the same. The advantage of relying on a single regression is that we can better estimate the role played by the additional controls and have a single standard error.

5. Homophily in Collaboration

We are now ready to report and discuss our empirical findings based the estimation of regressions (6) and (7).

5.1. Results

Table 3 presents the results from regressions (6) and (7) in columns (1) and (2) respectively. In these columns, by forcing the coefficient of $P_{o,d}$ and $\tau_{o,d}$ to equal 1, the effect of culture is estimated for the number of passes from the passer to the receiver relative to the former's total number of passes when both players are fielded together. In both models, team by half-season fixed effects absorb the total number of team passes. They also absorb team and team by half-season characteristics such as club history or current management. Cross-position dummies capture the relative roles of players.

Column (1) supports the naive prediction that the pass count between two players is positively related with their total passes made or received, and negatively related

[32] We cannot have passer by receiver fixed effect as sometimes used in the gravity equation literature because our variable of interest is time-invariant.

[33] In the gravity literature the analogous constraints are embedded in the so-called 'multilateral resistance terms'. As shown by Fally (2015), when the FE-PPML estimator is used, these constraints are automatically satisfied thanks to origin and destination fixed effects.

Table 3: Baseline results

	pass_count	
	(1)	(2)
Same culture (any) (0/1)	0.0131^{***}	0.0242^{***}
	(0.0020)	(0.0025)
Passer total passes (ln)	0.9143^{***}	
	(0.0011)	
Receiver total pass received (ln)	0.3022^{***}	
	(0.0034)	
Average length of passes (ln)	-0.6580^{***}	-0.7788^{***}
	(0.0052)	(0.0053)
Average forwardness Ind (0-1)	0.0743^{***}	0.0813^{***}
	(0.0049)	(0.0061)
Observations	669,022	668,105
Pseudo R^2	0.70850	0.76077
Team by half-season dummies	✓	
Cross-position dummies	✓	✓
Passer by half-season dummies		✓
Receiver by half-season dummies		✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Exposure variable is $\tau_{o,d}$, the sum of passes by passer when receiver is also on the pitch divided by total sum of passes by the passer. Total team pass count is captured via team by half-season fixed effects.

with pass distance. Passing pairs who move the ball forward tend to pass more as well. As for the homophily premium, relative to overall passing patterns players with same culture pass 1.31% more to each other than to players of different culture. Column (2) replaces total passes made and total passes received by a player pair with time-varying passer by half-season and receiver by half-season fixed effects. This still allows player characteristics to change over time, and implies that the estimated coefficient of same culture is close to what would be the average of coefficients if estimated one by one for teams and time periods. The estimated homophily premium is 2.42%. To interpret this coefficient, consider the passes made by a player in a half-season, conditioning on constant and time-varying receiver characteristics, passer-receiver position pair and other pass features. This player is expected to pass 2.42% more to teammates of same culture than to teammates of different culture. This is our preferred estimate of the 'homophily premium'.

To put the estimated choice homophily into context, we estimate how homophily affects passes without distinguishing between its choice and induced aspects. We do so by estimating the effect of $SameCult_{o,d}$ as an unconditional average difference within team and half-season in the following Poisson model:

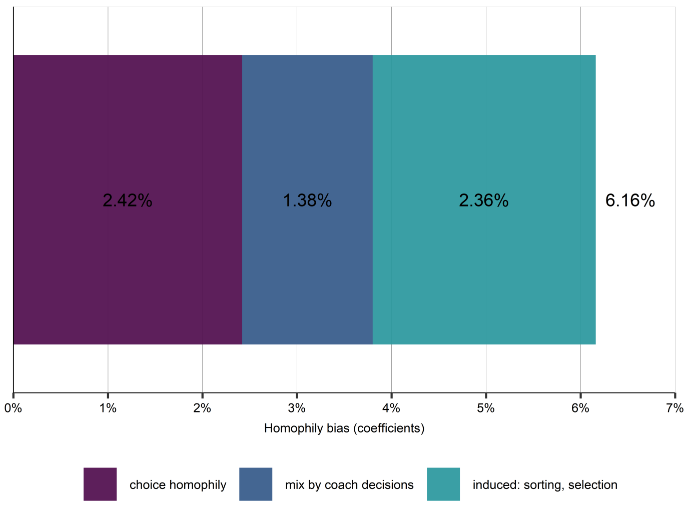
$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + \nu_{team,t}) \tag{8}$$

where $\nu_{team,t}$ is a team by half-season fixed effect.

The corresponding results are reported in Column (1) of Table 4, where, for ease of comparison, Column (3) recalls the baseline results from Column (2) of Table 3. Column (1) offers clear unconditional evidence of homophily: players with same culture tend to pass 6.16% more to each other than to players with different culture. Columns (2) and (3) introduce the full set of player fixed effects and passes controls with an exception. Specifically, while Column (3) controls also for the time a player pair spends together on the pitch in a half-season, Column (2) does not.

Comparing the three columns of Table 4 suggests that 2.36% of the overall homophily premium of 6.16% vanishes when controlling for player and pass characteristics, and a further 1.38% disappears when one considers players' time spent together on pitch. This further reduction reveals the role of managers' decisions in allowing the team to internalize the effects of homophily. As discussed in Section 2, endogenous team formation may mitigate the effects of homophily on collaboration, which is captured by the time a player pair spends together on the pitch. For example, as the manager observes his players in training, he may decide to field same culture players in a game because he sees them collaborating more. In this case, the manager acts as a mediator allowing the team to internalize the effects of homophily. Hence, the effects of homophily on collaboration we estimate can be seen as a lower bound estimate with respect to what would be found in a randomly composed team. In other words, as discussed in Section 2, endogenous team formation may mitigate the effects of homophily on collaboration, which is captured by $\tau_{o,d}$.

Figure 3: Dissecting total homophily bias



Note: Coefficient values from fixed effects regression from Table 4.

Figure 3 summarizes how the overall homophily premium of 6.16% can be decomposed in a choice homophily premium of 2.42%, a mitigation premium of 1.38% due to endogenous managerial decisions, and an induced homophily premium of 2.36% due to player and pass characteristics.

Table 4: From total homophily to choice homophily

	pass count		
	(1)	(2)	(3)
Same culture (any) (0/1)	0.0616***	0.0380***	0.0242***
	(0.0090)	(0.0052)	(0.0041)
Observations	669,022	668,105	668,105
Pseudo R^2	0.07813	0.67154	0.75929
Minutes shared together			✓
Pass features		✓	✓
Cross-position dummies	✓	✓	✓
Team by half-season fixed effects	✓		
Passer by half-season fixed effects		✓	✓
Receiver by half-season fixed effects		✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Column (2) and (3) includes ln(average pass distance) and forwardness index. Total team pass count is captured via team by half-season fixed effects.

5.2. Discussion

According to our estimated 'homophily premium', conditioning on constant and time-varying receiver characteristics, passer-receiver position pair and other pass features, a player is expected to pass 2.42% more to teammates of same culture. We now discuss the extent to which this finding might be affected by model specification, omitted variables, or the operalization of the notion of culture. We then comment on the importance of the premium for footballing outcomes.

5.2.1. Alternative Specifications

As already discussed, fixed effects PPML estimates can deal with the incidental parameter bias of fixed effects models better than non-linear OLS. It is, however, interesting to compare those estimates with the fixed effects OLS estimates of a log count model. This is done in Table 5, which shows that the homophily premium obtained using fixed effect OLS (2.22% in column (4)) is quite similar to the one relying on fixed effect PPML (2.42% in column (1).

In column (2) we also check whether our PPML results are driven by peculiar cases by excluding observations when: a player passes to a teammate but is never reciprocated in a half-season; two players are fielded less than 45 minutes together in a half-season; and either the passer or the receiver is a goalkeeper (as goalkeepers can pass, but their choice set is typically more limited). While the number of observations is reduced by

Table 5: Robustness I: Alternative specification

		pass count		ln(pass per min)
	(1)	(2)	(3)	(4)
	Poisson	Poisson	Poisson	OLS
Same culture (any) (0/1)	0.0242***	0.0225***	0.0237***	0.0222***
	(0.0025)	(0.0025)	(0.0027)	(0.0030)
Average length of passes (ln)	-0.7788***	-0.7861***	-0.8143***	-0.3918***
	(0.0053)	(0.0051)	(0.0062)	(0.0046)
Average forwardness Ind (0-1)	0.0813***	-0.0026	0.2868***	0.3089***
	(0.0061)	(0.0062)	(0.0073)	(0.0046)
Passer total passes when together		0.1403***		0.2838***
		(0.0029)		(0.0030)
Observations	668,105	668,105	432,125	666,230
Pseudo R^2 /R^2	0.76077	0.76038	0.71358	0.26120
Passer by half-season fixed effects	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓
Cross-position dummies	✓	✓	✓	✓

Column 1-3 Poisson, column 4 OLS regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Total team pass count is captured via team by half-season fixed effects.

36%, the point estimate for the homophily premium is essentially unchanged (2.37% vs. 2.42%).

Finally, in column (3), we enrich the PPML specification by adding the total number of passes when both players are on the pitch, that is, we do not restrict the exposure coefficient of $ln\tau_{o,d,t}$ to be equal to one. The fact that the resulting estimate is positive reveals a more-than-proportional effect of total passes on the bilateral pass count. Nonetheless, the coefficient estimate for the homophily premium hardly changes (2.25% vs. 2.42%).

5.2.2. Potential Confounders

Identification of choice homophily according to our definition of culture assumes orthogonality with respect to other possible dimensions of passing assortativity, but some of these dimensions might actually overlap with culture. To check whether this is indeed the case, we extend the fixed effects PPML specification (7) by including a battery of player characteristics other than culture that could arguably foster reciprocal passes between players: shared experience, quality, physical attributes, regulations, national styles, and common values. Each of them could be correlated with nationality, thereby inducing collaboration between same nationals and thus confounding our estimated homophily. The corresponding results are reported in Table 6, where column

(1) recalls the baseline homophily premium estimate for ease of comparison.

Shared experience. A player pair may collaborate more today because of previous interactions. If, for instance, these interactions were more likely between players from the same country because they grew up there, then our results would not reflect an actual relationship between same culture and collaboration. To check whether this matters, column (2) in Table 6 includes shared experience through two binary variables. Two players may pass more in the current team because they have been playing there together for a while or because they have played together somewhere else in the past. Accordingly, the first binary variable equals one if the passer and the receiver have spent at least a half-season at a club together, and zero otherwise.[34] The second binary variable equals one if the passer and the receiver played together earlier in a different team. For the latter variable, we collect team history for all the players in our sample.[35] For each passing pairs, we then compute how many half-seasons they played together before playing for their current team, including possible spells together in youth teams. Any shared past experience is actually rare: only 3% of player pairs have ever played together. While the point estimates for both variables are different from zero, neither influences the estimate of homophily. Interestingly, while the coefficient for shared experience at the current club is positive, the coefficient for shared past experience is negative.

Quality, physical attributes, regulations. Two players may collaborate more because of similar quality or physical attributes. That would happen, for instance, if players of similar quality passed more to each other and players of given quality came from the same country. In the same vein, height could be typical of players from a certain country. If long legs favored long kicks and long runs, tall players would pass more to each other due to physical complementarity. Moreover, assortativity may be induced by regulatory constraints on players' countries of origin, making it more likely for better or taller players to hail from the same country. In all these cases, our results would not reflect an actual relationship between same culture and collaboration. To check whether this is the case, column (3) in Table 6 considers assortativity in terms of quality, physical attributes, and regulations. First, we add a variable measuring the difference in the (log) of players' values. Second, as we have data on the height of all players, we control for the absolute height difference (in cm) between passer and receiver. Third, we condition on a regulatory aspect that restricts the fielding of players from non-EU countries with league-specific exemptions. Combing through these, we add a binary variable equal to one if passer and receiver are from the European Union or other exempted countries, and zero otherwise.[36] Column (3) shows that all these variables have only minor effects

[34] Results are robust to using longer period, or the log number of days spent together instead.

[35] Players are typically first observed around the age 14, when they get a semi-professional contract with a youth team.

[36] For details, see Appendix B: 20% of player pairs have at least one restricted player.

Table 6: Robustness II: Potential confounders

	pass count			
	(1)	(2)	(3)	(4)
Same culture (any) (0/1)	0.0238***	0.0248***	0.0214***	0.0247***
	(0.0028)	(0.0026)	(0.0026)	(0.0025)
Average length of passes (ln)		-0.7943***	-0.7823***	-0.7631***
		(0.0052)	(0.0052)	(0.0052)
Average forwardness Ind (0-1)		0.0142**	0.0113*	-0.1145***
		(0.0063)	(0.0062)	(0.0062)
Shared experience, 1sh+ (0/1)		0.0117**	0.0110*	
		(0.0056)	(0.0056)	
Experience w/ other team (0/1)		-0.0153***	-0.0165***	
		(0.0044)	(0.0043)	
Height difference (cm)			-0.0126***	
			(0.0002)	
Players value difference, (dlog)			-0.0008***	
			(0.0002)	
Both treated as EU player (0/1)			0.0107	
			(0.0080)	
Observations	668,105	668,105	668,105	668,096
Pseudo R^2	0.74289	0.75930	0.76074	0.75731
Passer by half-season fixed effects	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓
Cross-position dummies	✓	✓	✓	
Cross-position dummies by nationality				✓

Column 1-3 Poisson, column 4 OLS regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Total team pass count is captured via team by half-season fixed effects. Both players treated as EU+ reflect national regulations to play, see Appendix B.4. Experience with other team – different team prior to current team including youth teams. Difference in the (log) of players' values, absolute height difference (in cm) between passer and receiver.

on the point estimate of the homophily premium.

National style. Another reason why players with same nationality pass more to one another may be related to the fact that countries differ in terms of their style of football and this style is transmitted to local players at an early career stage. For instance, some countries may traditionally favor a fast 'vertical' style that tries to score goals by quickly moving the ball into scoring range through long forward passes, while others may prefer a slow 'horizontal' style, playing less quickly with short passes to find a weakness in the opposing team's tactics. Shared national style would then make it more likely for players from the same country to pass more to one another other without necessarily reflecting an actual relationship between same culture and collaboration.[37]

To deal with this possible confounder, in column (4) of Table 6 we introduce national styles as follows. Considering three aggregate positions (defender, midfielder, forward), for a country we look at its players in each position as an origin and compute all the passes they make to teammates in each position as a destination. We then calculate two national metrics: the average number of passes over ninety minutes ('pass intensity') and the frequency of passes from defenders to forwards divided by the number of passes by midfielders to other midfielders ('pass verticality'). Finally, we use a combination of the two metrics to define a country's national style. Figure 4 shows a large variation in national styles with players from different countries actually favoring different combinations of intensity and verticality.

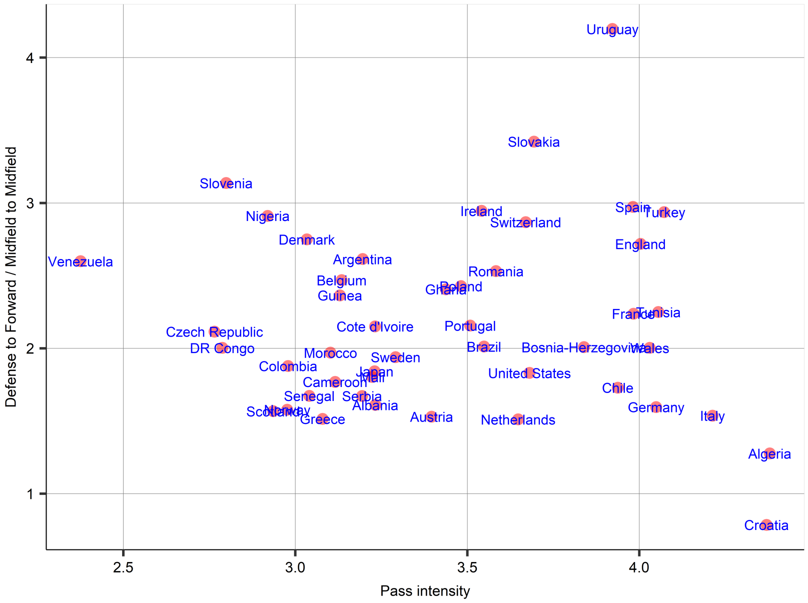
Turning to regressions, national pass intensity is already captured by player fixed effects. To partial out the national pass verticality (and any other potential defining features of national style), column (4) interacts the position by position set of dummies with the (first) nationality of players, which allows, say, Spanish defenders and midfielders to have a different pass intensity than Croatian defenders and midfielders. This interaction, however, does not affect our results as the point estimate for the coefficient of interest hardly changes (2.41% vs. 2.42%).

5.2.3. Operalization of Culture

In our regressions two players belong to the same cultural group if they share either nationality, colonial legacy, federal legacy or language. The underlying idea is that, while nationals of the same country are more likely to share a common heritage, common language and shared history due to bygone political ties may still allow individuals of different nationality to share some aspects of common heritage. A possible source of concern is, however, that treating (as we do) all cultural groups as equally different from one another may be someway too coarse. In the literature on culture it has been indeed

[37] In the same vein, an additional confounder could be national specialization in positions. For example, if all Italians were defenders, all Spaniards were midfielders and all Brazilians were forwards, and passes within position were more frequent, the patterns of national specialization would induce collaboration between same nationals and bias estimated homophily. In Table 6 player by half-season fixed effects and nationality by cross-position dummies take care of this possible concern.

Figure 4: Simple comparison of national style



Note: Countries with at least 25 players observed. Averages across players, unweighted by passes.

argued that there are dimensions along which some cultural groups can be considered more closely related to one another other than to other cultural groups.[38]

We address this source of concern in four ways. First, we disentangle the different components of culture to allow for asymmetries among cultural groups depending on whether they share nationality, colonial legacy without nationality, federal legacy without nationality, or just language without nationality or any historical legacy. Specifically, recall that our measure of cultural affinity $SameCult_{o,d}$ takes value 1 if a passer o and a receiver d share either nationality, colonial legacy, federal legacy or language, and 0 otherwise. We now define the following variables corresponding to the categories described in Section 3.3: $SameNat_{o,d} = 1$ if o and d have the same nationality, and $SameNat_{o,d} = 0$ if they have different nationality; $SameCol_{o,d} = 1$ if o and d have the same colonial legacy but different nationality, and $SameCol_{o,d} = 0$ if they have different colonial legacy and different nationality; $SameFed_{o,d} = 1$ if o and d have the same federal legacy but different nationality, and $SameFed_{o,d} = 0$ if they have different federal legacy and different nationality; $SameLan_{o,d} = 1$ if o and d have the same language but different colonial legacy, different federal legacy and different nationality; $SameLan_{o,d} = 0$ if they have different language as well as different colonial, different federal legacies and different nationality. The benchmark again consists of

[38] See, e.g., Melitz and Toubal (2014) and Desmet and Ortuño-Ortín (2017).

pairs with different nationality, different colonial legacy, different federal legacy, and different language. We then reestimate the fixed effects model (7) disentangling the four components of $SameCult_{o,d}$ as follows:

$$E(pass\_count_{o,d,t}|...) = exp(\delta_1 SameNat_{o,d} + \delta_2 SameCol_{o,d} + \delta_3 SameFed_{o,d} \\ + \delta_4 SameLan_{o,d} + PassFric_{o,d,t} + \ln \tau_{o,d,t} + \upsilon_{o,t} + \upsilon_{d,t}) \quad (9)$$

where estimated δ's larger than zero would again reveal the presence of choice homophily based on the corresponding cultural aspects.

Second, we look at linguistic proximity, relying on the CEPII dataset to create a similar language indicator variable based on the common language index developed by Melitz and Toubal (2014).[39] This indicator equals 1 if language similarity is above the median value 0.5 (as, e.g., for Italy and Spain, Denmark and Sweden, Croatia and Bulgaria) and 0 otherwise. According to the indicator, in our data 6.7% of player pair by half-season observations can be classified as speaking similar languages despite their different nationality, different colonial legacy, different federal legacy and different language.

Third, we consider geographical proximity as measured by another indicator variable, which equals 1 if two countries share a land border and 0 otherwise. In our data 7.5% of observations have a shared border despite their different nationality, different colonial legacy, different federal legacy and different or not even similar language.

Finally, another reason why some cultural groups might be considered more closely related is that their values are more aligned. We look into this issue by using the combined Wave7 of the World Value Survey (WVS) + European Values Survey (EVS) exercise to create a simple distance metric between countries. Specifically, following Spolaore and Wacziarg (2016) and Desmet and Ortuño-Ortín (2017), we assess the similarity of countries' answers to 72 selected questions. Two countries are considered to have similar values if similar shares of their populations give the same answers to questions concerning religion, sexuality, trust in institutions, politics, membership in organizations, the role of work or family. Distance between countries is computed as the average Euclidean distance between the shares of nationals giving the various admissible answers.[40] The distance ranges between 0 when two countries are perfectly aligned (i.e. they exhibit the same distribution of answers across all questions) and $\sqrt{2}$ when they are fully misaligned. In our dataset the mean and median distances are 0.25 and 0.17, while the minimum and the maximum are 0.08 and 0.56. To check the importance of values, we create a binary variable for similar and dissimilar countries, defined as those below and above the median distance respectively. Finally, we cut the sample in three categories, depending on whether player pairs exhibit same nationality, similar values,

[39] See the "Common language Index based on the level specification" in Melitz and Toubal (2014) at a bilateral country level (mean=0.13).

[40] We treat this distance as constant in our period of observation. Details are discussed in Appendix C.3.

and dissimilar values (base).

Results are reported in Table 7. Column (1) shows that different cultural components have different effects on homophily. Compared to pairs not sharing any aspect of culture, we find a homophily premium of 2.84% for same nationality as well as for same colonial legacy without same nationality. Same colonial legacy is, therefore, as consequential as same nationality. Interestingly, we also find a negative homophily premium for player pairs from formerly federated countries (such as countries former Soviet Union or Yugoslavia).[41]. In contrast, we find no correlation for the relatively few cases of having just shared language without same colonial legacy or same nationality (such as Austria and Germany).

In column (2) we add linguistic proximity, partitioning players into four categories: same nationality, same language but different nationality, similar language, and dissimilar language (base). Compared with pairs who speak different and dissimilar languages (such as Russian and English), players of shared nationality (and language) pass 3.02% more; those with same language but different nationality (such as Spanish and Argentinian, Austrian and German or Irish and English) pass 1.56% more; pairs with similar rather than same language (such as Dutch and German) pass 1.11% more. In column (3), we consider geographical proximity (being neighbors) on top of language variables. We find results with respect to language proximity broadly in line with those in column (2), only with slightly smaller point estimates. However, for countries with the same (or similar) language that are also neighbors, coefficients should be added up. Geographical proximity alone contributes a little to passing. Lastly, column (4) shows a point estimate for similar values that is actually negative: player pairs from countries with similar values seem to pass less.

To summarize, beyond nationality, we find choice homophily to be about similar language (interpreted as a cultural marker rather than a means of communication) and shared history (especially colonial ties) and unlikely to be driven by shared values.

5.2.4. Importance of Homophily

Beyond statistical significance, another interesting question to ask about our findings is whether the estimated homophily premium makes a difference in terms of footballing outcomes. This is a hard question to answer directly if one has in mind the overall impact of homophily on a team's performance as measured, for instance, by its end-of-season position in the league standings. The reason is that, to answer the question, one would have to estimate not only by how much choice homophily fosters bilateral passes as we do ('pass creation'), but also by how much it diverts passes from alternative options ('pass diversion'), and whether the resulting reallocation of passes fosters or hampers team performance. Only by netting out the opposite effects of pass creation and pass diversion, it would then be possible to assess by how much a team's performance is affected by homophily thorough passes. The quantification of these effects

[41] See Appendix C.1 for details

Table 7: Dissecting culture

	pass_count			
	(1)	(2)	(3)	(4)
Same nationality (0/1)	0.0284***	0.0302***	0.0315***	0.0183***
	(0.0030)	(0.0031)	(0.0031)	(0.0031)
Same colonial legacy (0/1)	0.0284***			
	(0.0041)			
Same federal legacy (0/1)	-0.0223**			
	(0.0106)			
Just shared language (0/1)	-0.0046			
	(0.0070)			
LC: diff country, same language (0/1)		0.0156***	0.0140***	
		(0.0039)	(0.0040)	
LC: diff country, similar language (0/1)		0.0111**	0.0094*	
		(0.0044)	(0.0045)	
Geographical proximity (neighbors) (0/1)			0.0064*	
			(0.0031)	
WVS: similar values (0/1)				-0.0077***
				(0.0025)
Observations	668,105	668,105	668,105	668,105
Pseudo R^2	0.76078	0.76077	0.76077	0.76076
Passer by half-season fixed effects	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓
Cross-position dummies	✓	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Column 1: Same cultural background is separated into aspects (language, colonial, federal legacy or nationality. Because of top-coding "Just shared language" means no colonial ties, just language such as Belgium and France. Column 2: keeps same nationality but and adds linguistic similarity categories as per the 'LC' index of Melitz and Toubal (2014) (base is different country and not similar language), Column 3 adds geographic proximity. Column 4 has value similarity based on the 2021 WVS/EWS, cut at median (base = different values). *tau* is the sum of passes by passer when receiver is also on the pitch divided by total sum of passes by the passer. All models include pass distance and the forwardness index. Total team pass count is captured via team by half-season fixed effects.

would require a ‘partial equilibrium’ model of team performance. Moreover, to map differential passes into a team’s rank in the league standings, one would also have to deal with the fact that points gained or lost by a team are lost or gained by its competitors. The quantification of these aggregate effects would call for a ‘general equilibrium’ model of league competition. While this is a very interesting direction of research, it clearly goes beyond the scope of the present paper.

That said, it could still be useful to provide some benchmark against which to assess, at least indirectly, the quantitative relevance of the estimated homophily premium for team performance. We do so in two ways, providing back-of-the-envelope calculations based on the estimation of the correlation of points gained with passes and the calculation of a player value equivalent of the homophily premium.

To investigate the correlation between passes and points, we aggregate our dataset at the level of teams and half-seasons. We end up with N=1’568 observations, given that we have 16 half-seasons, 20 teams for England, France, Italy and Spain, and 18 teams for Germany. We then estimate the relation between the average number of points gained by teams and their (log) average pass count per game in a half-season.[42] In particular, we run two types of regressions. In the first, we estimate the cross-section correlation for a single half-season (specifically, the first half of season 2015-16) including league dummies. In the second, we perform a panel estimation with league-half-season and team fixed effects. Table 8 presents the cross-section OLS results in column (1) and the panel results in columns (2), (3) and (4). While in columns (1) and (2) the dependent variable is points per game, in columns (3) and (4) it is log(points per game) for easier interpretation. Column (4) differs from column (3) by also including (log) average player value.

[42] Teams earn 0 points for a loss, 1 point for a draw, and 3 points for a win.

Table 8: Team level performance and passes

	points per game		ln(points per game)	
	(1)	(2)	(3)	(4)
Pass count per game (ln)	1.195***	0.4434***	0.3487***	0.2894***
	(0.1900)	(0.0707)	(0.0616)	(0.0619)
Passer valuation (ln)				0.1194**
				(0.0486)
Observations	98	1,568	1,568	1,568
Pseudo R^2	0.39971	0.62163	0.81025	0.81789
League fixed effects	✓			
League by half-season fixed effects		✓	✓	✓
Team fixed effects		✓	✓	✓

OLS regressions. Standard errors, clustered at the team level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Team * half-season level aggregated data. Top 5 soccer leagues. Column 1: First half of 2015/16 season, columns 2 and 3: 8 seasons: 2011-2019. Half-season is 16-20 games before and after 1 January. Passer valuation is average transfer value per player who played in the half-season.

Column (1) reveals a strong positive cross-sectional correlation between average points and average pass count per game. Conditioning on league, when a team exhibits an average pass count per game 10% higher than the average team (which makes about 365 passes per game), it also obtains 0.12 more points per game than the average team. Over 38 games per season, that adds up to 4.54 points, which corresponds to 9.1% more points given an average of 50 points per team. This point difference is equivalent to a difference of two to three positions in a typical league's standings. Let s_{non} denote the share of non-homophilous passes for the average team. Then, given our estimated homophily premium, another team made of same culture players, but otherwise identical to the average team, would make $s_{non}\times$ 2.42% more passes. In the extreme case in which the average team had no players with the same culture ($s_{non} = 1$), the counterfactual team would make 2.42% more passes, placing itself around one position above the average team. For the panel fixed effect models of column (2), (3) and (4), we see a smaller but still relevant positive correlation. Conditioning on league specific aggregate trends, in half-seasons when a typical team passes 10% more than its average pass frequency, it tends to win 0.044 (or 3.5% or 2.9%) more points than average. Over 38 games per season, this corresponds to almost 2 points or one position difference in a typical league's standings. If in the typical team all players had different culture, in a counterfactual half-season in which all its players shared the same culture it would make between zero and one points more than its usual points. For a team usually in the top or the bottom parts of the standings, one point more could still make a big difference in terms of sporting outcomes by allowing the team to participate to lucrative and prestigious international competitions or to avoid costly and humbling relegation to a lower league.

The other back-of-the-envelope calculation we make to investigate the importance of the homophily premium is based on the estimation of an alternative, but closely related specification to the fixed effects model (7), in which we replace player by half-season fixed effects with player values obtained from `www.transfermarkt.com`. We recover a player's estimated transfer market value, defined as the "expected value of a player in a free market" as determined by a group of experts. This estimate is based on how much a player may contribute to the team's success, how well he plays, and how valuable he may be to another team. A player's transfer value is considered a consensus summary measure of the quality of his football skills accounting for all observable (to the experts but not necessarily to us) circumstances. The most important of such circumstances include development potential, experience level, future prospects, injury susceptibility, league-specific features, marketing value, performance at the club and national team, performance potential, number and reputation of interested clubs, prestige. Accordingly, a player's value is a more general proxy of footballing skills than his fixed effect estimated from passes. For the same reason, however, it is also less tightly linked to passing performance than his fixed effect estimate. Yet, what is particularly interesting for our purposes is that a player's value also takes into account many of the player characteristics that, being unobservable to us, we soak up with player by half-season fixed effects, including the composition of the player's team in a half-season, and thus both his passing options (Λ^o and Λ^d) and his coach's decisions (τ_{od}).

Replacing player fixed effects in (7) with player values ($value_{o,t}$, $value_{d,t}$) and other observable player characteristics ($playerchar_{o,t}$, $playerchar_{d,t}$), we estimate:

$$\begin{aligned} E(pass\_count_{o,d,t}|...) = exp(&\delta SameCult_{o,d} + PassFric_{o,d,t} + \eta_1 value_{o,t} + \eta_2 value_{d,t} \\ &+\theta_1 playerchar_{o,t} + \theta_2 playerchar_{d,t}) + \phi_{c,t} \end{aligned} \quad (10)$$

where other player characteristics include age, height, time (in days) elapsed since joining the team, and a binary indicator for being on loan.[43] We also include position by half-season dummies, and nationality by half-season dummies. This specification does not feature the share of passes made by player o when player d is on the pitch (i.e. τ_{od} in the theoretical model) because, by construction, passer's and receiver's values also reflect complementary footballing skills that are intertwined with their chances of being fielded together.

[43] For additional details on loans see in Appendix, Section B.3.

Table 9: Benchmarking homophily

	pass count (1)
Same culture (any) (0/1)	0.0436***
	(0.0047)
Passer player valuation (ln)	0.3407***
	(0.0042)
Receiver player valuation (ln)	0.3461***
	(0.0041)
Average length of passes (ln)	-0.3669***
	(0.0054)
Average forwardness Ind (0-1)	0.0332***
	(0.0067)
Observations	668,982
Pseudo R^2	0.35227
Team by half-season dummies	✓
Passer: position by league and half-season dummies	✓
Receiver : position by league and half-season dummies	✓
Passer: nationality by league and half-season dummies	✓
Receiver: nationality by league and half-season dummies	✓
Cross-position dummies	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Player values (euro million in ln) are measured at the start of period. 'Personal controls' are (for both players): height, age, time since with club (in days), binary if on loan. Total team pass count is captured via team by half-season dummies.

Table 9 reports the result from estimating (10). It confirms that the number of passes made or received by a player depends on his value. It also reveals a statistically significant homophily premium. More to the point here, it allows us to put a monetary equivalent value on the homophily premium. In particular, contrasting the estimated 4.36% premium with the estimated 34.61% coefficient of log receiver valuation implies that passing to a receiver with same culture is equally likely as passing to a receiver with different culture but valued a remarkable 10.5% more.[44] This corresponds to 367'500 and 809'550 euros for the median and average players as these are valued 3.50 and 7.71 million euros respectively.

All in all, and with all their limits, both back-of-the-envelope calculations hint at the importance of the homophily premium for footballing outcomes.

6. Cost Saving vs. Favoritism

So far we have been interested in assessing whether choice homophily actually exists, while remaining agnostic about whether its existence represents an efficient outcome promoting team performance or rather a manifestation of inefficient in-group favoritism detrimental to the team. We now discuss whether the estimated homophily premium is more likely due to an objective cost for the team as passes are easier to coordinate within a group ('cost saving') or to a subjective cost for the passer who prefers to keep the ball within his own group ('favoritism').

6.1. Results

To answer the foregoing question, we rely on three identifying assumptions. First, favoritism should subside when the stakes are high for the passer as pondered decisions are more likely in this case. We should then observe less homophily than when stakes are low. Second, favoritism should be more visible when the passer belongs to a minority group, in which case we should observe more homophily if the passer belongs to a small rather than to a large group. Third, focusing on passes between players of different culture, favoritism should promote passes to smaller groups as these are less likely to keep the ball within them, whereas cost saving should promote passes to larger groups as these are more likely to keep the ball within them and thus reduce a team's passing frictions.

[44] The monetary equivalent value of the homophily premium is calculated by equating $E(pass\_count^{H}_{o,d,t}|...)$ to $E(pass\_count^{V}_{o,d,t}|...)$. The former is the predicted $E(pass\_count_{o,d,t}|...)$ using the receiver's actual value ($value_{d,t}$) and the estimated homophily premium ($\delta > 0$) from regression (10). The latter is the counterfactual $E(pass\_count_{o,d,t}|...)$ computed after setting the homophily premium to zero ($\delta = 0$) and the receiver's value to $value'_{d,t}$ such that $E(pass\_count^{H}_{o,d,t}|...)/E(pass\_count^{V}_{o,d,t}|...) = exp(\delta + \eta_2 \left(value_{d,t} - value'_{d,t}\right)) = 1$. Solving this equation gives the monetary equivalent value of the homophily premium as $value'_{d,t} - value_{d,t} = \delta/\eta_2$, with δ and η_2 estimated from regression (10).

6.1.1. High Stakes

Unintentional bias is more likely when one makes fast decisions or acts on the spur of the moment (Price and Wolfers, 2010). In contrast, there are circumstances in a game in which a passer might take a step back thus reducing bias. In this respect, as high stakes raise awareness and foster reasoning, one may expect favoritism to subside when stakes are high for the passer.[45] We consider several of such circumstances determined in terms of the type of passes or the type of players involved (younger passers; receivers of higher quality).

For the type of passes, we look at passes with longer average length and higher average forwardness (Table 10), complex pass sequences (Table 11), and passes before a shot on goal (Table 12). Specifically, column (1) of Table 10 reruns the baseline regression of Table 3 excluding average length and higher average forwardness. This exclusion has only a marginal effect on the homophily premium (2.38% vs. 2.42%). Columns 2 and 3 look at long passes directly. Long passes are identified by an indicator variable valued 1 when their length is above the median. The indicator is then interacted with the homophily premium. Column 2 reveals that homophily is stronger for long passes (1.80% vs. 3.29%). Column 3 shows that including average pass length changes the estimated coefficients only marginallly.

We then distinguish between simple pass sequences (in which player o passes to player d and the ball does not come back) and complex pass sequences (in which the ball goes back and forth between the two players at least once). On average, as already mentioned, player pairs make 15.98 passes per half-season. A vast majority of them (87%) consists of simple pass sequences, but 13% are complex pass sequences featuring 3.54 passes on average. Half of player pairs are involved in at least a complex pass sequence in our sample. Conditional on having joined at least a complex pass sequence in a half-season, on average player pairs are involved in 3.88 complex pass sequences in that half-season.

Table 11 presents the results for the number of pass sequences instead of the number of passes. For cleaner comparison, Column (1) re-estimates our baseline model using as dependent variable the number of pass sequences (simple and complex) between two given players. In contrast, Column (2) re-estimates the baseline model using as dependent variable the number of complex pass sequences in which the two players are involved. Comparing the two columns reveals that the homophily premium is more than twice as large for complex pass sequences: 4.81% in Column (2) vs. 2.04% in Column (1). It may be argued that a pass sequence is truly complex only if it is part of a forward movement. Generally speaking, two players exchange the ball back and forth in three situations: the passer sends the ball to the receiver and moves forward leaving an opponent behind to receive the ball back; the receiver is not in a good

[45] For example, as discussed by Thaler (1986), a common view - especially among economists - is: “If the stakes are large enough, people will get it right”. See Enke et al. (2023) for a recent critical assessment of this view.

Table 10: Long passes

	pass count		
	(1)	(2)	(3)
Same culture (any) (0/1)	0.0238***	0.0180***	0.0191***
	(0.0028)	(0.0031)	(0.0029)
Long pass (0/1)		-0.3219***	-0.0389***
		(0.0036)	(0.0035)
Average forwardness Ind (0-1)		0.0594***	0.0815***
		(0.0067)	(0.0061)
Same culture (any) (0/1) × Long-pass (0/1)		0.0149***	0.0120***
		(0.0042)	(0.0039)
Average length of passes (ln)			-0.7465***
			(0.0061)
Observations	668,105	668,105	668,105
Pseudo R^2	0.74508	0.75180	0.76082
Passer by half-season fixed effects	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓
Cross-position dummies	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Long pass is passes with length above the median. Total team pass count is captured via team by half-season fixed effects.

Table 11: All and complex pass sequences

Dep var:	All pass sequences			
	Forward pass sequences			
	all	complex	all	complex
	(1)	(2)	(3)	(4)
Same culture (any) (0/1)	0.0204***	0.0481***	0.0225***	0.0565***
	(0.0025)	(0.0048)	(0.0031)	(0.0073)
Observations	668,105	644,539	649,210	542,172
Pseudo R^2	0.74759	0.56028	0.74971	0.39504
Cross-position dummies	✓	✓	✓	✓
Passer by half-season fixed effects	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team by half-season fixed effects. Columns 1,2: Sequence count is the number of pass sequences, complex seq count is the number of at least 2 pass-long sequences. Columns 3,4: only sequences with no backward passes. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team by half-season fixed effects.

Table 12: Heterogeneity by passes: high stake passes

	pass_count (1)	pass_neargoal (2)
Same culture (any) (0/1)	0.0242***	0.0244***
	(0.0041)	(0.0079)
Observations	668,105	480,264
Pseudo R^2	0.75929	0.40891
Passer by half-season effects	✓	✓
Receiver by half-season fixed effects	✓	✓
Cross-position dummies	✓	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Same cultural background is equality of either cultural aspect (language, colonial, federal legacy or nationality). High stake passes are those creating a chance or precede a shot on goal. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team by half-season fixed effects.

position to benefit from the ball and thus sends it back to the passer; the two players want to "waste time" to hold on to a favorable result for their team. Only the first situation should be classified as a "complex" pass sequence as it involves a coordinated movement of the first passer to receive the ball back. In the other two situations giving the ball back to the passer may simply be the easiest alternative.[46] As we know the pitch coordinates of each pass, we can filter out sequences with backward passes and focus only on the remaining ones. When we do that, columns (3) and (4) reveal an even bigger difference in the homophily premium: 2.25% vs. 5.65%. This confirms that homophily is especially important for more complex collaboration.

Lastly, we study close-to-goal passes that either create a big chance or lead to a shot on goal, which are 2.5% of all passes in our data. Table 12 reports the results for the number of close-to-goal passes instead of the number of passes. The point estimate of the homophily premium for the former is the same as for the latter.

Turning to the type of players, more pressure may be felt by younger passers and passers targeting receivers of higher quality as in both cases a passer might feel he is more likely to be assessed critically for possible mishaps. To investigate the possible relevance of these aspects, we compare young passers with experienced ones as well as passers targeting low or high quality receivers. We do so by interacting binary indicators with the same culture variable. Specifically, we define an indicator variable Het_t equal to

[46] We thank an anonymous referee for making this point.

$Het\_age_{o,t} = \{young, experienced\}$ in the case of age and $Het\_quality_{d,t} = \{low, high\}$ in the case of receiver's quality. We then estimate the following specification

$$\begin{aligned} E(pass\_count_{o,d,t}|...) = exp(&\delta_1 SameCult_{o,d} + \delta_2 Het_t + \delta_3 SameCult_{o,d} \times Het_t \qquad (11)\\ &+ PassFric_{o,d,t} + \ln \tau_{o,d,t} + \upsilon_{o,t} + \upsilon_{d,t}) \end{aligned}$$

where δ_3 captures the difference in homophily gap between types, with $\delta_2 = 0$ for $Het_t = Het\_age_{o,t}$ as age is stable over a half-season.[47] A passer is 'young' ($Het\_age_{o,t}$) if he belongs to the lowest quartile of the age distribution (below 23.2 at the start of the half-season), and 'experienced' otherwise ($Het\_age_{o,t} = 0$). A receiver is high quality ($Het\_quality_{d,t} = 1$) if he is one of the top 2 teammates in terms of market value players per squad, and 'low' quality otherwise ($Het\_quality_{d,t} = 0$).[48]

Results are reported in Table 13. In column (1) the homophily premium is higher for young passers (3.19%) than for experienced ones (2.23%). In column (3) the homophily premium is unaffected by receiver quality.

Taken together, these results suggest that the homophily premium is unlikely due to favoritism because homophily is not less (and in some case is actually more) pronounced when stakes are high.

6.1.2. Group Size

Relative group size may affect the homophily of a group's members (Jackson et al., 2017). Being in a relatively small group may increase in-group favoritism (Porter and Washington, 1993). In our case, favoritism should be more visible when the passer belongs to a small group.

To check whether this happens, we measure group size as the number of same culture receivers a player faces each time he passes on average over a half-season. This measure ranges between 1 when on average no receiver has the same culture as the passer, and 11 when on average all receivers share his same culture. We then define an indicator variable $Het\_grsize_{o,t} = \{small, large\}$ that records whether the passer belongs to a small or a large group. His group is 'large' ($Het\_grsize_{o,t} = 1$) if its average size is larger than 4, and 'small' ($Het\_grsize_{o,t} = 0$) otherwise.[49] In our sample, 66.6% of passes are started by passers that are part of large culture groups (i.e. on average at least 4 players of their group are fielded).[50] We finally run regression (11) with Het_t equal to $Het\_grsize_{o,t}$.

[47] In specification (11) the three aspects are treated moderators. An econometric challenge to using interaction terms in panel data is that, despite the fixed effects, they may be confounded by other potential moderators. A potential solution is to use split samples – interacting all variables, including fixed effects with the moderator variable. Split sample results confirm those with simple interactions.

[48] Results are robust to taking the top 4 rather than the top 2 teammates by value. They are also robust to considering as high quality all players valued more than 15 million euros.

[49] Results are robust to alternative cutoffs.

[50] For details see E in the Appendix.

Table 13: Heterogeneity by passer: age, experience, group size

Dep.var: Pass count	(1)	(2)	(3)
Same culture (any) (0/1)	0.0317***	0.0174***	0.0236***
	(0.0045)	(0.0043)	(0.0027)
Same culture (any) (0/1) × Passer age (0/1, 1=Experienced)	-0.0094**		
	(0.0048)		
Same culture (any) (0/1) × Passer group size (1/1, 1 when N¿=4)		0.0146***	
		(0.0059)	
Same culture (any) (0/1) × Receiver quality (0/1, 1= top 2)			0.0044
			(0.0057)
Passer group size (1/1, 1 when N¿=4)		-0.0358***	
		(0.0067)	
Receiver quality (0/1, 1= top 2)			0.0129
			(0.0081)
Observations	668,105	668,105	668,105
Pseudo R^2	0.76077	0.76078	0.76077
Passer by half-season fixed effects	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓
Cross-position dummies	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The average length of passes and average forward index are included. Total team pass count is captured via team by half-season fixed effects. Receiver quality is based on transfer values, highest two per team. Group size is based on same culture groups. Experienced player is aged 23.2 on the starts of the half-season.

The corresponding results are reported in column (2) of Table 13, which shows that, against the minority hypothesis the homophily premium is higher for players in large groups (3.20%) than for those in small groups (1.74%). Hence, the homophily premium is again unlikely due to favoritism.

Moreover, focusing on passes to receivers of different culture than the passer's, favoritism should promote passes to smaller groups, whereas cost saving should promote passes to larger groups (Currarini and Mengel, 2016; Jackson et al., 2017). To investigate whether this is the case, we look at passes going from passer o to receiver d recording whether o and d belong to small or large groups as defined before. We then rerun the fixed effects regression (7) controlling whether their passes entail passing from a small to a large group, from a large to another large group, from a large to a small group, or from a small to another small group. The last type of passes is taken as base, so that all other variables (including the same culture dummy) are measured against it.

Table 14 reports the corresponding results, which show that passes are more likely from small to large groups than from large to large groups. In turn, passes from large to large groups are more likely than those to small groups, no matter whether these originate from large or small groups.[51] Hence, also according to the evidence on out-group passes, the homophily premium is unlikely due to favoritism. In addition, the higher frequency of passes to large groups is consistent with the idea that keeping the ball in those groups minimizes a team's passing frictions.[52]

6.1.3. Shared Experience

The contact hypothesis in psychology suggests that prejudice and conflict between groups can be reduced if members of the groups interact with each other (Pettigrew and Tropp, 2006). Common features of prejudice include having negative feelings and holding stereotyped beliefs about members of the group, as well as a tendency to discriminate against them. Frequent instances involve prejudices based on characteristics like race, sex, religion, and also culture. Prejudice often supports in-group favoritism and out-group discrimination, which nonetheless should be mitigated through intergroup contact.

To check whether this happens in our data, we study the evolution of the homophily premium over time as players repeatedly interact. In particular, consider a player new to a team: all teammates are initially new to him. Over time, new players will arrive and there will be some teammates with whom he will have shared experiences and others who will be new to him. To measure shared experience, we use a threshold of 215

[51] As passes from a small to another small group is the base, its coefficient is zero.

[52] This is reminiscent of the conclusions of Giannetti and Yafeh (2012) with regard to the observed negative effects of cultural differences between contracting parties on the terms of syndicated bank loans. Their favorite interpretation of this finding is that cultural differences make negotiations more cumbersome and thus increase contracting costs.

Table 14: Pass from and to larger groups

	pass_count (1)
Same culture (any) (0/1)	0.0358***
	(0.0048)
Different culture pass: Small to large (0/1)	0.0283***
	(0.0073)
Different culture pass: Large to large (0/1)	0.0176***
	(0.0059)
Different culture pass: Large to small (0/1)	-0.0068
	(0.0073)
Different culture pass: Small to small (base)	0
	0
Observations	656,776
Pseudo R^2	0.74018
Passer by half-season fixed effects	✓
Receiver by half-season fixed effects	✓
Cross-position dummies	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team by half-season fixed effects. Cutoff for group size: large group is defined as an average size of at least 3 members.

Table 15: Homophily over time: shared experience

	pass_count				
	(1)	(2)	(3)	(4)	(5)
Same culture (any) (0/1)	0.0166***	0.0163***	0.2325	0.0131*	0.0206***
	(0.0053)	(0.0053)	(0.2156)	(0.0078)	(0.0050)
Same culture (any) (0/1) × Experience	0.0117**	0.0127**	-0.1372	0.0191**	
	(0.0059)	(0.0060)	(0.1924)	(0.0088)	
Same culture (any) (0/1) × Experience long					0.0073
					(0.0059)
Observations	457,838	443,641	13,530	219,178	384,818
Pseudo R^2	0.76317	0.76431	0.83248	0.76578	0.76699
Early experience w other team	Include	Exclude	Only	Include	Include
Time with team capped	No	No	No	Yes	No
Passer by half-season fixed effects	✓	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓	✓
Cross-position dummies	✓	✓	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer by half-season and receiver by half-season level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Shared experience is binary, 1 if passer and receiver has spent at least 215 at current team together. Shared experience long is 364 days or more. Previous experience is at past clubs including youth teams. In all models, the average length of passes and average forward index are included. Total team pass count is captured via team by half-season fixed effects.

days (7 months), which is roughly equivalent to a half-season including summer (May to December/June to January) and very close to the median by days spent together at a team. We define a passer and a receiver as having shared experience if they have spent at least 215 together at their current team. We then consider only passers who have been in the team for more than 215 days and compare their passes to receivers with and without shared experience. This gives a 68% subsample of the original data featuring 29.5% passing pairs without shared experience and 70.5% passing pairs with shared experience.

The corresponding results are shown in column (1) of Table 15, which reveals that homophily is more pronounced for pairs with shared experience. This goes against the contact hypothesis and suggests that favoritism based on prejudice is not what drives homophily.

We carry out two robustness tests. First, it is possible that some pairs have prior experience having played together before in other teams, including youth teams. In columns (2) and (3) we then repeat the exercise for pairs without any previous shared experience and only previous shared experience respectively. In the latter case the sample is rather small and having only previous shared experience has no significant effect. Second, in Column (4) we constrain the dataset to players in the first two years with their current team. This is a much smaller dataset, with a higher share of new partnerships (40.5% vs 29.5%). Here we find an even stronger relative role of shared experience.

Finally, we show that the time horizon for shared experience is fairly short. In column (5), we replace the 7 months threshold with 12 months, only to find that the (imprecisely) estimated effect of shared experience becomes smaller. This suggests that the impact of shared experience materializes fairly fast, though the coefficients in columns (1) and (4) cannot be told apart statistically.

Overall, we find evidence that homophily is higher for player pairs with some shared experience with the team. This finding supports the idea that homophily is not driven by prejudice.

6.2. Discussion

While the body of evidence presented in the previous section does not offer us a definite proof that the homophily premium is about cost saving rather than favoritism, it still provide us with a set of clues that consistently point in that direction. In light of these clues, we review several potential mechanisms behind the homophily premium that have been highlighted in the vast literature surveyed in Section 2.

Identity. People who share a national identity may want same nationals to do well. This would be mostly a national issue and not a cultural one, that is, it would explain Brazilian but not Portuguese-Brazilian homophily. Our results showing how post-colonial links or other forms of cultural proximity matter do not support the national identity mechanism. It may, nevertheless, support a mechanism working through broader cultural identity. Yet, our finding that inter-group passes favor receivers in large groups is inconsistent with in-group favoritism as large homophilous groups are more likely than small homophilous groups to keep the ball within them.

Prejudice: People may have a false overly confident belief in the ability of same culture peers and underestimate the ability of different culture ones. According to the contact hypothesis, the prejudice mechanism can be detected if homophily declines with shared experience. This not what we find as we see the opposite happening.

Salience. Being in a small group leads to salience of cultural affiliation as people are more likely to be aware of belonging to the same group when the group is small. This is not what we find when we look at a passer's group size. It is also inconsistent with the finding that inter-group passes favor receivers in large groups even when the passer belongs to a small group.

Friendship. People may have no personal preference when they start collaborating, but will build friendships over time. Friends outside work will then help each other in the workplace (i.e. on the pitch). Stronger collaboration may then ensue (on purpose or even involuntarily). Friendship may be easier to build with people from the same culture, as they typically speak the same language, have the same social cues, like the same cuisine, listen to the same music, watch the same tv series or sports, and so on.[53] Friendship as a mechanism may then be detected if homophily increases with time spent together with teammates. This is consistent with our finding that shared experience amplifies the homophily premium.

To sum up, our results can be consistently interpreted as hints at cost saving as the source of homophily and off-pitch familiarity as a possible mechanism through which same culture leads to homophily.

7. Conclusions

We have investigated how homophily based on cultural traits affects collaboration in superstar multinational teams. In doing so, we have collected and exploited a newly assembled exhaustive dataset recording all passes by professional European football players in all teams competing in the top five men's leagues over eight sporting seasons, together with full information on players' and teams' characteristics. The outcome we have chosen as our measure of collaboration is the 'pass rate'. Passes represent how players work together for the common objective of scoring a goal and are positively correlated with team performance.

The main cultural traits we have based our investigation on as antecedents of homophily are nationality, colonial legacy, federal legacy, and language, and we have measured culture through their combination. We have used a discrete choice model of players' passing behavior as a baseline to separately identify homophilous collaboration due to cultural preferences (”choice homophily”) from homophilous collaboration due to opportunities (”induced homophily”).

We have found strong evidence of choice homophily with respect to some cultural traits. Players do have a preference to pass more to players of their same culture than to other players: player pairs of the same nationality have a pass rate 2.42% higher than player pairs of different nationalities. Among the alternative operationalizations of culture, only same colonial legacy has been revealed as consequential as same nationality. In particular, we have found no evidence supporting the positive relevance of shared values as measured by the World Values Survey.

We have also shown that the estimated homophily is important for team outcomes in terms of sporting outcomes, and that passing to a receiver with the same culture is as likely as passing to a receiver with different culture but valued a conspicuous 10.5% more.

Our estimated homophily could be due to an objective cost for the team as passes are easier to coordinate within a group or to a subjective cost for the passer who prefers to

[53] We are grateful to an anonymous referee for flagging this mechanism. Interestingly, Kovacs and Kleinbaum (2020) find that even similar linguistic style will lead to friendship formation in college students.

keep the ball within his own group. To tell these alternative explanations apart, we rely on three identifying assumptions inspired by the literature: favoritism should subside when the stakes are high for the passer; favoritism should be more visible when the passer belongs to a minority group; for inter-group passes, favoritism should promote passes to smaller groups as these are less likely to keep the ball within them, whereas cost saving should promote passes to larger groups as the fact that these are more likely to keep the ball within them reduces a team's passing frictions. In our data, evidence on all three aspects pointed in the same direction: the observed choice homophily is unlikely to be due to favoritism.

Finally, we have discussed possible mechanisms behind the observed cultural homophily that have been highlighted in the literature. Players' experience does not change the results. When time spent together on the pitch (a component of induced homophily) is not controlled for, measured homophily increases: same culture players are selected to play together at 1.38 percent higher frequency than players with different cultural backgrounds. This reveals the managers' role in making their teams internalize players' homophilic preferences. The fact that players' experience does not change the results suggests that choice homophily is not due to prejudice, limited familiarity with diverse environments, or lack of professional experience. Moreover, we have found that homophily increases with the time players spend in the same clubs. While we do not know how much time players pass together outside the workplace, we have conjectured that such a result hints at off-pitch familiarity as a possible mechanism through which the same culture leads to homophily.

To summarize, we have shown that choice homophily based on culture is pervasive and persistent even in teams of very high-skill individuals with clear common objectives and aligned incentives, who are involved in interactive tasks that are well-defined, readily monitored, and not particularly language intensive. Cultural similarity affects collaboration also in superstar teams of professionals at the top of their industry. This is not by itself necessarily inefficient as it appears to allow the teams to minimize objective barriers to collaboration rather than stemming from team members' overestimation of the abilities of similar others, or simply from a fundamental irrational indication of in-group partiality.

We see at least two promising directions of future research to overcome some of the limitations that we have highlighted in our analysis. First, our assessment of the importance of homophily for team performance has relied on the correlation of teams' passes with their points in league rankings and on the back-of-the-envelope calculation of a monetary equivalent value for the estimated homophily premium. In both cases the question of whether the estimated homophily premium makes a difference in terms of footballing outcomes has been only indirectly answered. As discussed, a direct answer would require a full-fledged structural model of team performance and league competition. While such a full-scale structural analysis was beyond the scope of this paper, we view our assessment of the importance of homophily as a first step in that direction. Second, we have interpreted our finding that shared experience amplifies the homophily premium as hinting at off-pitch familiarity as a possible mechanism through which same culture leads to homophily. In future work, this conjecture could be tested by matching our data with the players' social media profiles to gauge their reciprocal friendship ties and their intensity.

Appendix

A. Model details

In this appendix we provide additional details on the derivations of the discrete choice model presented in Section 4. The team's payoff benefit maximized by player o is given by:

$$(1-\beta)\,U^o+\beta\max_{\{d\}_{d=1}^{N}}\left\{\varphi^d[U^d]-\widetilde{c}^{o,d}+z^d\right\}, \tag{12}$$

where z^d is the realization of a random variable Z following a Gumbel distribution (Type-I Extreme Value distribution)

$$G(z)=\exp\left(-\exp(-\kappa z)\right),$$

with mode 0 and concentration around the mode inversely related to $\kappa>0$. After defining $V^d\equiv\exp U^d$ and $c^{o,d}\equiv\exp\widetilde{c}^{o,d}$, the *ex ante* probability that player o in possession of the ball successfully passes to teammate d is

$$\pi^{o,d}=\left(V^d\right)^{\kappa\varphi^d}\left(c^{o,d}\right)^{-\kappa}(\Lambda^o)^{-\kappa}\ \text{ with }\Lambda^o\equiv\left[\sum_{s=1}^{N}(V^s)^{\kappa\varphi^s}\left(c^{o,s}\right)^{-\kappa}\right]^{\frac{1}{\kappa}}, \tag{13}$$

which *ex post* becomes (approximately) the average share of successful passes that player o makes to player d in a half-season in the subset of passing episodes $T^{o,d}$ when both o and d are fielded and player o has ball possession. The fact that also $s=o$ is included in the sum $\sum_{s=1}^{N}(V^s)^{\kappa\varphi^s}\left(c^{o,s}\right)^{-\kappa}$ implies that $\sum_{s=1}^{N}\pi^{o,s}=1$ holds.

As in the main text, let us use P^o to denote the number of passing episodes involving player o as the passer, P^d to denote the number of passing episodes involving player d as the receiver, and $P^{o,d}$ to denote the number of passing episodes involving involving player o as the passer and player d as the receiver over a half-season. Accordingly, we have $P^o=\sum_{d=1}^{N}P^{o,d}$, $P^d=\sum_{o=1}^{N}P^{o,d}$ and $P=\sum_{o=1}^{N}\sum_{d=1}^{N}P^{o,d}$. Moreover, let $\tau^{o,d}$ be the fraction of passing episodes involving player o as the passer when player d is on the pitch (i.e. player d is an option as receiver). Then, we can write:

$$P^{o,d}=P^o\tau^{o,d}\pi^{o,d}.$$

Using expression (13) to substitute for $\pi^{o,d}$, we have

$$P^{o,d}=\frac{\left(V^d\right)^{\kappa\varphi^d}\left(c^{o,d}\right)^{-\kappa}\tau^{o,d}P^o}{(\Lambda^o)^{\kappa}}, \tag{14}$$

which can be summed across $d = 1, ..., N$ to obtain

$$P^o = \sum_{d=1}^{N} \frac{P^{o,d}}{\tau^{o,d}} = \sum_{d=1}^{N} \frac{\left(V^d\right)^{\kappa\varphi^d} \left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}.$$

From this expression we can retrieve

$$\left(\Lambda^o\right)^{\kappa} = \sum_{d=1}^{N} \left(V^d\right)^{\kappa\varphi^d} \left(c^{o,d}\right)^{-\kappa}. \tag{15}$$

Analogously, summing (**??**) across $o = 1, ..., N$ gives

$$P^d = \sum_{o=1}^{N} \frac{P^{o,d}}{\tau^{o,d}} = \sum_{o=1}^{N} \frac{\left(V^d\right)^{\kappa\varphi^d} \left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}.$$

From this expression we can retrieve

$$\left(V^d\right)^{\kappa\varphi^d} = \frac{P^d}{\sum_{o=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}} \tag{16}$$

Substituting (16) in (15) yields

$$\left(\Lambda^o\right)^{\kappa} = \sum_{d=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^d}{\sum_{o=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}} \tag{17}$$

Expression (16) can be used to rewrite (14) as

$$P^{o,d} = \frac{\frac{P^d}{\sum_{o=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}} \left(c^{o,d}\right)^{-\kappa} \tau^{o,d} P^o}{\left(\Lambda^o\right)^{\kappa}}$$

and thus

$$P^{o,d} = \frac{\frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}}{\sum_{o=1}^{N} \frac{\left(c^{o,d}\right)^{-\kappa} P^o}{\left(\Lambda^o\right)^{\kappa}}} \tau^{o,d} P^d. \tag{18}$$

Collecting terms that are specific to player o, bilateral, or specific to player d leads to the expression (3) in the main text:

$$P^{o,d} = \frac{P^o}{\left(\Lambda^o\right)^{\kappa}} \left(c^{o,d}\right)^{-\kappa} \tau^{o,d} \frac{P^d}{\left(\Lambda^d\right)^{\kappa}}$$

with

$$\left(\Lambda^{d}\right)^{\kappa}=\sum_{o=1}^{N}\frac{\left(c^{o,d}\right)^{-\kappa}P^{o}}{\left(\Lambda^{o}\right)^{\kappa}}.$$

B. Football rules

B.1. Institutional setup in European Association Football

Association football in Europe is organized at the country (UEFA association member) level in several tiers. The English Premier League includes English and Welsh teams, but not Scottish or Norther Ireland teams. In this paper, we only look at men's football at the top tier in each country.

Teams are owned by groups of supporters (the main approach in Germany) or investors (the main approach in England). In terms of sporting management, coaches train players and select the players to play at every game and decides on tactics. Coaches are mostly (but not always) managers as well, deciding on player transfers. In most cases hiring and termination and sales transactions are decided in a team that includes the coach (manager), the technical director and owners.

In what follows we also summarize key rules, transfers and nationality rules across leagues.

B.2. Key football rules

This subsection describes the key rules in football (soccer). Association football, such as our leagues, is governed by the Laws of the Game.[54] In this section, we review some relevant aspects of the game.

In a league, all teams play all other teams twice: in a home and in an away game. A team gets 3 points for winning, 1 for drawing and 0 for losing. There is churning season by season: every year the worst few (2 or 3) teams are relegated, while a few are promoted from the lower division to replace them.[55]

Due to the flow of the game, almost two thirds of the events are passes and about 75% of these passes are successful. The rest of the events include shots on goal, goals, free-kicks, actions to contest the ball, yellow and red cards for disciplinary action, and substitutions.

In a game, there are twenty-two (2x11) players on the pitch. All decisions on who plays are down to the head coach, who is sometimes also the manager of the team. For each game, the coach nominates a "starting 11" - 11 players who start the game. In the period of observation there are up to 3 substitutions per team/game (these tend to occur in the last third of the game). Substitutions may happen because of an injury or due to some tactical decision. At any time during a game, there are 11 players on

[54] For details see https://en.wikipedia.org/wiki/Laws_of_the_Game_(association_football)

[55] Readers unfamiliar with soccer may find additional details here: https://en.wikipedia.org/wiki/Ted_Lasso.

the pitch unless a player gets a red card and is sent out (permanently). However, this rarely happens - about once every five games.

There is freedom in selecting the players on the pitch, but mostly they consist of 1 goalkeeper, 3-5 defenders, 3-5 midfielders and 1-3 forwards (strikers). There are some typical passes that come from how football is played: goalkeepers mostly pass to defenders or kick the ball far ahead; midfielders make a lot of passes among themselves, forwards pass relatively less. Teams may have distinctive styles: some play focusing on possession with a great deal of passing activity, while others wait and rely on counter-attacks. Some teams will try to have many shots at goal while others will pass more waiting for an ideal opportunity. As balls may be contested (dribbles, duels, tackles), better players are expected to control the ball more; and teams with better players to pass more. As a result the total number of passes by a team in a game depends on both quality and style.

B.3. Teams and transfers

Football teams are organizations with 25-30 players (also called the squad). Churning in the squad composition is high from one season to another, typically 20-40% of a team changes. A transfer means that a player leaves or arrives after being sold or bought by the team. In Europe, transfers happen twice a year. The main opportunity to get new players, or sell existing ones, is between 1 July and 1 September, also called the summer transfer window. Over 90% of deals in a season happen during this period. The winter window is shorter - from 1 January to 1 February- and much smaller. Transfers may include loan deals. A player is 'on loan' when playing temporarily for a club other than the club holding his contract. The typical length of a loan contract is one or two half-seasons (and in rare cases it may be longer).

Games are also held during the transfer windows, which generates complications with respect to measurement - see in the Appendix section D.

B.4. Nationality rules in leagues

Some leagues do not limit the number of nationalities playing for a team on the pitch, while others restrict the number of non-EU, especially South American players. In addition, some leagues have rules regarding the composition of the squad (e.g. squads must have home grown 'academy' players), but this has very little effect on initial selection of players to be fielded in a match ('starting 11').

Regarding the five leagues in our data, there are two types of regulations.

France, Italy and Spain have restrictions about the number of 'foreign' players, defined as players coming from non-EU countries. In France their number is capped at 4, in Spain at 3, and in Italy at 2. Among these three countries, the definition of non-EU varies only marginally, but in all cases they include a set of African, Caribbean and Pacific countries (such as Nigeria, Ivory Coast, Guyana) with which the European

Union has eased labor laws under the “Cotonou” agreement.[56]

In Italy and Spain proving ancestry can fast track getting citizenship. In Spain, South American players are able to obtain citizenship after 2 years instead of 5, if they can show Spanish ancestry. Many Argentinian and Uruguayan players have been able to become citizens due to their ancestry in Italy.

These non-EU restrictions are binding mostly for South American players (without double citizenship). As a result of these regulations, in France, Italy and Spain, two Brazilians or a Uruguayan and an Argentinian player are less likely to play together than two Europeans.

While England and Germany do not have restrictions on players coming from non-EU countries, both countries (but especially Germany) have preference for home grown (also called ‘academy’) players. In England, visa restrictions favor players who play or have the potential to play for their national team.

We have coded all of these restricting regulations. Overall, in our estimation dataset, 89% of observations have a passing player who is considered to be unrestricted in the European Union. In a robustness check, we condition on these regulations and find them having no effect on our results.

Finally, all personal information on the players is dated back to the summer of 2021. This might give rise to bias, as a few players may get new citizenship over their career, but we may only see it for older players who have already got it. For young players who have ancestry and will get nationality in the future, we may not see it. This may downward bias our same nationality estimates marginally.

C. Measuring culture

In this subsection, we describe the major decisions.

C.1. Defining player nationality

We kept nationalities as defined by FIFA, the international football’s governing body.[57] In practice, a nationality is defined as being a polity with a national football team. In most cases a country would form a nationality. However, there are some exceptions: the United Kingdom has four national teams (Wales, Scotland, Northern Ireland, and England), and small nations like the Faroe Islands (a constituent country of Denmark) or Jersey (a British Crown Dependency) are also treated as separate nations. Apart from the UK, which is treated separately in the paper, all others have only a handful of players.

In the period of observation some players changed nationality as their countries, such as Czechoslovakia, Yugoslavia and the USSR, dissolved and gave birth to new

[56] https://www.footballmanagerblog.org/2018/04/football-manager-squad-registration-rules.html. See https://en.wikipedia.org/wiki/Cotonou_Agreement for details on the list of countries.

[57] See Article 5 principles in FIFA (2021)

countries. Hence, for country of birth we had to occasionally make edits to match the current list of countries.

C.2. Measuring player colonial legacy, federal legacy, and language

Regarding colony and official language definitions, we followed CEPII data.[58] Similar (but not the same) languages like Danish and Norwegian were considered as different. For the definition of same culture, we had to consider pairs with multiple nationalities. For instance, it is possible that two players have both the same nationality and the same colonial legacy (for example, P1 is Moroccan and French and P2 is French). In such cases, we adopted a '*top-coding*' approach, and considered them to be of the same nationality. This is actually a large set of player pairs: 54% of those who share a colonial past also share a citizenship. As for colonial legacy, the majority of the same colonial legacy category comes from a link between a ruler and a former colony. Some links are derived from having the same colonial ruler. It is possible for two players to have a ruler-colony legacy and a colonial sibling legacy (for example, if P1 is a citizen of Ivory Coast and P2 is a citizen of Senegal and France). In 86% of the cases, the same colonial history means the same language as well. However, this is not true for all country pairs (e.g. England and Egypt or Russia and Georgia). Some countries had multiple colonial rulers (such as Cameroon with France and England, so linked to both).

Beyond colonial linkages a small group (1.47%) is formed by countries that used to belong to a political union and now have separate teams: the USSR, Yugoslavia, the USSR, Yugoslavia, countries of the British Isles (Ireland, Northern Ireland, England, Scotland, Wales, as well as Jersey, Gibraltar). These formerly federated or currently partially federated countries are considered to have the same federal legacy.

C.3. Defining cultural distance from World values Survey

Cultural distance in terms of values is based on well known World Values Survey (WVS), which has been running for many years. We use the Wave 7 round because it extended with a European version that covers many smaller European countries like Croatia.

We define cultural distance based on values as an average Euclidean distance between the share of replies to categorical variables. For $n = 1...N$ survey questions, each with $K|n$ answers (ie $K = 2$ for a binary, and $K = 5$ for a typical categorical question), we have

$$D_{WVS=} \sum_{n=1}^{N} \left(\sum_{k=1}^{K} \left(y_{n,k}^{A} - y_{n,k}^{B} \right)^2 \right)^{0.5} \tag{19}$$

[58] The 25 most frequent official languages (in order of frequency in the estimation dataset) are Spanish, Italian, English, French, German, Arabic, Dutch, Portuguese, Russian, Polish, Serbo-Croat, Bulgarian, Turkish, Czech-Slovak, Swedish, Hungarian, Georgian, Macedonian, Norwegian, Albanian, Ukrainian, Finnish, Danish, Slovene and Greek.

The simple euclidean distance has some relevant characteristics. It works both for binary and categorical variables, and they may be added. It is symmetric with respect to country A to B and B to A. It has a zero value when country A = country B. It ranges between 0 (same country or at least same values) and $\sqrt{2}$ (total polarization). We selected 72 relevant questions[59] and assume the stability of values for the decade we use it for.

Data was available for 80 countries compared to 132 in our data. For missing countries we imputed values based on geographical and historical similarity[60], distance between imputed country and its reference was set at the median distance value.

In our data for different countries, the mean is 0.25, the median is 0.17, and minimum is 0.08 and the maximum is 0.56. In the empirical work, we create a binary variable for more similar pairs and different ones, as below and above the median value.

D. Data cleaning and entity resolution

D.1. Matching players from two sources and entity resolution

Data on football players come from two different sources: passing data and player information. In each datasets, players are identified via their names. To combine them, we developed an entity coreference algorithm to match players based on variations on their names, and some background information.[61]

Our method improves upon a standard fuzzy matching algorithm. First, even for ten thousand players as in our sample, it takes a lot of computing power to calculate all possible similarities and find the best ones. Second, simply matching the players by themselves is not precise enough, and thus we must use additional information, such as their teams or first nationality. However, even player features (such as team names) are also not precise and unique.[62] Third, data quality problems also mean that in one dataset some players might have two or more different records. Fourth, we added an algorithmic checkup, because re-examining and correcting the possible matches for over ten thousand players by hand is simply not feasible.

Our improved solution relied on introducing 'motifs': a combination of player features. Instead of simply matching players from the two datasets, we match motifs in a network of players, matches, seasons and teams. This way, we can utilize the already discovered coreferences in order to narrow the search space. In addition, the noise in the data can be mitigated as we rely on more than one similarity to establish a coreference.

The algorithmic matching is not perfect as players may use different names, and

[59] We used "EVS_WVS_Joint_Stata_v4_0.dta", see details in the Online Appendix

[60] See details in the Online Appendix

[61] We thank Endre Borza. For additional details, see https://github.com/endremborza/encoref.

[62] This problem can be illustrated by the names of two teams. In one dataset, two clubs are called 'Athletic' and 'Atletico Madrid', while in the other the same clubs are referred to as 'Athletic Bilbao' and 'Atletico'. Hence, the solution must be open to the possibility, that the two entities, 'Athletic' and 'Atletico' are different even though they are very similar in name.

accents may be incorrectly used as well. When the matching score was low, we checked the match by hand and corrected player names if deemed necessary - reaching about 1% of total player names.

D.2. Detailed cleaning steps and decisions

One important aspect is the possibility of zero passes. Due to aggregation, all passer-receiver pairs in the pass data have non-zero passes. However, there are 52,092 player-pairs*time (7.8%) where only one direction of the pass is recorded. As clearly a pass was possible, we added zeroes for these pairs for the opposing direction.

There are several additional steps of data wrangling:

- We dropped observations (N=340), when a player had only a single partner in half-season.
- Player age for every season was defined as the number of days to the 1 of September in the current year. When a player age was missing, we created sample means by teams and seasons and replaced the missing with that mean.
- When player position was missing, we replaced it with 'Central Midfielder'.
- Player ID was missing in 0.1% of cases and in 62 cases the passer and receiver were the same. We dropped these observations.
- When a player value for one season was missing, we imputed his average valuation over time. When player valuations were missing, we imputed 100,000 euros. This happened almost entirely for young and new players.
- There were 6 player pairs who moved together to a different club within a time period. We dropped them.
- As noted earlier, games are not suspended during the transfer windows. Hence players moving within a window may end up playing for more than a team in a half-season. In our sample, we observed 954 events when players played for two teams within the same half-season.[63] We kept them only once, in the team where they had the longer spell.[64]

E. Counting culture groups

In this subsection, we document how same cultural groups were defined and counted. Counting same culture groups is not a trivial process for two reasons. First, there is 11 players on the pitch at once, but typically 13 will play in a game, and a player will pass to

[63] There were 374 players who not only moved teams, but also moved leagues.

[64] Very rarely (10 directional player pairs) we observed a given player pair passing in two different teams in the same half-season.

15-20 others in a half-season. Second, players have multiple cultural background, and a Uruguay - Italy player will share cultural group with another Italian (same nationality) or an Argentinian (same colonial heritage) player.

We implemented a new procedure to create $n\_cultureq_any$, the number of players on the pitch (out of 11) sharing the same culture (including nationality, colonial legacy, federal legacy or language). This is computed for both the passer and the receiver.

E.1. Counting events by game-segments

The procedure first considers the eleven players on the pitch at any point in time, thus group size varies over time to match substitutions. In a game, up to 3 substitutions were possible in this period, but red cards may also alter groups.

Let us define a game segment between any change in the team. This yields a handful (median =3) of segments in a game, with unchanged team composition in a segment but different between any two.

For each game-segment, we can compute the size of culture group i as the number of players who belongs to that group: $1 =< n\_culture\_any\_group_i <= 11$

These segment information then combined with pass data. Thus, for every single pass, we know know the size of same culture groups for all players.

E.2. Dealing with multi-culture players

A key difficulty comes from players with more than a single culture (because of bi-cultural country like Belgium or multiple, different culture nationality like Spanish and French. There is no completely neutral way of dealing with it, but we developed a method that allows the largest possible group size value to manifest. We deal with this with the following procedure.

For each player, based on his nationalities, create a list of cultures he belongs to. A dual culture passer would belong to both culture groups, and his group size will be the larger one unless passing to the other (smaller) culture group. For example, in a team with this player and 3 Spanish and 1 Italian, he would be part of the Spanish group (N=4) except when passing to the other Italian, when he is part of that group (N=2). We count the maximum possible group size regarding multi-culture players. As a consequence of players with multiple group memberships, the total number of players in groups may be above 11.

E.3. Defining large group

Another aspect of same culture groups is to map the share of passes that a passer has (i) being in a large group or not, to (ii) a receiver being in a large group or not. This setup creates possibly 4 options. Same culture players are defined to be in a group, so this distinction is only relevant from different culture group passer-receiver pairs.

We start by creating a binary indicator being in a large same culture group. We define a cutoff c to classify $large\_culture\_group$ as a group with at least c members from a culture. The simplest case is $c = 2$, so for a passer or a receiver, $large_culture_group = 1$ is a group with at least two same culture members. In practice $c = 2, 3, 4$.

E.4. Aggregation to season-half level

Our data remained at the pass level. As part of aggregation, we now average over passes for a half-season. For the same culture group size, we simply calculate the average. As a result of the procedure this average is weighted by the number of passes. As for bilateral in/out of large culture group values, we once again create a share, ie a weighted average of 0 and 1. Because players often play with similar partners, these average shares are often (91% of the cases) 0 or 1, with some values in between.

E.5. WMS

This is the list of questions we used[65]: a001, a002, a003, a004, a005, a006, a008, a027, a029, a030, a032, a034, a035, a038, a039, a040, a041, a042,a165, c001, c001_01, c002, c002_01, d081, d026_03, d026_05, c038, c039, c041, d001_b, g007_36_b, d054, d059, d060, e035, e036, e037, e039, e069_01, e069_02, e069_04, e069_05, e069_06, e069_07, e069_08, e069_11, e069_12, e069_13, e069_14, e069_17, e224, e235, f025, f028b_wvs7,f028, f066_evs5, f034, f050, f051, f053, f054, f114a,f115, f116,f117, f118,f119, f120, f121, f122, f123, f132

[65] A table on WVS imputations is available at ** LINK HERE **

F. Additional Tables and Results

Table 16: Selection into play: culture detailed

	pass count (1)	Total passes in shared mins (2)	pass count (3)
Same nationality (0/1)	0.0285***	0.0154***	0.0445***
	(0.0048)	(0.0027)	(0.0061)
Same colonial legacy (0/1)	0.0283***	0.0127***	0.0408***
	(0.0065)	(0.0036)	(0.0085)
Same federal legacy (0/1)	-0.0227	0.0048	-0.0185
	(0.0151)	(0.0084)	(0.0202)
Same language (0/1)	-0.0047	0.0055	0.0021
	(0.0123)	(0.0065)	(0.0163)
Average length of passes (ln)	-0.7944***		-0.8389***
	(0.0094)		(0.0108)
Average forwardness Ind (0-1)	0.0142		0.1094***
	(0.0099)		(0.0104)
Observations	668,105	668,114	668,105
Pseudo R^2	0.75931	0.86281	0.67156
Passer by half-season fixed effects	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓
Cross-position dummies	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. In column 2, the dependent variable is total pass count by player 1 in minutes when both are fielded. Same cultural background is equality of either cultural aspect (language, colonial legacy or nationality). Total team pass count is captured via team by half-season fixed effects.

Table 17: Results on Robustness

		pass count		pass count (ln)	pass count
	(1)	(2)	(3)	(4)	(5)
	Poisson	Poisson	Poisson	OLS	Poisson
Same nationality (0/1)	0.0269***	0.0249***	0.0265***	0.0238***	0.0298***
	(0.0055)	(0.0048)	(0.0047)	(0.0048)	(0.0051)
Same colonial legacy (0/1)	0.0314***	0.0264***	0.0266***	0.0231***	0.0272***
	(0.0075)	(0.0065)	(0.0064)	(0.0063)	(0.0069)
Same federal legacy (0/1)	-0.0304*	-0.0307**	-0.0237	-0.0093	-0.0294*
	(0.0176)	(0.0150)	(0.0147)	(0.0161)	(0.0154)
Same language (0/1)	-0.0038	-0.0087	-0.0052	0.0152	-0.0147
	(0.0145)	(0.0121)	(0.0120)	(0.0111)	(0.0124)
Shared experience, 1sh+ (0/1)		0.0105*			
		(0.0056)			
Height difference in cm		-0.0126***			
		(0.0004)			
Players value difference, d(ln)		-0.0008***			
		(0.0002)			
Both treated as EU player (0/1)		0.0064			
		(0.0117)			
Passer total passes when together			1.140***	0.2838***	
			(0.0048)	(0.0036)	
Average length of passes (ln)		-0.7824***	-0.7861***	-0.3918***	-0.8141***
		(0.0094)	(0.0091)	(0.0071)	(0.0114)
Average forwardness Ind (0-1)		0.0112	-0.0027	0.3089***	0.2866***
		(0.0098)	(0.0099)	(0.0052)	(0.0115)
Observations	668,105	668,105	668,105	666,230	432,125
Pseudo R^2	0.74290	0.76074	0.76039	0.26121	0.71361
Passer by half-season	✓	✓	✓	✓	✓
Receiver by half-season fixed effects	✓	✓	✓	✓	✓
Cross-position dummies	✓	✓	✓	✓	✓

Column 1-3 Poisson, column 4 OLS regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Same cultural background is equality of either cultural aspect (language, colonial legacy or nationality). Player values are start of the half-season. Total team pass count is captured via team by half-season fixed effects. Both players EU+ reflect national regulations to play, see Appendix B.4. Similar valuation and height: both below/above median.

References

Adams, R. B. and Ferreira, D. (2009), 'Women in the boardroom and their impact on governance and performance', *Journal of financial economics* **94**(2), 291–309. 4

Adams, R. B. and Funk, P. (2012), 'Beyond the glass ceiling: Does gender matter?', *Management Science* **58**(2), 219–235. 4

Ahern, K. R. and Dittmar, A. K. (2012), 'The changing of the boards: The impact on firm valuation of mandated female board representation.', *The Quarterly Journal of Economics* **127**(1), 137–197. 4

AlShebli, B. K., Rahwan, T. and Woon, W. L. (2018), 'The preeminence of ethnic diversity in scientific collaboration', *Nature communications* **9**, 1–10. 13

Anderson, S., De Palma, A. and Thisse, J.-F. (1992), 'Discrete choice theory of product differentiation', *MIT Press* . 22

Andrews, M., Gill, L., Schank, T. and Upward, R. (2012), 'High wage workers match with high wage firms: Clear evidence of the effects of limited mobility bias', *Economics Letters* **117**(3), 824–827. 13, 14

Apesteguia, J., Azmat, G. and Iriberri, N. (2012), 'The impact of gender composition on team performance and decision making: Evidence from the field', *American Economic Review* **58**(1), 78–93. 4

Arcidiacono, P., Kinsler, J. and Price, J. (2017), 'Productivity spillovers in team production: Evidence from professional basketball', *Journal of Labor Econonomics* **35**(1), 191–225. 15

Berge, L. (2018), Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm, Working Paper 13. 27

Bertrand, M. and Duflo, E. (2017), Field experiments on discriminationa, *in* A. V. Banerjee and E. Duflo, eds, 'Handbook of Field Experiments', Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, chapter 10, pp. 309–393. 12

Bobowik, M., Valentim, J. P. and Licata, L. (2018), 'Introduction to the special issue: Colonial past and intercultural relations', *International Journal of Intercultural Relations* **62**, 1–12. 17

Bonhomme, S., Lamadon, T. and Manresa, E. (2019), 'A distributional framework for matched employer employee data', *Econometrica* **87**(3), 699–739. 14

Buchholz, M. (2021), 'Immigrant diversity, integration and worker productivity: uncovering the mechanisms behind 'diversity spillover' effects', *Journal of Economic Geography* **21**(2), 261–285. 11

Burt, R. (1992), *Structural Holes: The Social Structure of Competition*, Harvard University Press. 10

Calder-Wang, S., Gompers, P. A. and Huang, K. (2021), Diversity and performance in entrepreneurial teams, Working Paper 28684, National Bureau of Economic Research. 12, 13

Castilla, E. J. (2011), 'Bringing managers back in: Managerial influences on workplace inequality', *American Sociological Review* . 10

Coleman, J. (1958), 'Relational analysis: The study of social organizations with survey methods', *Human Organization* **17**(4), 28–36. 5

Cross, R. and Cummings, J. (2004), 'Tie and networks correlates of individual performance in knowledge-intensive work', *Academy of Management Journal* **47**, 928–937. 10

Cullen, Z. and Perez-Truglia, R. (2023), 'The old boys' club: Schmoozing and the gender gap', *American Economic Review* **113**(7), 1703–1740. 10

Currarini, S., Jackson, M. O. and Pin, P. (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045. 5, 9, 10

Currarini, S., Jackson, M. O. and Pin, P. (2010), 'Identifying the roles of race-based choice and chance in high school friendship network formation', *PNAS* **107**(11), 4857–4861. 9

Currarini, S. and Mengel, F. (2016), 'Identity, homophily and in-group bias', *European Economic Review* **90**, 40–55. 8, 11, 12, 53

Desmet, K. and Ortuño-Ortín, I.and Wacziarg, R. (2017), 'Culture, ethnicity, and diversity', *American Economic Review* **107**(9), 2479–2513. 4, 17, 38, 39

D'Orsogna, P. and Ottaviano, G. (2011), 'Football Economics and Policy by Stefan Szymanski', *Journal of Economic Literature* **49**(4), 1294–1297. 15

Dovidio, J. F. and Gaertner, S. L. (2010), Intergroup bias, *in* S. T. Fiske, D. Gilbert and G. Lindzey, eds, 'Handbook of social psychology', 5 edn, Vol. 2, Wiley, New York, pp. 1084–1121. 8

Earley, C. P. and Mosakowski, E. (2000), 'Creating hybrid team cultures: An empirical test of transnational team functioning', *Academy of Management Journal* **43**(1), 26–49. 14

Eeckhout, J. and Kircher, P. (2011), 'Identifying Sorting—In Theory', *The Review of Economic Studies* **78**(3), 872–906. 14

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T. and Ven, J. V. D. (2023), 'Cognitive biases: Mistakes or missing stakes?', *Review of Economics and Statistics* **105**(4), 1–15. 47

Ertug, G., Brennecke, J., Kovacs, B. and Zou, T. (2021), 'What does homophily do? a review of the consequences of homophily', *Academy of Management Annals* . 4, 8, 9, 10

Fally, T. (2015), 'Structural gravity and fixed effects', *Journal of International Economics* **97**(1), 76–85. 27, 29

Feld, S. L. (1981), 'The focused organization of social ties', *American Journal of Sociology* **86**, 1015–1035. 10

Feld, S. L. (1982), 'Social structural determinants of similarity among associates', *American Journal of Sociology* . 10

Feld, S. L. (1984), 'The structured use of personal associates', *Social Forces* **62**, 640–652. 10

FIFA (2021), *Commentary on the Rules Governing Eligibility to Play for Representative Teams*, FIFA: International Federation of Association Football. 64

Freeman, R. B. and Huang, W. (2015), 'Collaborating with People Like Me: Ethnic Coauthorship within the United States', *Journal of Labor Economics* **33**(S1), 289–318. 13

Gauriot, R. and Page, L. (2019), 'Fooled by performance randomness: Overrewarding luck', *The Review of Economics and Statistics* **101**(4), 658–666. 15

Giannetti, M. and Yafeh, Y. (2012), 'Do cultural differences between contracting parties matter? evidence from syndicated bank loans', *Management Science* **58**(2), 365—383. 53

Head, K. and Mayer, T. (2014), Gravity equations: Workhorse, toolkit, and cookbook, *in* G. Gopinath, E. Helpman and K. Rogoff, eds, 'Handbook of international economics', Elsevier, chapter 3, pp. 131–195. 9, 19

Hinz, J., Stammann, A. and Wanner, J. (2021), State Dependence and Unobserved Heterogeneity in the Extensive Margin of Trade, CEPA DP 36, Center for Economic Policy Analysis. 27

Hjort, J. (2014), 'Ethnic divisions and production in firms', *The Quarterly Journal of Economics* **129**(4), 1899–1946. 12, 14

Horwitz, S. K. and Horwitz, I. B. (2007), 'The effects of team diversity on team outcomes: A meta-analytic review of team demography', *Journal of management* **33**(6), 987–1015. 10

Ingersoll, K., Malesky, E. J. and Saiegh, S. M. (2017), 'Heterogeneity and team performance: Evaluating the effect of cultural diversity in the world's top soccer league', *Journal of Sports Analytics* **3**(2), 67–92. 14

Jackson, M. O., Rogers, B. W. and Zenou, Y. (2017), 'The economic consequences of social-network structure', *Journal of Economic Literature* **55**(1), 49–95. 8, 9, 10, 11, 51, 53

Jackson, S. E., Joshi, A. and Erhardt, N. L. (2003), 'Recent research on team and organizational diversity: SWOT analysis and implications', *Journal of Management* **29**(6), 801–830. 3, 11

Joshi, A., Labianca, G. and Caligiuri, P. M. (2002), 'Getting along long distance: understanding conflict in a multinational team through network analysis', *Journal of World Business* **37**(4), 277–284. 3

Kahane, L., Longley, N. and Simmons, R. (2013), 'The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and Language Diversity in the National Hockey League', *The Review of Economics and Statistics* **95**(1), 302–314. 12, 14

Keane, M. P., Todd, P. E. and Wolpin, K. I. (2011), The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications, *in* O. Ashenfelter and D. Card, eds, 'Handbook of Labor Economics', Vol. 4, Elsevier, pp. 331–461. 22

Keane, M. and Wolpin, K. I. (2009), 'Empirical applications of discrete choice dynamic programming models', *Review of Economic Dynamics* **12**(1), 1–22. 22

Kleven, H. J., Landais, C. and Saez, E. (2013), 'Taxation and international migration of superstars: Evidence from the european football market', *The American Economic Review* **103**(5), 1892–1924. 15

Kovacs, B. and Kleinbaum, A. M. (2020), 'Language-style similarity and social networks', *Psychological Science* **31**(2), 202–213. 57

Lang, K. (1986), 'A language theory of discrimination', *Quarterly Journal of Economics* **101**(2), 363–382. 11

Laurentsyeva, N. (2019), From friends to foes: National identity and collaboration in diverse teams, Working Paper 226. 12

Lawrence, B. S. and Shah, N. P. (2020), 'Homophily: Measures and meaning', *Academy of Management Annals* **14**(2), 513–597. 2, 4, 5, 8, 9, 10

Lazarsfeld, P. F. and Merton, R. K. (1954), Friendship as a social process: A substantive and methodological analysis, *in* M. Berger, ed., 'Freedom and Control in Modern Society', Van Nostrand, New York, pp. 18–66. 8, 9

Lazear, E. (1999*a*), 'Language and culture', *Journal of Political Economy* **107**(6), S95–S126. 1, 11

Lazear, E. P. (1999*b*), 'Globalisation and the market for team-mates', *The Economic Journal* **109**(454), 15–40. 1, 11

Lix, K., Goldberg, A., Srivastava, S. B. and Valentine, M. A. (2022), 'Aligning differences: Discursive diversity and team performance', *Management Science* **68**(11), 8430—-8448. 13

Marsden, P. V. (1987), 'Core discussion networks of americans', *American Sociological Review* **52**(1), 122–131. 5

Matsa, D. A. and Miller, A. R. (2013), 'A female style in corporate leadership? evidence from quotas', *American Economic Journal: Applied Economics* **5**(3), 136–169. 4

McFadden, D. (2001), 'Economic choices', *American Economic Review* **91**(3), 351–378. 22

McFadden, D. and Train, K. (2000), 'Mixed mnl models for discrete response', *Journal of Applied Econometrics* **15**(5), 447–470. 22

McFarland, D. A., Moody, J., Diehl, D., Smith, J. A. and Thomas, R. J. (2014), 'Network ecology and adolescent social structure', *American Sociological Review* **79**(6), 1088–1121. 13

McPherson, J. M. and Smith-Lovin, L. (1987), 'Homophily in voluntary organizations: Status distance and the composition of Face-to-Face groups', *American Sociological Review* **52**(3), 370–379. 5, 13

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001), 'Birds of a feather: Homophily in social networks', *Annual Review Sociology.* **27**(1), 415–444. 2, 5, 8, 9, 10

Melitz, J. and Toubal, F. (2014), 'Native language, spoken language, translation and trade', *Journal of International Economics* **93**(2), 351–363. 38, 39, 41

Neeley, T. (2015), 'Global teams that work', *Harvard Business Review* . 1

Nüesch, S. and Haas, H. (2013), 'Are multinational teams more successful?', *International Journal of Human Resource Management* **23**(15), 3105–3115. 3, 14

Opper, S., Nee, V. and Brehm, S. (2015), 'Homophily in the career mobility of china's political elite', *Social Science Research* . 10

Ottaviano, G. I. and Peri, G. (2005), 'Cities and cultures', *Journal of Urban Economics* **58**(2), 304–337. 11

Ottaviano, G. I. and Peri, G. (2006), ‘The economic value of cultural diversity: evidence from US cities’, *Journal of Economic Geography* **6**(1), 9–44. 11

Parsons, C. A., Sulaeman, J., Yates, M. C. and Hamermesh, D. S. (2011), ‘Strike three: Discrimination, incentives, and evaluation’, *American Economic Review* **101**(4), 1410–1435. 14

Petrin, A. and Train, K. (2010), ‘A control function approach to endogeneity in consumer choice models’, *Journal of Marketing Research* **47**(1), 3–13. 22

Pettigrew, T. F. and Tropp, L. R. (2006), ‘A meta-analytic test of intergroup contact theory’, *Journal of Personality and Social Psychology* **90**(5), 751–783. 53

Porter, J. and Washington, R. (1993), ‘Minority identity and self-esteem’, *Annual Review of Sociology* **19**, 139–161. 51

Price, J. and Wolfers, J. (2010), ‘Racial Discrimination Among NBA Referees’, *The Quarterly Journal of Economics* **125**(4), 1859–1887. 14, 47

Santos-Silva, J. and Tenreyro, S. (2022), ‘The log of gravity at 15’, *Portuguese Economic Journal* **21**, 423–437. 27

Spolaore, E. and Wacziarg, R. (2016), Ancestry, language and culture, *in* ‘The Palgrave Handbook of Economics and Language’, Springer, pp. 174–211. 4, 17, 39

Szymanski, S. (2010), *Football Economics and Policy*, Springer. 15

Terenzini, P. T., Cabrera, A. F., Colbeck, C. L., Bjorklund, S. A. and Parente, J. M. (2001), ‘Racial and ethnic diversity in the classroom’, *Journal Higher Education* **72**(5), 509–531. 8

Thaler, R. (1986), ‘The psychology and economics conference handbook: Comments on simon, on einhorn and hogarth, and on tversky and kahneman’, *Journal of Business* **59**(4), S279–S284. 47

Todd, P. and Wolpin, K. I. (2010), ‘Structural estimation and policy evaluation in developing countries’, *Annual Review of Economics* **2**, 21–50. 22

Tovar, J. (2020), ‘Performance, Diversity And National Identity Evidence From Association Football’, *Economic Inquiry* **58**(2), 897–916. 2, 14

Weidner, M. and Zylkin, T. (2021), ‘Bias and consistency in three-way gravity models’, *Journal of International Economics* **132**, 103513. 27

Zipf, G. K. (1949), *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA. 10