Collaboration and Homophily in Global 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

How do cultural barriers related to nationality, colonial legacy and language affect collaboration in multinational teams? We address this question by creating and exploiting an exhaustive dataset recording all 10.7 million passes by 7 thousand professional European football players from 132 countries fielded by all 154 teams competing in the top 5 men leagues over 8 sporting seasons, together with full information on players' and teams' characteristics. We use a discrete choice model of players' passing behavior as a baseline to separately identify collaboration due to biased cultural preferences ('choice homophily') from collaboration due to opportunities ('induced homophily'). The outcome we focus on is 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 in a half-season. We find strong evidence of choice homophily. Relative to the baseline, player pairs of same culture have a 2.50 percent higher pass rate. Passes between players of same nationality, same colonial legacy and same language are associated respectively with 3.01, 2.34 and 0.66 percent higher pass rates. These findings show that choice homophily based on culture is pervasive 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.

Keywords: Organizations, teams, culture, homophily, diversity, language, globalization, border effect, big data, text as data, football

JEL-codes: J15, J44, J71, F23, L83, C81

First Draft: 14 January 2022. See latest version HERE.

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

**Bekes thanks the support of the 'Firms, Strategy and Performance' Lendület grant of the Hungarian Academy of Sciences. We thank Gábor Kézdi, Miklós Koren, Alice Kugler, Balázs Muraközy, Balázs Lengyel, Ádám Szeidl and seminar participants at CERS-HAS, University of Reading, IMT Lucca, and CEU for useful comments and suggestions. We are grateful to Endre Borza, Bence Szabo for outstanding and extensive research assistance.

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. On the one hand, a multinational workforce allows companies to build teams that offer the best expertise 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 face several hurdles. When team members have from different cultural background, communication can rapidly deteriorate, misunderstanding can ensue, and cooperation can degenerate into distrust.

In this paper we systematically investigate how barriers related to diversity in cultural background affect collaboration in multinational teams (Lazear, 1999*b*,*a*). We define 'collaboration' as the situation of two or more people working together to create or achieve the same thing (Cambridge Dictionary), and study teams that are not geographically dispersed, as is often the case in multinational companies, since dispersion may *per se* inhibit collaboration (Joshi et al., 2002). 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 language as well as norms, values and attitudes that are transmitted intergenerationally, which we proxy through nationality and colonial legacy. We show that a 'border effect' between team members of different culture may indeed hamper collaboration, pretty much like different language and colonial past hamper international trade in goods and services between the regions of different countries (Head and Mayer, 2014). Team members of same culture collaborate more than team members of different culture.

We base our investigation on the unique features of 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 full information on players' and teams' characteristics as well as their performances.[1] The dataset records 10.7 million passes made in 14,608 games by 7,000 players from 132 countries fielded by 154 teams. Passes are aggregated by passer-receiver player pairs and half-seasons as these are time periods in which football clubs have stable squads. The analysis is then carried out on the resulting three-dimensional (unbalanced) panel of 669 thousand passer-receiver pairs over 16 half-seasons. We measure collaboration as the average number of passes per minute between a pair of players in a half-season (which we call their 'pass rate'), and study how it is affected by the pairs' nationalities, languages and colonial legacies. Passes are the essential building blocks of football. They represent how players work together for the common objective of scoring or preventing the opponent from scoring a goal. Importantly, passes are

[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) .

also positively correlated with winning more league points and achieving higher league standings.[2]

This type of sports data has several advantages. First, the European football industry is very globalized: fans are spread around the world, and multinational teams are the rule in the top five leagues.[3] Second, 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. Third, 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). Fourth, 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. Fifth, while team composition is exogenous to players' decisions, collaboration with other team members is mostly up to their individual choices. Sixth, 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.

All these features allow us to 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 from issues of gender diversity.[4]

We are not the first to exploit team sports data to analyze the potential gains and losses from employing culturally diverse work teams. In the case of the top North

[2]In our dataset, regressing points per game (in levels) on log total passes, team and league by half-season fixed effects and conditioning on league specific aggregate trends shows that, in a half-season when a team passes 10% more than its average pass frequency across half-seasons, it wins 0.025 points (or 2.1%) more than its average points across half-seasons. Over a typical league's season of 38 games, this sums up to 1 point (compared to an average of 50 points per team in a season). This difference is equivalent to one position difference in final standings. See Appendix D for details.

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

[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.

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 the 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. 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 and we can measure it accurately through the pass data[5].

The key methodological challenge that our investigation faces has been highlighted in the studies on homophily, defined as the tendency to associate with similar others (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 underlying 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

[5]Beyond diversity, team sports data are increasingly used to study various issues in labor and public economics. For example, Parsons et al. (2011) exploit data from the top North American baseball league (Major League Baseball) on the way umpires judge 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.

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

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 during the matches they play together in a given time period. It is implemented empirically by a Poisson regression with player characteristics as well as player-time fixed effects as controls. 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. Specifically, the outcome we focus on is 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. In a regression with passer by half-season fixed effects as well as receiver by half-season fixed effects, conditioning on pass features (such as length) shows that player pairs of same culture have a pass rate 2.50 percent higher than player pairs of different culture. Accordingly, sharing the same cultural background is about as likely to lead to more passes as doubling the player pair's valuation (a consensus measure of their skills). As for the different cultural traits, passes between players of same nationality, same colonial legacy and same language are associated respectively with a pass rate that is 3.01, 2.34 and 0.66 percent higher than the pass rate between players without shared language, shared colonial legacy and shared nationality. Choice homophily is more pronounced for complex pass sequences in which the ball goes back and forth between a given player pair. For these sequences, the pass rate for pairs of same culture is 5.0 percent higher than for pairs of different culture, compared with 2.1 percent for single passes. Shared experience does not affect these results: once individual experience is controlled for, shared experience is irrelevant. This suggests that culture does not simply capture knowing each other.

These findings show that homophily based on cultural traits is pervasive even in

[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).

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 language intensive.

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. Section 5 presents the model estimation, whose results are then discussed in Section 6. Section 7 concludes.

2. Related literature

This paper is related to various research streams of the vast literature on cultural diversity and performance in teams, which spans from management (see e.g. Earley and Mosakowski (2000) and (Jackson et al., 2003)) to education studies (see e.g. Terenzini et al. (2001)).

Four streams are particularly relevant to what we do. The first is concerned with 'diversity spillovers', which improve team performance in a diverse environment, but not necessarily in a team that is itself diverse (Ottaviano and Peri, 2006, 2005). This stream highlights four main mechanisms (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) through increasing the specialization, variety of skills and 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'), (iv) when simply through exposure to a diverse range of knowledge and approaches to problems workers learn and become more productive ('exposure effects'). The evidence on US Metropolitan Statistical Area reported in Buchholz (2021) supports exposure effects as the main mechanism, 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). 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 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 are typically based on field experiments (Bertrand and Duflo, 2017). 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[7].

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. Moreover, the flower plant and the football team setups have different pros and cons. The former can exploit an essentially random rotation process to assign workers to positions for identification, but its external validity may be limited. In the latter setup rotation is arguably not random as it depends on the manager's choices, but the richness of information from which to obtain all sorts of individual and team controls makes the case for external validity stronger. Be it as it may, non-random rotation due to endogenous team formation leads to known biases. 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. 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. 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. 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.

The third research stream analyzes homophily in scientific publications. Looking into scientific papers written by US-based authors from 1985 to 2008, Freeman and

[7]Conflicts exacerbate discrimination. 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.

Huang (2015) find evidence of choice homophily as persons of similar ethnicity co-author together more frequently than predicted by their proportion among authors; and that 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. To further isolate the effects of ethnic diversity, they use randomized baseline models and again find a clear link between diversity and impact. Differently from these studies, we use a discrete choice model rather than randomized models to separate choice homophily from induced homophily.

Finally, the fourth research stream is concerned with the formation of social networks (Jackson, 2008). In particular, Currarini et al. (2009) and Currarini et al. (2010) 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. Differently from these studies, our baseline is derived from a discrete choice model in which forward-looking behavior allows us to highlight the role of biased preferences ('passing to teammates') after netting out also the implications of biased meeting rates ('passing to teammates through teammates') through the model's structure and the dataset's richness.

3. Data

In this section, we'll describe the scope of the unique data we use and briefly summarize how the data was collected, cleaned and transformed for the purpose of our analysis. After an overview, we'll separately describe the players' data, the dataset on passing events, and the combined estimation dataset used for the model estimation.

3.1. Overview

To estimate how homophily may affect collaboration we collected, curated, and cleaned an immense dataset, we call "raw football dataset". This dataset is a collection of several data tables, covering events such as passes, as well as player characteristics. Data in the raw football dataset has been collected by webscraping – an automated data collection algorithm that extracts information from websites. The output of these algorithms is structured text, which we parsed – transformed – into tabular data. In

what follows we describe the tabular versions, where each observation is an event or a player.

The combined dataset consists of all passes made by professional European football players in the top five men leagues over eight sporting seasons, together with full information on players' and teams' characteristics as well as their performances. It is a relational dataset linked via player names and additional information[8].

The top five leagues are the German Bundesliga, the French Ligue 1, the Spanish La Liga, the English Premier League, and the Italian Serie A. We have selected these leagues 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, for which data quality is the highest. (These were not interrupted by the coronavirus.) 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.

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. Exact cutoff-dates vary by leagues. The Premier League, La Liga, Serie A and Ligue 1 are all composed of 20 teams (playing 20×19=380 games), 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*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.

Data quality and coverage are both very high in our datasets. Nevertheless, a few small data cleaning steps were needed and we discuss these issues discussed in Appendix B.

3.2. The players dataset

The information on players and their characteristics are compiled by Transfermarkt[9]. 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. They also include a player's estimated transfer value, that is, 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, how valuable he may be to another team. As such, a player's transfer value is considered a consensus measure of the quality of his football skills. Transfer values are estimated twice a year in correspondence with the transfer windows.

We have 6,998 players in our sample, for whom we can fully map their entire career,

[8]See replication options in Appendix E.

[9]See https://www.transfermarkt.com/

with a typical team relying on a squad of about 30 players.

European football is truly globalized as there are players from 132 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.

Table 1: Most frequent nationalities

Country	share (%, all players)
Spain	13.5
France	12.0
Italy	9.8
Germany	8.4
England	7.0
Brazil	4.4
Argentina	3.4
Portugal	1.8
Netherlands	1.6
Senegal	1.6
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.

3.3. Determining shared cultural background

To determine whether two players have the same culture, we consider three cultural traits: nationality, language and colonial legacy.

First, nationality is defined based on citizenship of a country. As some players have multiple citizenships, we define two players as co-national if they share at least one of them, or have the same country of birth.

Second, for language we rely on Melitz and Toubal (2014) to ascertain whether or not two countries share one or more common official and widely spoken languages. For example, the official and widely spoken languages in Morocco are Arabic and French. Accordingly, players from Morocco and Egypt share Arabic, those from Morocco and

France share French, while those from Egypt and France do not share any language. We then assume that a player speaks the official and widely spoken languages of his country of citizenship at the beginning of his career as mother tongues (so a Moroccan player speaks Arabic and French as 'mother tongue'). Hence, a player starting his career in Morocco shares a common language ('mother tongue') with players starting their careers in Egypt or France, whereas a player starting his career in Egypt (France) shares a common language ('mother tongue') with players starting their careers in Morocco but not with those starting their careers in France (Egypt).

Third, to ascertain common colonial legacy, we use colonial links data from CEPII as Head and Mayer (2014). For some players (such as South American and Spanish, Brazilian and Portuguese, French and Senegalese) same colonial legacy subsumes same language. Other players (such Croatian and Serbian, Belgian and French, German and Austrian) share the same language but do not share the same colonial legacy. We consider languages that are very close, even if not identical (such as Serbian and Croat, Czech and Slovakian, or Dutch and Flemish) as one language.

Multiple citizenships pose a challenge in determining same language or same colonial legacy as well. Similarly to nationality, we say that two players speak share colonial legacy if they have nationalities that are linked via colonial legacy. We say players speak the same language only, if they have a nationality sharing the same language, but no shared nationality or one sharing colonial legacy. Thus, we have categorical variable on culture with four categories.

1. Same nationality (e.g. two Argentinian players)
2. No same nationality, but same colonial legacy (e.g. Argentina and Spain)
3. No same nationality, no same colonial legacy, but same language (e.g. Belgium and France)
4. No same nationality, no same colonial legacy and no same language (e.g. Argentina and Brazil).

In our dataset, 37.8% of the players have the same nationality, 8.0% have the same colonial legacy but different nationality, and 4.1% have the same language but different colonial legacy and different nationality. We define all these players as having the same culture. According to this definition, 49.9% of the players in our sample have the same culture, whereas 50.1% of them do not.

3.4. The event feed and passes datasets

Information on passes comes from OPTA, a sports technology company, and is available via third party sites such as whoscored.com. OPTA's data are generated by people watching games and coding events helped by cameras. Events are recorded one-by-one and cross-checked to create highly reliable and widely used depiction of games. This is called the "event feed data". Our event dataset has been scraped from a third party website.

A pass is an event defined as "any pass attempted from one player to another", including free kicks, corners, throw-ins, goal assists[10]. The event feed data include a timestamp, team id, x and y coordinates of where the pass took place in the pitch, and its outcome (e.g., a flag for a successful pass). Using the next event, the receiver and his coordinates can be recovered. We filtered the event feed data on passes only and call it the "passes dataset".

Thus, one row in the passes dataset has the following important columns: passer (and receiver) player ID and his team ID, position information on the pass, time into the game, nature of the pass. There are 17 passes per minute, on average, and there are about 800 successful passes on average per game, so we have 10.73 million passes.

First we have aggregated the passes data to single games to generate variables such as the sum of passes between any two players in a game. We have then aggregated these observations by half-season. This division is suggested 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 halves: 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.

Under the reasonable assumption that, 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, we drop the corresponding player pairs. Only 7.8% of the observations have zero passes - this happens when player P1 passes to player P2, but there is no pass from P2 to P1.

With respect to aggregation, half-seasons have several advantages compared to seasons or games. In a single game, two players may not play together for various reasons as squads are large and only eleven players can be fielded at the same time. This may raise a selection issue, which can be tackled by considering all the games in a season instead. If two players never pass to each other during an entire season, either they are never fielded together or one can safely assume they are never fielded in positions that interact. Moreover, considering a half-season allows one to investigate the role of common experience as players who spend more time together on the pitch may learn to pass more.

On the other hand, the presence of the winter transfer window implies that during a season a team's squad may change composition. Our assumption of unchanged player quality makes more sense in a half-season than in a season, especially as younger players may evolve. In this respect, a half-season strikes a balance between mitigating the selection issue and keeping squad composition fixed. Finally, 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.

[10]See https://www.statsperform.com/opta-event-definitions.

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	310,493
-	yes	yes	player-pair's number of passes in half-season, shared experience with club	669,025

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.

3.5. Combination: creating the estimation dataset

The final task to prepare our estimation dataset is to combine player information and pass 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 an essential albeit difficult task. First, there are players who are recorded differently across datasets - especially when they have diacritical marks (such as "é"), are translated from a non-latin alphabet, or have many middle names. Second, players may have the same name, especially with frequent family names. To solve this issues, we developed a matching algorithm based on player names and additional information. The procedure is detailed in Appendix B.

We have also made a few decisions regarding data cleaning: dropped players who only had a single passing partner or those we could not identify. These minor steps are detailed in B. All results are robust to making them.

Variables may be aggregated at different levels, and Table 2 presents examples for the four different types of variables.

We will call this combined and aggregated passes and player dataset the "estimation dataset". It is a directional pass level dataset, which means, we have a separate row for player 1 as passer and player 2 as receiver and player 2 as passer and player 1 as receiver. All features, such as pass counts, are aggregated to the half-season level. This gives us 669,025 observations at the passer, receiver and half-season level.

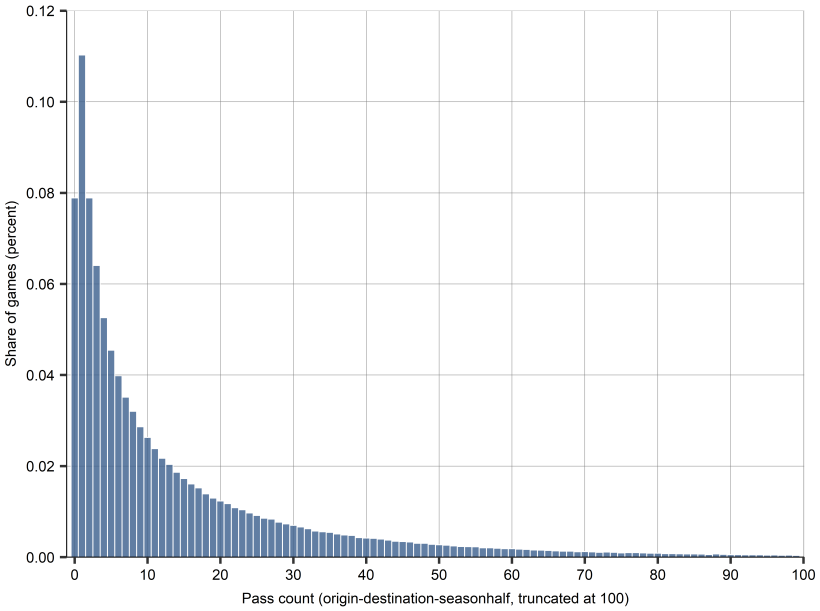
In a half-season, a passer makes a total of 650 passes on average (ranging between 2 and 3750, median is 570). On average, he has 19.31 receiver partners (ranging between 2 and 35, median is 20).

Taking the estimation dataset (N=669,025), the average pass count from passer to receiver is 15.92 (ranging between 0 and 488, median is 8)[11]. The distribution looks

[11]Zero pass comes in when we added them for symmetry: player A passed to B, but not the other way around.

highly skewed to the right as shown in Figure 1. (The distribution is truncated at 100 (98.63% of observations) for better visibility.)

Figure 1: Distribution of passes



4. A Discrete Choice Model of Passing Behavior

A crucial challenge in assessing how common nationality and common language affect 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 unrelated 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, valuation, citizenship) and pass features (such as average distance).

Consider a football team of $N = 11$ players, indexed from 1 to N, engaged in a half-season consisting of P passing episodes.[12] 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. Let us focus on two players, labeled o and d, and on the subset of passing episodes $T^{o,d}$ in which both players are on the pitch

[12]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.

with player o having ball possession. 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. We are interested in characterizing the probability that player o passes to player d rather than to any of the other nine teammates.

A passing episode consists of two periods: when the pass is initiated by o (t) and when the pass is received by d ($t+1$). The passer 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 ability and position of that player, and by some randomness due to the vagaries of the game. These may include, for instance, the performance of the opposing team, the referee's decisions or the weather conditions. We use $\ln u_t^d$ to denote the deterministic part of the team's benefit as determined by player d's characteristics, and z_t^d to denote the realization of its random part ('shock') due to match contingencies. Specifically, for each receiver d, z_t^d is the realization of a random variable Z with continuous differentiable c.d.f. $\Pr[Z \leq z] = G(z)$ over the support $(-\infty, +\infty)$. Passer o also understands the challenges he faces in passing the ball to receiver d. We call $\widetilde{c}^{o,d}$ the associated 'passing cost' capturing such challenges in terms of physical and mental effort. In particular, this cost may be high if the pass is difficult due to the positions of players and their reciprocal distance or o and d find it hard to collaborate due to different cultural traits. Finally, passer o realizes the difficulty receiver d may face in taking control of the ball, which depends on the receiver's characteristics. We use φ^d to denote the probability that receiver d takes control of the ball. We call this the probability of a successful pass. Hence, any difference in outcomes across the $T^{o,d}$ passing episodes ultimately depends on different success of attempted passes and different realizations of the shock due to match contingencies.

The passer's decision can be characterized as the problem of passing the ball to the receiver who generates the highest expected benefit for the team. The value function of this problem is written recursively as

$$U_t^o = \ln u_t^o + \max_{\{d\}_{d=0}^N} \left\{ \beta \varphi^d E\left[U_{t+1}^d\right] - \widetilde{c}^{o,d} + z_t^d \right\} \tag{1}$$

where the team's benefit U_t^o of controlling the ball in period t is split into two components: the benefit of player o currently controlling the ball (e.g. in the current period the player could try to score a goal; or he could decide to kick the ball out of play to allow the team to reorganize) and the option value of player o passing (or keeping control of) the ball at the beginning of the future period. These two components correspond to $\ln u_t^o$ and $\max_{\{d\}_{d=0}^N} \left\{ \beta \varphi^d E\left[U_{j+1}^d\right] - \widetilde{c}^{o,d} + z_t^d \right\}$ respectively, with expectation $E\left[U_{t+1}^d\right]$ taken over the realizations z_t^d of the shock. The parameter $\beta \in [0,1]$ measures the relative importance the team attaches to passing in general, independently from 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 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.

We assume that the random variable Z follows the 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 $\kappa > 0$. 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. Smaller κ then implies more intense training to reduce variation in their individual assessments. The Gumbel assumption leads to a simple expression for the probability of player o passing to teammate d in period t. Specifically, after having taken expectations on both sides of (1), defining $V_t^o \equiv \exp E[U_{t+1}^o]$ and $c^{o,d} = \exp \widetilde{c}^{o,d}$ allows one to express the *ex ante* probability that player o in possession of the ball in period t successfully passes to teammate d at the beginning of period $t+1$ as

$$\pi_t^{o,d} = \left(V_{t+1}^d\right)^{\kappa\beta\varphi^d} \left(c^{o,d}\right)^{-\kappa} \left(\Lambda_{t+1}^o\right)^{-\kappa} \text{ with } \Lambda_{t+1}^o \equiv \left[\sum_{s=1}^{N} \left(V_{t+1}^s\right)^{\kappa\beta\varphi^s} \left(c^{o,s}\right)^{-\kappa}\right]^{\frac{1}{\kappa}}, \qquad (2)$$

which *ex post* becomes (approximately) the average share of successful passes that player o makes to player d per episode over 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.[13] The probability that player o successfully passes the ball to player d in period t is thus determined by the team's expected benefit from player d controlling the ball in period $t+1$ (V_{t+1}^d), his probability of taking control of the ball (φ^d) and the difficulty of passing the ball to him ($c^{o,d}$), relative to the team's average benefit from its players $s = 1, ..., N$ controlling the ball in period $t+1$ (V_{t+1}^s), weighted by their probability of taking control of the ball (φ^s) and the difficulty of passing the ball to them ($c^{o,s}$).

In the data we observe the total number of team passing episodes (P), the number of passing episodes involving a pass from o to d ($P^{o,d}$), and the number of passing episodes when both o and d are fielded and player o has ball possession ($T^{o,d}$) over a half-season. If we define the half-season 'pass rate' as $p^{o.d} = P^{o,d}/P$, the model then implies $p^{o,d} = T^{o,d}\pi_t^{o,d}/P$, and thus

$$\log p^{o,d} = \log T^{o,d} + \log\left(\Lambda_{t+1}^o\right)^{-\kappa} + \log\left(V_{t+1}^d\right)^{\kappa\beta\varphi^d} + \log\left(c^{o,d}\right)^{-\kappa} - \log P. \qquad (3)$$

The distinction between distance and culture related challenges in passing from player

[13] The fact that also $s = o$ is included in the sum $\sum_{s=1}^{N} \left(V_{t+1}^s\right)^{\kappa\varphi^s} \left(c^{o,s}\right)^{-\kappa}$ implies $\sum_{s=1}^{N} \pi_t^{o,s} = 1$.

o to player d embedded in $c^{o,d}$ can be made explicit by specifying the bilateral passing cost multiplicatively as

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

In (4) $g^{o,d}$ is the physical distance between the two players' positions so that $\left(g^{o,d}\right)^{\gamma}$ captures all distance-related frictions that make it difficult 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 difficult 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.

Substituting (4) into (3) gives

$$\log p^{o,d} = \log T^{o,d} - \kappa \log \Lambda^{o}_{t+1} + \kappa\beta\varphi^{d} \log V^{d}_{t+1} - \kappa\gamma \log g^{o,d} - \kappa\lambda \log l^{o,d} - \log P \tag{5}$$

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, two remarks are in order. First, (5) distinguishes the role of biased preferences ('passing to teammates'), which work through $\log l^{o,d}$, from the implications of biased meeting rates ('passing to teammates through teammates'), which work through the forward looking terms $\log \Lambda^{o}_{t+1}$ and $\log V^{d}_{t+1}$. Second, such distinction allows us to argue that in (5) the cultural dimensions of $l^{o,d}$ determine choice homophily (i.e. biased preferences), while induced homophily is determined by all other terms on the right hand side of (5).

5. Model estimation

Before discussing details of estimating choice homophily from our theoretical model, we will first estimate chance and choice homophily as an unconditional average within team (and half-season) difference. We estimate this relative difference with a Poisson model:

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = exp(\gamma SameCult_{p1,p2} + \nu_{team,t}) \tag{6}$$

We implement our discrete choice model empirically by comparing the 'pass rate' for a passer-receiver pair by the same culture indicator variable ($g^{o,d}$ in the model) in the estimation dataset.

In detail, let us index the passer as $p1$, the receiver as $p2$, and half-season time period as t. The pass count from passer to receiver in half-season t is then given by $pass\_count_{p1,p2,t}$. Passes from $p1$ to $p2$ are considered relative to the total number of passes by $p1$ when both $p1$ and $p2$ are on the pitch denoted as $T_{p1,p2,t}$. Accordingly, the pass rate is defined as:

$$pass\_rate_{p1,p2,t} = \left(\frac{pass\_count_{p1,p2,t}}{T_{p1,p2,t}} \right)$$

Our theoretical model needs a generalized linear estimator with a log link function. In such setup, there is a large literature on the benefits of preferring a Poisson model to a log(count) model with a large number of fixed effects[14]. While the Poisson model seems the best choice for us, robustness checks will be provided with the log(count) as outcome.

To estimate our model, we use a Poisson model with the pass count ($pass\_count_{p1,p2,t}$) as dependent variable. We have an exposure variable for the total passes by the passer while both players are on the pitch ($\ln T_{p1,p2,t}$). The log of this exposure variable as handled as an offset variable in the regression (ie it's coefficient is forced to be 1). The use $team \times half\_season$ fixed effects will also absorb the total number of passes per team in a half-season (P in the theoretical model).

Note that in the model, the relative value of passing (to anyone) compared to using the ball is captured via the parameter β. Empirically *beta* depends on both team and player characteristics, especially on the position and style of players and tactics used by the team. These characteristics may actually change over time, in our case, may vary by half-seasons. Player characteristics and team fixed effects, and later on player fixed effects will soak up β as well.

The passing friction ($PassFric$) beyond player characteristics will come the average distance between passes ($PassDist$), the share of passes with a forward direction ($Forwardness$). We acknowledge that these two variables may actually be a mechanism instead of a confounder and we show results with and without them. In all models we will have a set of dummies for each pair of passer and receiver broad positions (such as forward to defender) ($position_{p1}position_{p2}$). These are not play specific, and set to control the general system of football.

$$PassFric_{p1,p2,t} = \lambda_1 PassDist_{p1,p2,t} + \lambda_2 Forwardness_{p1,p2,t} + \eta position_{p1}position_{p2}$$

We estimate three versions of our model. First, we consider a version of the Poisson model with a variety of player characteristics as controls:

[14]For a discussion with regards to gravity models in international trade literature and the use of fixed effects Poisson Pseudo Maximum Likelihood estimation (FE-PPML) methods, see Fally (2015) and Santos-Silva and Tenreyro (2021). The procedure we use is described in (Berge, 2018). In particular, 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. See (Hinz et al., 2021). It turns out that FE-PPML estimates can manage this type of bias better than non-linear OLS - one more reason for our preference of the model (Santos-Silva and Tenreyro, 2021). Weidner and Zylkin (2021) shows that the Poisson model still leave some room for potential bias, but with no double player fixed effects and a large number of observations, the bias should be small in our case.

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = exp(\beta SameCult_{p1,p2} + PassFric_{p1,p2,t}+ \\ + 1\ln T + \sum_{j=1}^{2}\left(\eta_j value_{p(j)),t} + \theta_j playerchar_{p(j)),t}\right)) \quad (7)$$

where $SameCult_{p1,p2}$ is the same culture indicator , $PassFric_{p1,p2,t}$ is a set of distance and position-dependent passing frictions. Potential confounding variables are captured for both players, with player valuation ($value_{p(j)),t}$) and other player characteristics ($playerchar_{p(j)),t}$): age, height, age, time since with club (in days), and binary indicator for being on, as well as $position \times half\_season$ dummies, and $nationality \times half\_season$ dummies. In addition, we have $team \times half\_season$ dummies. Our exposure variable is the number of passes by the passer while both players were on the pitch ($\ln T$).

Second, we estimate another Poisson model with fixed effects.

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = exp(\beta SameCult_{p1,p2} + PassFric_{p1,p2,t}+ \\ + 1\ln T_{p1,p2,t} + \gamma_{p1,t} + \gamma_{p2,t}) \quad (8)$$

where $SameCult_{p1,p2}$ is the same culture indicator, $PassFric_{p1,p2,t}$ is a set of distance- and position-dependent passing frictions, $\gamma p1,t$ and $\gamma p2,t$ are ($p_1 \times half\_season$ and $p_2 \times half\_season$) fixed effects [15]. Our exposure variable is the number of passes by the passer while both players were on the pitch ($\ln T$).

Cultural distance, $SameCult_{p1,p2}$ may thus be considered as the sum of three mutually exclusive binary variables: $SameNat$ as same nationality=1, $SameCol$ as colonial history=1 and same nationality=0, and $SameLan$ as same language=1 but colonial history=0 and same nationality=0. It is time-invariant (does not depend on the half-season.)

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = \beta_1 SameNat_{p1,p2}+\beta_2 SameCol_{p1,p2}+\beta_3 SameLan_{p1,p2}+ \\ + PassFric_{p1,p2,t} + 1\ln T_{p1,p2,t} + \gamma_{p1,t} + \gamma_{p2,t}) \quad (9)$$

Finally, in the robustness check section, we also run OLS regressions with the log of pass rate as the dependent variable:

$$E(\ln pass\_rate_{p1,p2,t}|...) = \beta SameCult_{p1,p2} + PassFric_{p1,p2,t}+ \\ + 1\ln T_{p1,p2,t} + \gamma_{p1,t} + \gamma_{p2,t} \quad (10)$$

In all estimated models, standard errors are clustered at the passer level[16].

[15] We cannot have passer - receiver fixed effect as in some gravity models because our variable of interest is time-invariant.

[16] Another standard in the gravity literature is clustering at passer-receiver level: these will be

6. Results: homophily in collaboration

In this section, we present our model estimation with some extensions to better understand the underlying mechanisms. Empirical work is carried out on our estimation dataset, described in Section 3. It is at the level of passer–receiver pairs and half-seasons (N=669,025).

6.1. Core results

Let us start with the estimation results of the same cultural background (or simply, same culture) indicator coefficient in our Poisson model. When the estimate is greater than zero, we will refer to it as the 'homophily premium'.

Column (1) in Table 3 looks at the unconditional difference: when we look at teams (team*half-season) in our data, we find evidence of a homophily premium: players of the same culture tend to pass 6.55% more to each other than to a player of a different cultural background. With an average of 15.98 passes between players in a half-seasons, this unconditional mean corresponds to an average difference of 1.05 passes.

The estimated unconditional difference of 6.55% corresponds to the average overall homophily premium in teams, and includes both induced and choice homophily[17]. Our model allows us to separate these two, and calculate choice homophily only.

In Column (2), the model is estimated as described in equation 7 with several player controls. In Column (3), equation 9 is estimated double with player fixed effects. In both Column 2 and 3, the count of passes is estimated relative to the player's total passes when both players are fielded together (by forcing the coefficient of the log of total passes to be 1.). Note that team * half-season fixed effects are absorbing the number of total passes in games by a team from the theoretical model. They also absorb team and team * half-seasons characteristics such as history or current management.

Conditioning on observable player characteristics (such as team, position, valuation, citizenship), pass features (such as average distance) in Column (2), player-pairs of the same culture tend to have a pass rate of 2.04% higher than player-pairs of different culture. Note that all player characteristic variables vary over time, valuations change every half-season, and team, position, citizenship fixed effects are interacted with time period fixed effects. We find that receiver valuation and age are positively related to the pass rate. Pass distance is negatively related to the pass rate, while forwardness has a small positive correlation.

Note that Table 3 also allows for benchmarking the estimated coefficient by comparing it to player valuations. When we estimate our baseline model (Column (2)) with player features instead of fixed effects, we find that the coefficient of the same culture

somewhat smaller. With hundreds of thousand observations, standard errors have limited meaning anyway.

[17]The measure is conditional on being in the team. Homophily may have played a role in assembling teams, too.

Table 3: Baseline results

	pass_count		
	(1)	(2)	(3)
Same culture (any) (0/1)	0.0655***	0.0204***	0.0250***
	(0.0091)	(0.0038)	(0.0042)
Average length of passes (ln)		-0.6759***	-0.7944***
		(0.0077)	(0.0094)
Average forwardness Ind (0-1)		0.0066	0.0143
		(0.0077)	(0.0099)
Passer valuation (ln)		-0.0088***	
		(0.0008)	
Receiver valuation (ln)		0.0103***	
		(0.0015)	
Observations	669,025	668,985	668,108
Pseudo R^2	0.07818	0.74163	0.75930
team-half_season fixed effects	✓	✓	
passer_position-league_half_season fixed effects		✓	
receiver_position-league_half_season fixed effects		✓	
passer_nationality-league_half_season fixed effects		✓	
receiver_nationality-league_half_season fixed effects		✓	
passer_position-receiver_position D		✓	✓
passer-half_season fixed effects			✓
receiver-half_season fixed effects			✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. 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 In column 1, additional 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 *half-season fixed effects.

indicator is 2.04% while that of the log average value of the receiver value is 1.03%. This suggests similar magnitudes: sharing the same cultural background is about as likely to lead to more passes as doubling the player pair's valuation (conditioning on other player characteristics). This result suggests that homophily has a very large effect on collaboration.

Column (3) replaces observable player characteristics with time varying passer*half-season and receiver*half-season fixed effects. This allows player attributes to change over time. It also implies that the estimated same culture coefficient is close to what would be the average of coefficients, as if estimated one by one for teams and time periods.

In the fixed effects model, the estimated homophily premium is 2.50%, this is our core result. To interpret the coefficient of interest, we may think of a player who compares two potential receivers who are identical (in terms of all features that are fixed in the given period) except for the fact that only one of them has his same culture. The player will pass 2.50% more per minute to his same culture teammate. As this is a within estimator model, this estimated coefficient is the average of players' homophily preferences.

To summarize, we estimated an overall homophily coefficient of 6.55% and choice homophily of 2.50%. This implies a positive induced homophily: players of the same culture tend to cluster in a way that is conducive to more collaboration, too. One aspect is playing time, as we will see in Section 6.3, players of the same culture tend to spend more time on the pitch together. But they may also play in positions and styles that are more conducive to passes regardless of partners.

Next, we repeat the same exercise measuring same nationality, colonial legacy and same language separately in Table 4. The core fixed effect model is repeated in Column (3). Against a baseline of player-pairs with no shared language or common colonial past, we find that language alone hardly matters (0.66%), while the coefficient of the same colonial legacy (2.34%) is close to the coefficient of the same nationality (3.01%). Thus sharing a language only seems less important for homophilous behavior, while same nationality and common colonial legacy are close to each other.

6.2. Extensions and robustness

In this section, we look at some extensions and robustness checks of the core model (equation 9) with results presented in Table 5.

In Column (1), we consider if passing cost measures such as the average distance between passer and receiver may be parts of the mechanism and exclude them. As results shows, these have only a marginal effect on the coefficient estimate (2.45% vs baseline of 2.50%).

In Column (2), we consider an extended model with some additional potential confounders. We modeled passing cost as the sum of two (log) additive parts culture and passing frictions. There could be additional differences confounding our results. We

Table 4: Baseline results: culture detailed

	pass_count		
	(1)	(2)	(3)
Same nationality (0/1)	0.0799***	0.0238***	0.0301***
	(0.0102)	(0.0042)	(0.0048)
Same colonial legacy (0/1)	0.0140	0.0227***	0.0234***
	(0.0149)	(0.0058)	(0.0066)
Same language only (0/1)	0.0501**	0.0008	0.0066
	(0.0213)	(0.0089)	(0.0095)
Average length of passes (ln)		-0.6759***	-0.7944***
		(0.0077)	(0.0094)
Average forwardness Ind (0-1)		0.0066	0.0142
		(0.0077)	(0.0099)
Passer valuation (ln)		-0.0088***	
		(0.0008)	
Receiver valuation (ln)		0.0103***	
		(0.0015)	
Observations	669,025	668,985	668,108
Pseudo R^2	0.07835	0.74164	0.75930
team-half_season fixed effects	✓	✓	
passer_position-league_half_season fixed effects		✓	
receiver_position-league_half_season fixed effects		✓	
passer_nationality-league_half_season fixed effects		✓	
receiver_nationality-league_half_season fixed effects		✓	
passer_position-receiver_position D		✓	✓
passer-half_season fixed effects			✓
receiver-half_season fixed effects			✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. 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 In column 1, additional 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 *half-season fixed effects.

Table 5: Results on Robustness

	pass count			pass count (ln)	pass count
	(1)	(2)	(3)	(4)	(5)
	Poisson	Poisson	Poisson	OLS	Poisson
Same culture (any) (0/1)	0.0245***	0.0212***	0.0232***	0.0208***	0.0249***
	(0.0048)	(0.0042)	(0.0041)	(0.0034)	(0.0044)
Average length of passes (ln)		-0.7824***	-0.7861***	-0.2898***	-0.8143***
		(0.0094)	(0.0091)	(0.0059)	(0.0114)
Average forwardness Ind (0-1)		0.0113	-0.0027	0.2446***	0.2868***
		(0.0098)	(0.0099)	(0.0043)	(0.0115)
Shared experience (0/1)		0.0105*			
		(0.0056)			
Height difference (cm)		-0.0126***			
		(0.0004)			
Similar valuation (1/0)		-0.0008***			
		(0.0002)			
Both players EU+ (1/0)		0.0100			
		(0.0116)			
Passer total passes when together			1.140***	-0.1309***	
			(0.0048)	(0.0026)	
Observations	668,108	668,108	668,108	669,025	432,125
Pseudo R^2	0.74289	0.76073	0.76038	0.27589	0.71358
passer-half_season fixed effects	✓	✓	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓	✓	✓
passer_position-receiver_position D	✓	✓	✓	✓	✓

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 *half-season fixed effects. Both players EU+ reflect national regulations to play, see Appendix A.2. Similar valuation and height: both below/above median.

included two such variables as shown in Column (2). First, a binary variable to capture when players' valuations are close to each other, signaling similar player quality. This is motivated by the possibility of positive assortative matching with higher quality players interacting more with each other. If quality is correlated with nationality, this could confound our estimate.

Second, there may be some physical attribute that is somehow typical of players of a certain country but not of others. We have data on the height of all players and thus created a passer-receiver -height-difference measure (absolute difference in cm).

Third, we condition on a regulatory aspect that restricts the fielding of players from some non-EU countries. Each league has a different set of regulations adding extra EU country nationals exceptions. Combing through regulations, we added a variable that is one if both players are from the EU or non-restricted countries. For details, see Appendix A. 20% of player pairs have at least one restricted player.

Finally, we add a binary variable that captures shared experience at the club: it is one if the passer and the receiver spent at least a year at a club together. Results are robust to log number of days together.

Most of these new variables are actually correlated with passes, for example, shared experience – being in the team for at least a year together– is associated with 1% more passes. At the same time, they all alter the results just marginally: our estimate drops to 2.12% vs the baseline of 2.50%.

In Column (3), we add the total number of passes when both players are on the pitch - ie not restricting the exposure coefficient to unity. It is actually higher than 1 suggesting a more-than-proportional effect of total passes on pass count. Once again, our coefficient estimate hardly changes (2.32% vs the baseline of 2.50%).

In Column (4), we repeat this model, but use OLS with $\ln passcount$ instead of the Poisson model, only to find a very similar coefficient estimate (2.08% vs the baseline of 2.50%).

Finally, in Column (5), we consider a smaller subset of our data, filtering out rows when there was no pass from the passer to the receiver in that half-season (there was in the opposite direction only), when the two players spent less than 45 minutes together in the half-seasons, and when either of them was the goalkeeper (the model may describe the behavior of the goalkeeper less). As a result, the number of observations is cut by 36%, but our point estimate is unchanged. (2.49% vs the baseline of 2.50%).

The same robustness checks with respect to same nationality, colonial legacy and language may be found in Table .10 in Appendix C.

6.3. Endogeneity of time spent together

A special feature of our setup is that teams change over time, players are replaced within games, and they are selected to play for some but not all games. In half-seasons, player-pairs spend on average 337 minutes or 20% of the maximum feasible amount of time together on the pitch.

Table 6: Selection into play

	pass count (1)	Total passes in shared mins (2)	pass count (3)
Same culture (any) (0/1)	0.0250***	0.0142***	0.0399***
	(0.0042)	(0.0023)	(0.0053)
Average length of passes (ln)	-0.7944***		-0.8390***
	(0.0094)		(0.0108)
Average forwardness Ind (0-1)	0.0143		0.1095***
	(0.0099)		(0.0104)
Observations	668,108	668,117	668,108
Pseudo R^2	0.75930	0.86281	0.67154
passer-half_season fixed effects	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓
passer_position-receiver_position D	✓	✓	✓

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 *half-season fixed effects.

We exploit this feature to estimate a complementary model, keeping all independent variables, but replacing the pass rate with $T_{o,d,t}$, the number of total passes by the passer in which both players are on the pitch (Column (2) in Table 6). It is very closely correlated with the number of minutes they spend together. While the pass rate is eventually determined by players' decisions, minutes played together on the pitch are not: they are decided by the manager (coach).

Of course the manager's decisions are not random, and strongly correlate with the expected joint performance of the fielded players. This, in turn, will be based also on both the coach's evaluation of the same culture premium and his observations of how actual players play together in training. The coach may even teach selected players to play together, thus improving their future passing activity. In this scenario, activity shared on the pitch is also driven by homophily.

If minutes spent together is a mechanism, we shall not partial it out. Thus, we estimate the homophily premium without conditioning on $T_{o,d,t}$, with results Column (3) if Table 6. In essence, in this model, homophily works directly as before but also indirectly: players of same culture work better together in training, and coaches will thus select them to play together in games, too.

We find that the total passes when playing together also shows a homophily pre-

mium, estimated at 1.42%. The homophily premium at the pass count rather than the pass rate level – thus including selection into play – is thus higher: 3.99% vs the baseline 2.50%.

The baseline model with the pass rate as the dependent variable is the one that corresponds to the theory and the same culture aspect of passing cost we care about. The modified model without the exposure in Column (3) may offer a better estimate of what a coach can expect when considering player-pairs. Indeed if we consider playing time together to be endogenous, and view it as a homophily mechanism, the true parameter estimate would closer to 3.99% showed in Column (3). Thus, for a causal interpretation of the effect of same culture, this figure could be a better approximation.

In terms of looking at the three types of cultural background, we see similar patterns for all variables (same nationality, colonial legacy and language). For details see Table .11 in Appendix C.

6.4. Complex collaboration

If same culture is helpful for collaboration in general, one would expect it to be even more so when collaboration is more complex. In our setup, this implies that same culture should matter more for more complex passing patterns. To investigate whether this is indeed the case, we re-estimate the baseline model focusing on complex passing patterns.

To compute complex passing sequences, we first change the dependent variable. Instead of adding up passes, we identify pass sequences and look at the number of passes in a sequence. An S-long pass sequence is simply a series of S consecutive passes between two players P1 and P2. The simplest is a single pass: P1-P2 or P2-P1.

We define complex pass sequences as a pass sequence that includes at least two passes (P1-P2-P1, P1-P2-P1-P2, etc)[18]. On average, player pairs carry out 16 passes per half-seasons. A vast majority, 87%, are single passes, but 13% are complex pass sequences. These complex pass sequences have 3.54 passes on average.

50% of player pairs have made at least one complex pass in our sample[19]. Conditional on having at least one complex pass in the half-season, player-pairs have on average 3.85 complex pass sequences in that half-season.

Before we estimate a model with complex pass sequences, we re-estimate our baseline model with using the count of pass sequences rather than passes, which allows for a precise comparison (the count of passes and pass sequences are very strongly correlated – the correlation coefficient is 0.99). Then, we count the complex pass sequences, i.e.

[18] We can identify pass sequences because our raw event data has timestamps that allow us to capture them.

[19] A large share of players who spend significant time on the pitch but do not produce complex passes are goalkeepers. In modern football they do pass quite a lot, but rarely take part in a sequence of passes.

Table 7: Pass sequences and complex pass sequences

	pass_count_sequences (1)	pass_count_complex (2)
Same culture (any) (0/1)	0.0209***	0.0501***
	(0.0039)	(0.0071)
Average length of passes (ln)	-0.7007***	-1.724***
	(0.0091)	(0.0140)
Average forwardness Ind (0-1)	0.0839***	-0.5174***
	(0.0102)	(0.0141)
Observations	668,108	644,533
Pseudo R^2	0.74590	0.55971
passer-half_season fixed effects	✓	✓
receiver-half_season fixed effects	✓	✓
passer_position-receiver_position D	✓	✓

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. Same cultural background is equality of either cultural aspect (language, colonial legacy or nationality). Total team pass count is captured via team *half-season fixed effects. Sequence count is the number of pass sequences, complex seq count is the number of at least 2 pass-long sequences. Total team pass count is captured via team *half-season fixed effects.

those with at least two consecutive passes, and use the count of complex pass sequences as the dependent variable.

Table **??** presents the results. When we only count the number of passes in complex sequences, we find a stronger effect: the homophily premium is over twice the size for complex passes 5.01% vs 2.09%. This suggests that homophily is especially important for more complex collaboration tasks.

As we look at nationality, colonial legacy and language separately in Table 8, we find that language actually does matter a great deal for complex pass sequences.

6.5. Additional results: Moderating effects, heterogeneity

There are several potential individual level moderator variables as suggested by a variety of economics, management and psychology literature summarized in Ertug et al. (2021). To learn more about key mechanisms we look at the heterogeneity of the estimated average effect accordingly. We focus on differences by (1) minority and majority group, (2) low and high status, and (3) life experience.

To understand the moderating effects, we created 2 or 3 groups for potential mod-

Table 8: Pass sequences and complex pass sequences: culture detailed

	pass_count_sequences (1)	pass_count_complex (2)
Same nationality (0/1)	0.0253***	0.0571***
	(0.0045)	(0.0083)
Same colonial legacy (0/1)	0.0202***	0.0413***
	(0.0064)	(0.0109)
Same language only (0/1)	0.0039	0.0339**
	(0.0092)	(0.0141)
Average length of passes (ln)	-0.7007***	-1.724***
	(0.0091)	(0.0140)
Average forwardness Ind (0-1)	0.0839***	-0.5175***
	(0.0102)	(0.0141)
Observations	668,108	644,533
Pseudo R^2	0.74591	0.55971
passer-half_season fixed effects	✓	✓
receiver-half_season fixed effects	✓	✓
passer_position-receiver_position D	✓	✓

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. Same cultural background is equality of either cultural aspect (language, colonial legacy or nationality). Total team pass count is captured via team *half-season fixed effects. Sequence count is the number of pass sequences, complex seq count is the number of at least 2 pass-long sequences. Total team pass count is captured via team *half-season fixed effects.

erators, and interacted each moderator variable with the same culture indicator. For all these comparisons, we consider the passer's characteristics. Table 9 displays the same culture coefficients for each subgroup, along with some basic information about subgroups. In the third column, we show how relative frequency in the dataset. The last column shows if interaction terms are different from zero at the 5% significance level.

Table 9: Heterogeneity

Heterogeneity source		Freq	Coeff (%)	Diff?
Nationality same as league	Home national	57%	1.70	
	Foreign national	43%	3.40	yes
Player values (euros)	Low (below 3.5m, avg=1m)	27.4%	2.17	
	Medium (1-16m, avg=4.3m)	48.4%	2.60	no
	High (16m+, avg=22m)	22.2%	2.62	no
Age category (ys)	Veteran (29.3+, avg=31.9)	25%	1.99	
	Experienced (23-29.3, avg=26.2)	50%	2.44	no
	Young (below 23ys, avg=21)	25%	3.38	yes
Experience club (days)	Low (below 164, mean: 75)	25%	2.29	
	Medium (165-959, mean 484)	50%	2.70	no
	High (960+, mean: 1850)	50%	2.37	no

Baseline Poisson fixed effect regression model, see Table 3. "Diff?" is statistical difference at 5%. Heterogeneity is defined by the passing player characteristic. Base is first line.

First, let us consider the minority versus majority group argument. Here evidence showed that shared cultural background is more important for minority groups[20].

In our setup, players playing at their home league form the "majority group", such as Spanish players in Spain or Germans in Germany, while minority group is composed of players playing abroad, such as Spanish players in Germany. To test this, we interacted the same culture variable with a binary variable of being a home vs foreign national. Our results confirm this theory: the same culture premium is about twice as high for players playing abroad (3.40%) than for those being at their home league (1.70%). Homophily bias seems to be greater when players are a relative minority.

Second, we look at low and high status, where evidence to date suggested that shared cultural background matters more for low status individuals. This observation comes from the idea that low-status individuals might benefit relatively more from the information shared by similar people. For instance, in an investment bank setting, Ertug

[20]For instance, Greenberg and Mollick (2017) argues that belonging to the minority group acts as a mediator variable increasing homophily, with evidence from gender homophily in crowdfunding.

et al. (2018) finds that homophily is linked to performance for high-status individuals only. In our setup, we used player valuations to distinguish among players of different status. Player valuations reflect skills and experience similarly to wages. We created three groups: low (the bottom 25%), medium (50%) and high (top 25%) valued players. In this case, more highly valued players may need to rely less on information or trust coming from the same culture channel. We found no significant difference (the point estimate is actually higher for high than low status players).

Third, the role of shared cultural background may decline over time. Players may become more familiar with working with different coworkers, and gain experience in different cultural settings, thus reducing the effect of homophily. At the same time, reinforced homophilious behavior – continued collaboration in same culture pairs – may exacerbate the effect.... In our setup we devised two metrics to inspect the moderating effect of time.

The most obvious measure of time is the player's age, and we created three groups by age: young (bottom 25%, below 23 years), experienced (middle half) and veterans (top 25%, above 29.3ys). Here we found a clear pattern, the homophily premium is higher for younger people (3.38%), and then is lower on average (1.99% for veterans and 2.44% for experienced players). Age is an important mediator.

An alternative measure of time is experience with the current club. This is the moderator that would best capture the idea that working together in a team would be enough to eliminate homophily bias. It is different to age: relative young players may be with the club for several years, while as veterans may join clubs anew. Once again we created three groups by the number of days spent together (briefest 25%, longest 25% and in in-between (50%) when the half-seasons starts[21]. Here, we saw no meaningful difference at all.

Taken these two results together, we found that homophily bias may not be reduced by simply spending time together in the team, but will decline as players learn more and become more experienced navigating a diverse workplace. Also, the premium is twice as large for players who are a minority in their country of work.

7. Conclusions

We have investigated how homophily based on cultural traits affect collaboration in 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 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', defined as the count of passes from a passer to a receiver relative the passer's total

[21]To be more precise: 1 September for the Fall, and 1 February for the Winter half

passes when both players are fielded together in a half-season.

We have used a discrete choice model of players' passing behavior as a baseline to separately identify collaboration due to biased cultural preferences ('choice homophily') from collaboration due to opportunities ('induced homophily').

We have found strong evidence of choice homophily. Relative to the baseline, player pairs of same culture have a 2.50 percent higher pass rate. Same culture is about as likely to lead to more passes as doubling the player pair's valuation, which is a consensus measure of players' skills. Moreover, passes between players of the same nationality, same colonial legacy and same language are associated respectively with 3.01, 2.34 and 0.66 percent higher pass rates. These findings show that choice homophily based on culture is pervasive even in teams of very high skilled 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.

Appendix

A. Football rules

A.1. 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[22].

For the purpose of this paper, let us review some key aspects of the game what matters.

In a league, all teams play all other teams twice: in a home and an away game. A team gets 3 points for winning, 1 for drawing and 0 for losing. There is churning season-by season: the worst few (2 or 3) teams every year will be relegated, while a few will be promoted from the lower division to replace them[23].

In a game, there are 2 times eleven players on the pitch. There is freedom in composition, but mostly: 1 goalkeeper, 3-5 defenders (left, center or full- and right-backs), 0-3 forwards, and midfielders. In our data we have very specific positions such as left-back.

The flow of the game is such, that almost two-thirds of the events are passes. In our sample, 62% of events are passes, 77% of which are successful. The rest of the events include shots on the goal, goals, free-kicks, yellow and red cards for disciplinary action, substitutions, tackles and more. There is some variation by teams, some teams pass more than others. Typically better teams pass more.

Each game has a "starting XI" - 11 players who start the game. There are up to 3 substitutions per team/game (this happens typically in the last third of the game). This may happen because of injury or any tactical decision. At all times there will be 11 players on the pitch unless some player gets a red card and is sent out (permanently) - this rarely happens - about once in 5 games. There is freedom in composition, but mostly: 1 goalkeeper, 3-5 defenders (left, center or full- and right-backs), 0-3 forwards, and midfielders. In our data we have very specific positions such as left-back.

Football teams have squads of about 25-30 players. In Spain, squads are typically smaller: 22-28 players, and in England, larger, with 25-33 players. All decisions on who plays is down to the manager (coach).

Churning is large: 20-40% of team changes season to season. Players leave and arrive, this is called a transfer. Transfers happen twice a year in Europe. The summer transfer window is the main opportunity to get new players, or sell existing one. It is between 1 July to 1 September. This is the main window with over 90% of deals in a season. The winter window is shorter, between 1 Jan to 1 Feb, and much smaller. Transfers may include loan deals, when a player spends one or a few half-seasons with

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

[23]For readers unfamiliar with soccer, we kindly recommend watching https://en.wikipedia.org/wiki/Ted_Lasso.

another team.

Note that there are games during the window. This generates a complication with respect to measurement - see in Appendix section B.

A.2. Nationality rules in leagues

In some leagues, there is no limit on use of players of any nationality on the pitch, while in others non-EU, especially South American players face some restrictions. Also, some leagues have rules regarding the squad - must have home grown (academy) players - this has very little effect on starting XI. For our five leagues, we have two types of regulations.

Spain, France, Italy do have restrictions on foreign players. Foreign is defined as non-EU. In Spain it is max 3, in France it is max 4 and in Italy, it is max 2 non-EU. For these countries, the non-EU definition varies marginally but include players from 70 countries "Cotonou" agreement + countries offered the same by home country[24]. In Spain, South Americans get citizenship after 5 years, 2 if they can show Spanish ancestry. In Italy, ancestry also allows a fast track to citizenship, which has helped many people from Argentina and Uruguay.

Non-EU restrictions bite for some African/Asian players but mostly South Americans. The result is that in Spain, France and Italy, this regulation will imply that two Brazilians or a Uruguay and Argentina players are less likely to play together than two Europeans.

England and Germany do not have non-EU player restriction. But both have preference for home grown / academy product players, especially Germany. In England, visa restrictions are managed in a way that gives a preference to players who play or have the potential to play for their national team.

We have coded all these rules. Overall, in our estimation dataset, 89% of observations have a passing player who is considered to be unrestricted in the European Union.

Finally note that all personal information for players are dated as of data collection: summer of 2021. This gives rise a tiny bias: as a few players may get a new citizenship overtime, we may only see it for older players who have already got it. This may downward bias our same nationality estimates.

B. Additional information on data and cleaning

In this subsection, we describe important decisions in the process of data wrangling with a focus on how we coded our key variables.

[24] https://www.footballmanagerblog.org/2018/04/football-manager-squad-registration-rules.html and on the list of countries, see https://en.wikipedia.org/wiki/Cotonou_Agreement

B.1. Player nationalities

If more than one citizenship, typically the first one is the most important one, matching playing for a national team. We kept nationalities as defined by the FIFA, ie a nationality is what has a national team. Hence, we have English and Welsh players not Team GB or Team UK. Our results are robust to having a single UK team. Similarly Feroer Island and some other geographies are also treated as separate nation. For country of birth, we sometimes made edits to match current list of countries. Most importantly, for multi-ethnic countries dissolved since (such as Yugoslavia and Soviet Union) born players were given their current nationality if it was part of federation, if not we imputed the largest country (like Russia).

National team may include U21 and U19 as well. In a very few cases, players switch allegiances, but that means they played no more than 3 games for the team they left. We only included the last one if more than one.

There are several small measurement error issues, all are negligible in impact. First, nationalities may be handed out mid-career. These are typically based on heritage (Italian speaker Argentinians getting Italian citizenship), and it should be a formality rather than a culture change. Second, regarding player values, small measurement errors may come from some low key players having a single year estimate - we replicated it for both half-seasons. Junior team member players promoted to senior team mid-season and would thus have no player value, and we imputed a minimum value here.

B.2. Matching players from two sources and entity resolution

Football players came from two different sources: from the event feed data and from the player information we developed an entity coreference algorithm to match players – find out which records in the two different datasets describe the same player - are coreferent[25].

A baseline solution would be fuzzy matching: use simply the names of the players, with any additional information available, like date of birth, height, nationality and match them based on similarity.

There are several complications for a standard fuzzy matching algorithm for this purpose. First, even for ten thousand players 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, mostly due to the noise in the data: player features are also not precise and unique[26]. Third, data quality problems mean that some players might have two or more different records in one dataset. Fourth,

[25]This section is based on the algorithm developed for this project by Endre Borza, see https://github.com/endremborza/encoref

[26]This problem can be demonstrated with an example of teams: in one dataset the names of two clubs are *Athletic* and *Atletico Madrid*, while in the other *Athletic Bilbao* and simply *Atletico*. So the solution must be open to the possibility, that the two entities, *Athletic* and *Atletico* even though are very similar in name.

algorithmic checkups are really important, as re-examining and correcting the possible matches for over ten thousand players is simply not feasible.

Our improved solution relied on introducing "motifs": a combination of player features. The core concept of the improved solution is not to match simply players, but match motifs in a network of players, matches, seasons and teams. This way, already discovered coreferences can be utilized to narrow the search space, and noise in the data can be mitigated by relying on more than one similarity to establish a coreference[27].

The algorithmic matching is not perfect - as players may use different names, especially South American players, and accents may be incorrectly used as well. When the matching score was low, we checked players by hand - about 1% of total names.

B.3. Detailed cleaning steps and decisions

A pass is recorded when it is "Successful" (this means we know which player received the ball), and when player information is not missing. Data quality is rather high. For example in the EP 2014-15 season we had 351,883 passes, of which were 270,118 successful passes (77%). Only 364 has missing IDs, and 62 cases when the passer and receiver is the same.

The estimation dataset is based on the count of passes, and hence contain non-zero counts only. But, there are 52,093 player-pairs*time (7.8%) where only one direction of 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=4000), when a player had a single partner.
- Player age for every season was defined as number of days to 1 September at the actual year. When a player age was missing, we created sample means by teams and seasons and replace missing with that mean.
- When player position was missing, we replaced it with "Central Midfielder".
- When player valuation for a season were missing, we imputed his average valuation overtime. When player valuations were missing, we imputed 100,000 euros. This happens almost entirely for young and new players.
- There are 6 player pairs that moved together to a different club within a time period. We dropped them.
- As noted earlier, the transfer windows are such that players may move within the window thus playing for more than one team in a half-seasons. In our sample,

[27]See more on the algorithm at https://github.com/sscu-budapest/football-data-project/blob/main/reports/coreference_description.md

we observed 954 occasions when players played for two teams within the same half-seasons (374 players who not only moved teams, but also moved leagues). Very rarely (10 player pair – directions), we observe a player pair passing in two different teams in a half-seasons. This is possible because of the transfer window - players may play in Team A for a few games before moving to another team in the same league. We tagged all these 954 events, and kept them only once (in the team they were more active, almost always the second, longer spell).

C. Additional Tables and Results

D. Team level evidence: passes and winning

This appendix shows the correlation between passing intensity and team performance. The data has passes aggregated to the level of teams and time periods (half-seasons). Thus, one observation is a team*period (i.e. Inter Milan in the first half of the 2015-16 season). We have N=1568 observations (16 time periods, i.e. 8 seasons and 2 half-seasons per season; 4x20 + 1x18 teams). Team performance is measured as the average number of points won in the time period. Teams get 0 for a loss, 1 for a draw, 3 for a win.

We look at how team performance measured by average points is correlated with *ln_pass* defined as log(average pass count per game). We estimate a simple model of correlations:

$$Average\_points = \beta * \ln pass + FEs \tag{.1}$$

First, we only include league dummies, and show a cross-section correlation for a single half-season (2015-16, H1). Then, we estimate a panel fixed effects model adding league-half-seasons and team fixed effects. Column (1) and 2 has points per game, Column (3) has log(points per game) as dependent variable for easier interpretation. Table .12 presents the results.

The first column reports the cross-section OLS results showing a very strong cross-sectional correlation between points per game and pass frequency. In the panel fixed effect models of Column (2) and 3, we see a smaller but economically significant relationship.

We find evidence that, when teams pass more, they also tend to win more. In Column (2), we regressed points per game (in levels) on log total passes, team and league*half-seasons fixed effects. Conditioning on league specific aggregate trends, in half-seasons when a team passes 10% more than its average pass frequency, it tends to win a 0.025 point (or 2.1%) more than average. Over 38 games, this is 1 point (compared to an average of 50 points per team in a season). This difference is equivalent to one position difference in a typical league's standings.

Table .10: Results on Robustness

	pass count			pass count (ln)	pass count
	(1)	(2)	(3)	(4)	(5)
	Poisson	Poisson	Poisson	OLS	Poisson
Same nationality (0/1)	0.0288***	0.0264***	0.0281***	0.0223***	0.0314***
	(0.0055)	(0.0049)	(0.0047)	(0.0040)	(0.0051)
Same colonial legacy (0/1)	0.0252***	0.0192***	0.0214***	0.0217***	0.0202***
	(0.0076)	(0.0065)	(0.0065)	(0.0052)	(0.0071)
Same language only (0/1)	0.0063	0.0039	0.0059	0.0133*	0.0045
	(0.0111)	(0.0093)	(0.0092)	(0.0072)	(0.0097)
Average length of passes (ln)		-0.7824***	-0.7862***	-0.2898***	-0.8143***
		(0.0094)	(0.0091)	(0.0059)	(0.0114)
Average forwardness Ind (0-1)		0.0112	-0.0027	0.2446***	0.2867***
		(0.0098)	(0.0099)	(0.0043)	(0.0115)
Shared experience (0/1)		0.0104*			
		(0.0056)			
Height difference (cm)		-0.0126***			
		(0.0004)			
Similar valuation (1/0)		-0.0008***			
		(0.0002)			
Both players EU+ (1/0)		0.0065			
		(0.0117)			
Passer total passes when together			1.140***	-0.1309***	
			(0.0048)	(0.0026)	
Observations	668,108	668,108	668,108	669,025	432,125
Pseudo R^2	0.74290	0.76074	0.76039	0.27589	0.71359
passer-half_season fixed effects	✓	✓	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓	✓	✓
passer_position2-receiver_position2 D	✓	✓	✓	✓	✓

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 *half-season fixed effects. Both players EU+ reflect national regulations to play, see Appendix A.2. Similar valuation and height: both below/above median.

Table .11: Selection into play: culture detailed

	pass count (1)	Total passes in shared mins (2)	pass count (3)
Same nationality (0/1)	0.0301***	0.0162***	0.0469***
	(0.0048)	(0.0027)	(0.0061)
Same colonial legacy (0/1)	0.0234***	0.0156***	0.0382***
	(0.0066)	(0.0038)	(0.0086)
Same language only (0/1)	0.0066	0.0047	0.0142
	(0.0095)	(0.0050)	(0.0124)
Average length of passes (ln)	-0.7944***		-0.8390***
	(0.0094)		(0.0108)
Average forwardness Ind (0-1)	0.0142		0.1094***
	(0.0099)		(0.0104)
Observations	668,108	668,117	668,108
Pseudo R^2	0.75930	0.86282	0.67156
passer-half_season fixed effects	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓
passer_position-receiver_position D	✓	✓	✓

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 *half-season fixed effects.

Table .12: Team level performance and passes

	points_per_game (1)	points_per_game (2)	ln_points_per_game (3)
Total pass count (ln)	1.142***	0.2471**	0.2095***
	(0.2039)	(0.1168)	(0.0800)
Standard-Errors	HC Robust	cluster: teamid	cluster: teamid
Observations	98	1,568	1,568
Pseudo R^2	0.32060	0.72565	0.95605
leagueseason_half fixed effects	✓	✓	✓
team fixed effects		✓	✓

OLS regression models. Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Team-period level data. German, French, English, Italian, Spanish top soccer leagues. Column 1: First half of 2015/16 season. Columns 2 and 3: 8 seasons 2011-2019. Time period is defined as a half-season, 16-20 games before and after 1 January.

E. Replication options

As data come from private sources and cannot be made publicly available. Thus, we cannot share the raw data publicly. However, we offer a way to reproduce our results by (i) sharing the estimation dataset with individual identifiers purged the dataset and (ii) sharing all codes, including data cleaning.

Analysis are carried out in R, replication codes are available[28].

Furthermore, we will offer access to the data cleaning process with all raw data, with a possibility to run scripts on the full dataset on a secure server. Data wrangling is done in Python, and in R. It will be available on request, and we are currently building a data infrastructure to ease the process.

[28]Codes are stored at the project github page (https://github.com/gbekes/homophily-collaboration), while the data is available from OSF.io (https://osf.io/yc7ux/). These are private repositories at *this* version, access as available on request.

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. 2

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

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. 2

AlShebli, B. K., Rahwan, T. and Woon, W. L. (2018), 'The preeminence of ethnic diversity in scientific collaboration', *Nature communications* **9**, 1–10. 7

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. 2

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. 3

Berge, L. (2018), Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm, Working Paper 13. 17

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. 6

Buchholz, M. (2021), 'Immigrant diversity, integration and worker productivity: uncovering the mechanisms behind 'diversity spillover' effects', *Journal of Economic Geography* **21**(2), 261–285. 5

Calder-Wang, S., Gompers, P. A. and Huang, K. (2021), Diversity and performance in entrepreneurial teams, Working Paper 28684, National Bureau of Economic Research. 6

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

Currarini, S., Jackson, M. O. and Pin, P. (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045. 4, 7

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. 7

Desmet, K. and Ortuño-Ortín, I.and Wacziarg, R. (2017), 'Culture, ethnicity, and diversity', *American Economic Review* **107**(9), 2479–2513. 1

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. 5

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* . 3, 27

Ertug, G., Gargiulo, M., Galunic, C. and Zou, T. (2018), 'Homophily and individual performance', *Organization Science* **29**(5), 912–930. 29

Fally, T. (2015), 'Structural gravity and fixed effects', *Journal of International Economics* **97**(1), 76–85. 17

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. 6

Gauriot, R. and Page, L. (2019), 'Fooled by performance randomness: Overrewarding luck', *The Review of Economics and Statistics* **101**(4), 658–666. 3

Greenberg, J. and Mollick, E. (2017), 'Activist choice homophily and the crowdfunding of female founders', *Administrative Science Quarterly* **62**(2), 341–374. 29

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. 1, 10

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. 17

Hjort, J. (2014), 'Ethnic divisions and production in firms', *The Quarterly Journal of Economics* **129**(4), 1899–1946. 6

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. 3

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. 2, 5

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. 1

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. 3, 6

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. 3

Lang, K. (1986), 'A language theory of discrimination', *Quarterly Journal of Economics* **101**(2), 363–382. 5

Laurentsyeva, N. (2019), From friends to foes: National identity and collaboration in diverse teams, Working Paper 226. 6

Lawrence, B. S. and Shah, N. P. (2020), 'Homophily: Measures and meaning', *Academy of Management Annals* **14**(2), 513–597. 3

Lazear, E. (1999*a*), 'Language and culture', *Journal of Political Economy* **107**(6), S95–S126. 1, 5

Lazear, E. P. (1999*b*), 'Globalisation and the market for team-mates', *The Economic Journal* **109**(454), 15–40. 1, 5

Marsden, P. V. (1987), 'Core discussion networks of americans', *American Sociological Review* **52**(1), 122–131. 4

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. 2

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. 3

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. 3, 4

Melitz, J. and Toubal, F. (2014), 'Native language, spoken language, translation and trade', *Journal of International Economics* **93**(2), 351–363. 9

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. 2, 3

Ottaviano, G. I. and Peri, G. (2005), 'Cities and cultures', *Journal of Urban Economics* **58**(2), 304–337. 5

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. 5

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. 3

Santos-Silva, J. and Tenreyro, S. (2021), The log of gravity at 15, Discussion Paper 1, School of Economics, University of Surrey. 17

Spolaore, E. and Wacziarg, R. (2016), Ancestry, language and culture, *in* 'The Palgrave Handbook of Economics and Language', Springer, pp. 174–211. 1

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. 5

Tovar, J. (2020), 'Performance, Diversity And National Identity Evidence From Association Football', *Economic Inquiry* **58**(2), 897–916. 1, 3

Weidner, M. and Zylkin, T. (2021), 'Bias and consistency in three-way gravity models', *Journal of International Economics* **132**, 103513. 17