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 barriers related to nationality and language affect collaboration in multinational teams? We address this question by creating and exploiting a 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 five men leagues over eight sporting seasons, together with full information on players' and teams' characteristics. We measure collaboration as the average number of passes per minute between a pair of players in a half season. We use a discrete choice model of players' passing behavior as a baseline to separately identify excess collaboration within nationality or language due to preferences ('choice homophily') from collaboration due to opportunities ('induced homophily'). Our dataset allows us to estimate the model using a rich set of player and play characteristics as well as player fixed effects. We find strong evidence of homophily: conditioning on players' and teams' characteristics, player pairs of same nationality exhibit an average number of passes per minute that is 2.5 percent higher than player pairs of different nationality. Same nationality is about as likely to lead to more passes as doubling the player pair's valuation, which is a consensus measure of players' skills. Shared language has about half the impact of same nationality. Pairs of same nationality are also more likely to engage in deeper collaboration, disproportionately participating in more complex pass sequences. These findings show that homophily based on nationality and language is pervasive even in 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.

Keywords: Homophily, organizations, teams, diversity, big data

Preliminary and incomplete: 5 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 Gabor Kezdi, Balazs Murakozy, Miklós Koren, Alice Kugler, Adam Szeidl, Balazs Lengyel and seminar participants at CERS-HAS, University of Reading, IMT

1. Introduction

To compete in the global economy, companies are increasingly relying on a multinational workforce (Neeley, 2015). This allows them 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. Teams like these, however, also face several hurdles. When team members come from different countries and cultural backgrounds, communication can rapidly deteriorate, misunderstanding can ensue, and cooperation can degenerate into distrust.

In this paper we systematically investigate how barriers related to nationality and language diversity 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 dispersion may *per se* inhibit collaboration (Joshi et al., 2002). We show that a 'border effect' between team members of different nationality or language may indeed hamper collaboration, pretty much like it hampers international trade in goods and services between the regions of different countries (Head and Mayer, 2014). Team members of the same nationality or language will collaborate more than team members of different nationality or language.

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 (England, France, Germany, Italy, 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]

We build a unique dataset recording all the 10.7 million passes in 14,608 games by 7 thousand players from 132 countries fielded by 154 teams. To measure same nationality difference, we aggregate our data to time periods with stable squads of players - half of seasons. Estimations will be carried out on a three-dimensional (unbalanced) panel dataset with a total of 669 thousand origin and destination player pairs over the 16 half-seasons.

We measure collaboration as the average number of passes per minute between a pair of players in a given time period (which we call their 'pass rate'), and study how it is affected by the pairs' nationality and language. 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 positively correlated with winning.[2]

Lucca, and CEU for useful comments and suggestions. We are grateful to Endre Borza, Bence Szabo for outstanding and extensive research assistance.

[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) .

[2]See Appendix D

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 workers hardly facing obstacles with integration outside the workplace. Moreover, while language matters for collaboration, 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 players and teams[4]

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 exploiting an extremely rich set of team and worker controls.

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 National Hockey League in North America, (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 European 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

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

[4]This last point is relevant because it is not that some nationality has an advantage in understand and exploiting unwritten/unclear rules, and thus it passes more independently of collaboration propensity.

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.

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 the same nationality or language collaborate more than team members of different nationality or language is a statement about homophily. It highlights common nationality or language 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 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'. Hence, to be of any interest at all, the statement that team members of the same nationality or language 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 a variety of player characteristics as controls. Results are then corroborated by a rich set of robustness checks.

We find that player pairs of same nationality do pass more between them. Conditioning on observable player characteristics (such as team, position, valuation, citizenship), pass features (such as average distance), player pairs of same nationality tend to have a passing rate (count of passes relative the player's total passes when both players are fielded together) of 2.5 percent higher than player pairs of different nationality. Same nationality is about as likely to lead to more passes as doubling the player pair's valuation (a consensus measure of their skills). Same language has about half the impact of same nationality, suggesting that same nationality has multiple components, with language being only one of them. Pairs of same nationality also engage in deeper collaboration, participating in complex pass sequences. These include two-pass sequences (ABA or BAB for players A and B), or longer ones. For such sequences, player pairs of same nationality tend to have a passing rate of 4.8 percent higher than player pairs of different nationality. Shared experience does not affect these results. Once individual experience is controlled for, shared experience has no relevant effect. This suggests that nationality is not a proxy for knowing each other.

These findings show that homophily based on nationality and language is pervasive even in 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 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)).

Three 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. [5]

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

[5] 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 microdata from GitHub, 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.

were allowed to choose their teams from among students in their section. In other cohorts, students were randomly assigned to teams based upon a computer algorithm. In the cohorts that were allowed to choose, Calder-Wang et al. (2021) find strong selection based upon shared attributes. [6] 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 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 using sports data to answer questions in labor economics is not novel. In particular, our data is similar in nature to the basketball data used in Arcidiacono et al. (2017), who use detailed data from the NBA, the major league basketball tournament in the United States during the 2006–9 regular seasons. Similar to ours, they use an event based (or as they call it, play-by-play) data linked to information about players' biographical information. Their approach differs in that their unit of observation is at the player level: an offensive possession, while as we focus on interactions: passes. [7]

[6] Currarini et al. (2009) study friendship formation in US schools when students have types and may see type-dependent benefits from friendships. They show that any matching process such that types are matched in frequencies in proportion to their relative stocks cannot replicate the generalized inbreeding they observe in Add Health 1994 Data, regardless of type sensitivity of preferences. On the contrary, a model with both type-sensitive preferences and a matching bias generates the observed patterns of inbreeding. See also Currarini et al. (2010).

[7] This is also related to very different nature of games: basketball has very frequent scoring (typically around 40-45 successful shots per team vs about 1.5 in football)

3. Data

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

3.1. Overview

To estimate how homophily may affect collaboration we collected, curated and cleaned an immense dataset. It is a combination of two raw datasets, one on events such as passes, and another on player characteristics. The data has been collected by webscraping - collecting information from websites. The original data is stored as a structured text, which we transformed into tabular data.

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.[8] It is a relational dataset linked via player names and additional information.

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 preeminence as the pinnacle of national football competitions. Moreover, for these leagues data availability is the most comprehensive.

The dataset covers all games played in sporting seasons 2011-12 to 2018-19, for which data quality is the highest and were uninterrupted by covid. 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 till 31 of December, the Winter-Spring runs till May. 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. 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.

3.2. Players dataset

Player information and characteristics are compiled by Transfermarkt[9]. They include country of birth, citizenship or citizenships if multiple, date of birth, height, and participation in a national team, which are 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

[8] See replication options in Appendix F.

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

a player may contribute to the team's success, how well he plays, how useful he may be to another team that would purchase the rights, and so on. As such, a player's transfer value is considered a consensus measure of the quality of his footballing skills. Transfer values are estimated twice a year in correspondence of the transfer windows.

Data quality and coverage are both high overall. Several small data cleaning steps were needed, these issues are discussed in Appendix B.

We have 6,998 players in our sample, for whom we can fully map their entire career, with a typical team relying on a squad of about 30 players.

European football is truly globalized and there are players from 132 countries of citizenship in our sample. French, Spanish and Italian players are 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 shares of the first nationality for countries with at least a 1% share.

Table 1: Most frequent player nationalities in passing pairs

Country	share (%)
Spain	14.1
France	12.6
Italy	10.5
Germany	8.6
England	6.8
Brazil	4.4
Argentina	3.9
Portugal	1.8
Senegal	1.6
Netherlands	1.6
Belgium	1.6
Serbia	1.4
Switzerland	1.3
Cote d'Ivoire	1.3
Croatia	1.1
Uruguay	1.1

To determine common nationality between players in the presence of multiple citizenships, we define two players as co-national if they share at least one citizenship, or have the same country of birth.

3.3. Passes dataset

Information on passes comes from OPTA and is available via third party sites. Events are recorded one-by-one and cross-checked to create highly reliable and widely used depiction of games[10]. This is called event feed data.

A pass is an event defined as "any pass attempted from one player to another", including free kicks, corners, throw-ins, goal assists[11]. 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 destination player and his coordinates can be recovered.

Thus, one row in our pass dataset has the following important columns: passing (and receiver) player ID and his team ID, position information on the pass, time into the game, nature of the pass.

There is 17 passes per minute, on average, and there are about 800 successful passes on average per game, so we have 10.73 million passes. In the data, we found 5,400 events when a player passes to himself, and dropped such events.

We have first aggregated the pass 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.

This aggregated pass dataset remains a directional pass level dataset, which means, we have a separate row for player A to player B and another one for player B to player A. This gives us 669,025 observations at origin-destination player and season-half level, which is the level of aggregation we will use for estimation.

With respect to aggregation, half-seasons have several advantages with respect to alternatives such as seasons, games, or half-games. In a game or half-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 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 younger players may evolve. In this respect, a half-season strikes a balance between mitigating the

[10]OPTA's data are generated by machines and humans coding events during games. The data has been scraped from a third party website.

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

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.

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 8% of of the observations have zero passes - this happens when player A passes to player B, but there is no pass from B to A. To balance the dataset, we kept zero passes for these pairs only.

3.4. Combination of datasets

The final task to prepare our work dataset is to combine player information and pass information. To match player and pass data, we had to identify players in both dataset, creating a unique identifier for players we can use to match two datasets. This process, called entity resolution has been an essential albeit a difficult task. First, there are players who are recorded differently across datasets - especially when they have accents, translated from non-latin alphabet, or have many middle names. Second, more than one player 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. Details are in B, these are minor steps, and all results are robust to them.

One observation in the work data is a player pair - an origin and a destination player. Variables may be aggregated at different levels, and table 2 presents examples for all the four different types of variables.

Table 2: Variable types - based on level of aggregation

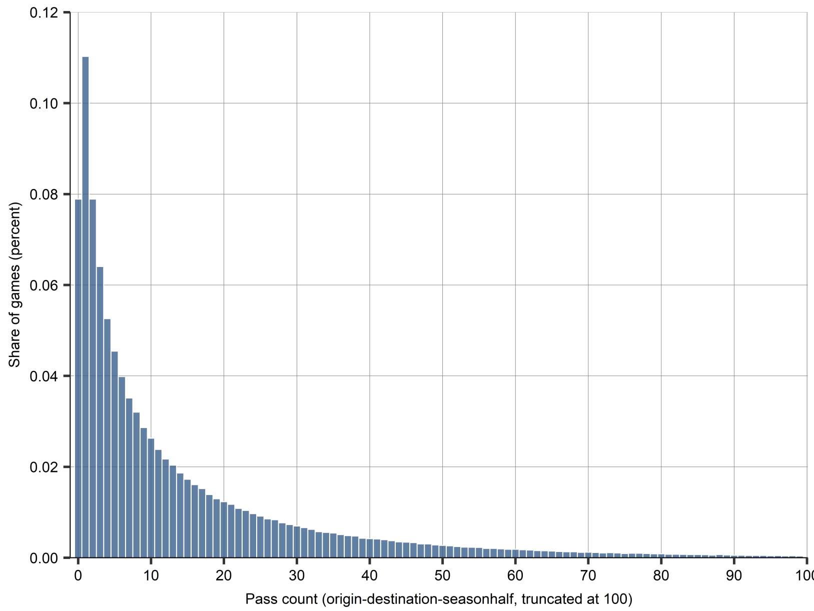
player specific	player-pair specific	season-half specific	Example variables	N
yes	-	-	player height, year of birth, nationality	6,995
yes	-	yes	players age, value, team id, half-season id, experience with the team	37,026
-	yes	-	player-pair's shared nationality indicator	310,493
-	yes	yes	player-pair's number of passes in half-season, shared experience with club	669,025

N refers to the number of different values, ie there are 7 thousand different players and 669 thousand different origib-destination player pair observed in a season-half.

In a half-season, a player will have on average 650 passes (between 2 and 3750, median is 570). He wil have on average 19.31 passing partners (between 2 and 35, median is 20) within a season-half.

Taking the work data (N=669,025), from one (origin) player to another (destination) one, the average pass count is 15.92 (ranging between 0 and 488, median is 8)[12]. The distribution looks highly skewed to the right as shown by 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.[13] During the half-season each player

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

[13]The model can be extended to allow for $N > 11$ and different selections of players being fielded along the half-season.

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 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 its random part ('shock') due to match contingencies. For each receiver this shock is the realization of a random variable Z with continuous differentiable c.d.f. $\Pr[Z \leq z] = G(z)$ over the support $(-\infty, +\infty)$. Any difference in outcomes across the $T^{o,d}$ episodes depends on this shock's realizations only.

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 particular, this cost may be high if o and d find it hard to collaborate due to different nationality (or language) or because the pass is difficult due to distance or the position of players.

Moreover, 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.

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 can be written recursively as

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

According to (1) 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. the player could try to score a goal; or, with the player in control of the ball, the opposing team cannot score) and the option value of player o passing (or keeping control of) the ball at the beginning of the future period. These two components are captured by $\ln u_t^o$ and $\beta \max_{\{d\}_{d=0}^N} \left\{ \varphi^d E\left[U_{j+1}^d\right] - \widetilde{c}^{o,d} + z_t^d \right\}$ respectively, with $\beta \in [0, 1]$ measuring the importance attributed to the latter option and expectation $E\left[U_{t+1}^d\right]$ being taken over future realizations of the shock.

Assuming that Z follows the Gumbel distribution (Type-I Extreme Value distribution)

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

leads to a simple expression for the probability of player o passing to teammate d in period t. Specifically, after taking expectations on both sides of (1), defining $\widetilde{V}_t^o \equiv E\left[U_t^o\right]$, $V_t^o \equiv \exp \widetilde{V}_t^o$ and $c^{o,d} \equiv \exp \widetilde{c}^{o,d}$ allows us 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

$$\pi_t^{o,d} = \frac{\left(V_{t+1}^d\right)^{\varphi^d k}\left(c^{o,d}\right)^{-k}}{\sum_{d=1}^N \left(V_{t+1}^d\right)^{\varphi^d k}\left(c^{o,d}\right)^{-k}} = \left(\Lambda_t^o\right)^{-k}\left(V_{t+1}^d\right)^{\varphi^d k}\left(c^{o,d}\right)^{-k} \tag{2}$$

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.[14] The probability that player o successfully passes the ball to player d in period t is an increasing function of the expected discounted payoff for the team if the ball is passed to player d (V_{t+1}^d), adjusted for the probability this player takes control (φ^d), and a decreasing function of the opportunity cost for the team if player o passes the ball rather than keeping it under control. This opportunity cost itself has two components: the 'passing cost' from o to d ($c^{o,d}$) and the option value of o keeping the ball (Λ_t^o).

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

$$\log p^{o,d} = \log T^{o,d} + \log\left(\Lambda_t^o\right)^{-k} + \log\left(V_{t+1}^d\right)^k + \log\left(c^{o,d}\right)^{-k} - \log P \tag{3}$$

This equation can be estimated using our data 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 independently of the identities of the passer and the receiver. The term $\left(l^{o,d}\right)^{\lambda}$ captures, instead, all non-distance-related frictions that make it difficult for player o to pass to receiver d, independently from the positions they are assigned. These non-distance-related fric-

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

tions may include, for instance, different cultural traits or limited experience in playing together.

In particular, we will measure $g(o, d)$ with $PassDist$, the average distance between passing players. $l^{o,d}$ will allow us measure homophily with $SameNatIndicator$ measuring shared cultural background.

With double player fixed effects, this leads to the following regression

$$\Pi_h^{o,d} = I_h^o + I_h^d + \alpha SameNatIndicator^{o,d} + \gamma PassDist^{o,d} + I_h^{team} + \Upsilon_h^{o,d} + \varepsilon_h^{o,d} \quad (5)$$

where h is the half-season label, $\Pi_h^{o,d}$ is log pass rate, I_h^o and I_h^d are passer and receiver quality measures or fixed effects, I_h^{team} is a team fixed effect (soaking up coach, style, pitch features), $\Upsilon_h^{o,d}$ is $\log T^{o,d}$, and $\varepsilon_h^{o,d}$ is an error term. We have $\gamma PassDist^{o,d}$, a measure of pass difficulty to capture $g(o, d)$.

As for $SameNatIndicator^{o,d}$, this is a time-invariant indicator variable equal to 1 if players o and d have the same nationality, and 0 otherwise. In this case, an estimated $\alpha_l > 0$ would measure a 'homophily premium' in passing between the two players. In this respect, the foregone homophily premium in passing between players of different nationalities would be analogous to the 'border effect' found in gravity regression for trade flows between regions belonging to different countries.

5. Model estimation

We implement our discrete choice model empirically by running a regression of the 'pass rate' for a player-pair on the same nationality indicator variable ($SameNatIndicator$ in the model).

The pass rate is defined as the number of passes ('count') divided by the number of minutes played together for a player-pair ('exposure') and the total number of passes in their teams during a half-season.

Specifically, we use a Poisson model that converts the pass rate into the pass count by multiplying both sides of the estimating equation by exposure, taking logs and thus featuring log(exposure) as a term added to the regression coefficients (offset). We then use $team \times half\_season$ fixed effects to absorb the total number of passes per team in a half-season.

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 log(count) model with a large number of fixed effects.[15] In particular, a drawback of

[15]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).

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.[16]. While the Poisson model seems the best choice for us, robustness checks will be provided with the log(count) as outcome. In detail, let us index player 1, player 2, and time period (half-season) by $p1$, $p2$, and t respectively. The pass count from player 1 to player 2 in period t is then given by $pass\_count_{p1,p2,t}$, while our variable of interest is denoted by $x_{p1,p2,t}$ with potential confounding variables $z_{p1,p2,t}$.

Passes from player A to B are considered relative to the total number of passes when both A and B are on the pitch denoted as *totalpasses_shared*. Accordingly, the pass rate is defined as:

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

In the Poisson model, *pass_count* will be the dependent variable and $\ln totalpasses\_shared$ will be treated as an exposure variable, the corresponding coefficient is set to unity. In most models, the pass count of a player will be soaked up by player fixed effects.

In the model, β denoted the relative value of passing (to anyone) compared to using the ball. Empirically *beta* depends on both team and player characteristics, especially 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.

In all models, the non-player specific passing costs will be captured by a set of dummies for position of player A to position of player B, such as central-defender to striker. These are not play specific, and set to control the general system of football. We also have player-pair specific variables such as the average distance between passing players and the direction of passes. These latter may actually be a mechanism instead of a confounder and we show results with and without.

We estimate two versions of our model. Our variable of inter-personal cost (such as same nationality) is denoted by $g_{p1,p2,t}$, while as passing costs are denoted by $g_{p1,p2,t}$.

First, we consider a version of the Poisson model with a variety of player characteristics as controls:

[16]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. In a robustness check, we used the algorithm developed by (Weidner and Zylkin, 2021) and indeed found only a small bias.

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = exp(\gamma g_{p1,p2,t} + \lambda l_{p1,p2,t} + 1 \ln totalpasses\_shared + \\ + \sum_{j=1}^{2} \left(\eta_j value_{p(j)),t} + \theta_j playerchar_{p(j)),t}\right) +) \tag{6}$$

where, for both players, we include player valuation and total passes in each half-season, as well as $position \times half\_season$, $nationality \times half\_season$ and $team \times half\_season$ dummies. Second, we estimate a version of the Poisson model with $player_1 \times half\_season$ and $player_2 \times half\_season$ fixed effects:

$$\mu_{p1,p2,t} := E(pass\_count_{p1,p2,t}|...) = exp(\gamma g_{p1,p2,t} + \lambda l_{p1,p2,t} + 1 \ln minutes\_shared + \\ + \gamma^1_{p1,t} + \gamma^2_{p1,t}) \tag{7}$$

where $\gamma^1_{p1,t}$ and $\gamma^1_{p1,t}$ are $player_1 \times half\_season$ and $player_2 \times half\_season$ fixed effects [17].

To check robustness, we also run OLS regressions with log(count) as dependent variable. With with $player_1 \times half\_season$ and $player_2 \times half\_season$ fixed effects, we have:

$$E(\ln pass\_rate_{p1,p2,t}|...) = \beta x_{p1,p2,t} + +\delta z_{p1,p2,t} + \gamma^1_{p1,t} + \gamma^2_{p1,t} \tag{8}$$

In all estimated models, standard errors are clustered at the player 1 level[18].

6. Homophily in collaboration

In this section, we present our core results along with extensions that help us better understand those results. We also discuss the robustness test based on OLS regressions with log(count) as dependent variable and player fixed effects. Our dataset described in Section 3 is at the level of directional player-pairs and season-halves (N=668,933), and has 1568 team*season-half units in our dataset.

6.1. Core results

Let us start with the estimation results of the same-nationality indicator coefficient in our Poisson model. When the estimate is greater than zero, we will refer to it as the

[17] We cannot have player1 - player 2 fixed effect as in some gravity models because our variable of interest is time-invariant (at least for the overwhelming majority of observations.)

[18] Another standard in the gravity literature is clustering at player 1 * player 2 level, these are somewhat smaller. With hundreds of thousand observations, SEs have limited meaning anyway.

’homophily premium’.

Table 3: Baseline results

	pass_count		
	(1)	(2)	(3)
Shared nationality (0/1)	0.0745***	0.0191***	0.0256***
	(0.0098)	(0.0040)	(0.0046)
Average length of passes (ln)		-0.6764***	-0.7961***
		(0.0077)	(0.0094)
Average forwardness Ind (0-1)		0.0075	0.0143
		(0.0078)	(0.0099)
P1 valuation (ln)		-0.0083***	
		(0.0008)	
P2 valuation (ln)		0.0112***	
		(0.0015)	
Observations	669,025	659,697	661,484
Pseudo R^2	0.07829	0.74032	0.75795
teamid-time fixed effects	✓	✓	
p1_position-league_time fixed effects		✓	
p2_position-league_time fixed effects		✓	
p1_citizenship-league_time fixed effects		✓	
p2_citizenship-league_time fixed effects		✓	
position_p1-position_p2 fixed effects		✓	✓
player_id1-time fixed effects			✓
player_id2-time fixed effects			✓

Poisson regression model. Standard errors, clustered at player 1 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 nationality is equality of either nationality. Player values are start of the time period. 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.

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 an homophily premium: players of the same nationality tend to pass 7.45% more to each other than to player of different nationality. With an average of 15.98 passes between players in a season-half, this an unconditional mean corresponds to an average difference of 1.07 passes.

The estimated unconditional difference of 7.45% corresponds to overall homophily, and includes both induced and choice homophily. Our model allows us separating these two, and calculating choice homophily only.

Our model is estimated in the rest of the table first with several player controls and then with player fixed effects. In both columns 2 and 3, count of passes are estimated relative the player's total passes when both players are fielded together (by forcing the coefficient of the log of total passes to be 1.). Column 2 partials out several player and pair level characteristics. Column 3 replaces player characteristics with double player-period fixed effects. Note that player-half-season fixed effects are absorbing the number of total passes in games from the theoretical model. As we kept players who moved teams only at the second team, they also absorb team-season-half characteristics such as management, too.

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 nationality tend to have a pass rate of 1.91% higher than player-pairs of different nationality. 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 destination player 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.

Columns 3 replaces observable player characteristics with time varying origin player*half-season and destination player*half-season fixed effects. This allows player attributes to change over time. It also implies that the estimated same nationality coefficient is close to what would be the average of coefficients, estimated one by one for teams and time periods.

The estimated homophily premium is 2.56%, To interpret the coefficient of interest, we may think of a player who compares two potential passing partners who are identical (in terms of all features that are fixed in the given period) except the fact that only one of them has his same nationality. The player will pass 2.56% more per minute to his same nationality teammate.

To summarize, we estimated an overall homophily coefficient of 7.45% and choice homophily of 2.56%. This implies a positive induced homophily: players of the same nationality tend to cluster in a way that is conducive to more collaboration, too. Examples may include that some nationalities provide more players for positions, as well as same nationality players may spend more time playing together.

6.2. Extensions and robustness

In this section, we look at some extensions and robustness checks of the main model result with results are presented in Table 4.

In column (1), we consider if passing cost measures may be part of the mechanism and exclude them. As results shows, this has only a marginal effect on the coefficient estimate.

Table 4: Results on Robustness

	(1) Poisson	pass count (2) Poisson	(3) Poisson	pass count (ln) (4) OLS	pass count (5) Poisson
Shared nationality (0/1)	0.0240***	0.0230***	0.0239***	0.0212***	0.0275***
	(0.0054)	(0.0047)	(0.0045)	(0.0038)	(0.0049)
Average length of passes (ln)		-0.7840***	-0.7870***	-0.4432***	-0.8143***
		(0.0094)	(0.0091)	(0.0057)	(0.0114)
Average forwardness Ind (0-1)		0.0114	-0.0039	-0.0762***	0.2867***
		(0.0099)	(0.0099)	(0.0070)	(0.0115)
Height difference (0/1)		-0.0126***			
		(0.0004)			
Value difference (0/1)		-0.0008***			
		(0.0002)			
Both EU/unrestricted (0/1)		0.0063			
		(0.0117)			
Total passes by p1 when together			1.151***	0.7770***	
			(0.0048)	(0.0030)	
Observations	661,484	661,484	661,484	662,132	432,125
Pseudo R^2	0.74135	0.75940	0.75920	0.51834	0.71358
player_id1-time fixed effects	✓	✓	✓	✓	✓
player_id2-time fixed effects	✓	✓	✓	✓	✓
position_p1-position_p2 fixed effects	✓	✓	✓	✓	✓

Column 1-3 Poisson, column 4 OLS regression model. Standard errors, clustered at player 1 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 nationality is equality of either nationality. Player values are start of the time period. Total team pass count is captured via team *half-season fixed effects.

In column (2), we consider an extended model, adding some additional potential confounders. We modeled passing cost as the sum of two (log) additive parts. Variables that would be in our model (section 4 as part of the friction term $\left(l^{o,d}\right)^{\lambda}$ capture non-distance-related frictions that would make it difficult for player o to pass to receiver d. We 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. We do not see substantive evidence of this. 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 player height and thus created a player-height-difference measure - it is indeed correlated with pass count but changes our coefficient only very marginally. Third, we condition on a regulatory aspect that restricts the play of players from some countries, by adding a variable that is one if both players are from the EU or non-restricted countries. For details, see Appendix A. Neither of these make any meaningful difference to our estimate.

Column (3) we add total number of passes when both players are on the pitch - ie not fix coefficient at one. It is actually higher than 1 suggesting a more-than-proportional effect on shared activity on pass count. Once, again our coefficient estimate hardly changes.

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.

Finally, in column (5), we consider a smaller subset of our data, filtering out rows when there was no pass from the origin to the destination player in that season-half (there was on the opposite direction only), when the two players spent less than 45 minutes together in the season-half, and when either was the goalkeeper (the model may describe goalkeeper behavior less). These cuts the number of observations by 36%. Our point estimate is a bit larger in this case. We believe this a more realistic match to our model, but kept the baseline, being more conservative.

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.

We exploit this feature to estimate a complementary model, keeping all independent variables, but replacing the pass rate with *pass_time_together*, the number of total passes by the origin player in which both players are on the pitch. It is very closely related to the time 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

Table 5: Selection into play

	pass count (1)	Total passes in shared mins (2)	pass count (3)
Shared nationality (0/1)	0.0256***	0.0131***	0.0393***
	(0.0046)	(0.0026)	(0.0060)
Average length of passes (ln)	-0.7961***		-0.8390***
	(0.0094)		(0.0108)
Average forwardness Ind (0-1)	0.0143		0.1094***
	(0.0099)		(0.0104)
Observations	661,484	668,117	668,108
Pseudo R^2	0.75795	0.86281	0.67153
player_id1-time fixed effects	✓	✓	✓
player_id2-time fixed effects	✓	✓	✓
position_p1-position_p2 fixed effects	✓	✓	✓

Poisson regression model. Standard errors, clustered at player 1 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 nationality is equality of either nationality. Total team pass count is captured via team *half-season fixed effects.

expected joint performance of the fielded players. This, in turn, will be based also on both the coach's evaluation of the same nationality 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, minutes shared is caused by same nationality via the same mechanisms we cared about.

In Table 5 we repeat our core results for the pass rate along with estimating a similar model but now also for number of passes shared, and the pass count - without conditioning for passes-shared.

We find that minutes shared on the pitch are also determined by same nationality, and so the coefficient estimate in the model without minutes-shared is higher. While the baseline model with the pass rate as the dependent variable is the one that corresponds to the theory and the object we care about, the modified model with passes in minutes-shared as the dependent variable, in Column (2), may offer a better estimate of what a coach can expect when considering player-pairs.

Indeed if we consider minutes shared together to be endogenous, and view it as a homophily channel, the true parameter estimate would closer to 3.9% showed in Column (3). For a causal interpretation of the effect of same nationality, this figure could be a better approximation.

6.4. Benchmark for the premium

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 nationality indicator is 1.91% while that of the log average value of the destination player value is 1.12%. This suggests similar magnitudes: the difference in pass frequency between same and different nationality players is similar than between players of twice the valuations (conditioning on other player characteristics). This result suggests that homophily has a very large effect on collaboration. The effect may be large partially due to the fact that the estimate is relative to the team average for players from a certain country playing at a certain position, which are features that are correlated with the effect of valuation.

6.5. "Deep" collaboration

If same nationality 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 nationality 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 sequence, we first make a change to the dependent variable. Instead of adding up passes, we identify pass sequences and look at the number of passes in a sequence. A N-long pass sequence is simply a series of N consecutive passes between two players A and B. The simplest is a single pass: AB or BA.

We define complex pass sequences as a pass sequence that includes at least two passes (ABA, BAB, ABAB, BABA, ABABA, etc).[19]. On average, player pairs carry out 16 passes per season-half. 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 directed player pairs have made at least one complex pass in our sample[20]. Player pairs who have made at least one complex pass, had 3.85 complex pass sequences in the 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 neater comparison. The count of passes and pass sequences are 99% correlated. Then, we count the complex sequences, i.e. those with at least two consecutive passes, and use the count of complex passes as dependent variable.

[19] We may identify pass sequences is possible because our raw event data has timestamps that allow us to capture them.

[20] 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 6: Pass sequences and complicated pass sequences

	count sequences (1)	count complex seq. (2)
Shared nationality (0/1)	0.0215***	0.0478***
	(0.0044)	(0.0080)
Average length of passes (ln)	-0.7021***	-1.726***
	(0.0091)	(0.0140)
Average forwardness Ind (0-1)	0.0842***	-0.5173***
	(0.0102)	(0.0142)
Observations	661,484	639,394
Pseudo R^2	0.74454	0.55820
player_id1-time fixed effects	✓	✓
player_id2-time fixed effects	✓	✓
position_p1-position_p2 fixed effects	✓	✓

Poisson regression model. Standard errors, clustered at player 1 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 nationality is equality of either 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.

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 4.78% vs 2.15%. This suggests that homophily is especially important for deeper collaboration tasks.

6.6. Nationality and culture

Shared culture could go beyond shared citizenship, it may be based on shared language or history. Even in football, shared history may implied similar approach to training.

First, our data allowed us to define mother tongues of players, allowing for players to have a background with multiple languages. To deal with it, 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, Morocco and Egypt share Arabic, Morocco and France share French, 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. Hence, a player starting his career in Morocco shares a common language with players starting their careers in Egypt or France, whereas a player starting his career in Egypt (France) shares a common language with players starting their careers in Morocco but not with those starting their careers in France (Egypt). Multiple citizenships created a challenge in determining common language. Similarly to nationality, we defined common language if there is a common language across any of players' nationalities.

Comparing nationality and language for the largest groups of players by nationality, Table 1 reveals an incomplete overlap that arises from former colonial links, or participation in the same legal nationality entity or empire: Spanish is spoken in Argentina and Uruguay, Portuguese is spoken in Brazil, French is spoken in Senegal and Cote d'Ivoire. Another reason for incomplete overlap is that some small countries share a language with neighbors: Croatia with Serbia; Switzerland with France, Germany and Italy, Belgium with France and the Netherlands. This suggests that speaking the same language is not only about verbal interaction, but also about common cultural traits.

To summarize, comparing nationality and language reveals an incomplete overlap between them allowing us to disentangle their effects. Based on the discussion here, we created three groups of player-pairs.

1. No same nationality and no same language (Example: Argentina and Brazil)
2. No same nationality but same language as mother tongue (example: Argentina and Spain)
3. Same nationality (example: two Argentina players)

In the $player - pair \times season - half$ dataset, 38.3% of observations are of the same nationality, 11.4% do not share a same nationality but has the same language as a mother tongue, while 50% do have the same mother tongue.

Table 7: Language, colonial links

	pass_count		
	(1)	(2)	(3)
Shared nationality (0/1)	0.0256***	0.0285***	0.0302***
	(0.0046)	(0.0048)	(0.0048)
Average length of passes (ln)	-0.7961***	-0.7961***	-0.7960***
	(0.0094)	(0.0094)	(0.0094)
Average forwardness Ind (0-1)	0.0143	0.0144	0.0143
	(0.0099)	(0.0099)	(0.0099)
Shared language		0.0128**	
		(0.0064)	
Shared language, no colonial (0/1)			0.0066
			(0.0095)
Has colonial tie (0/1)			0.0234***
			(0.0066)
Observations	661,484	661,484	661,484
Pseudo R^2	0.75795	0.75795	0.75796
player_id1-time fixed effects	✓	✓	✓
player_id2-time fixed effects	✓	✓	✓
position_p1-position_p2 fixed effects	✓	✓	✓

Column 1-3 Poisson, column 4 OLS regression model. Standard errors, clustered at player 1 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 nationality is equality of either nationality. Player values are start of the time period. Total team pass count is captured via team *half-season fixed effects.

We introduce two binary variables, one for the case when two players have the same nationality, and one for the case when they share a language but not nationality. Accordingly, the shared nationality coefficient is not expected to be necessarily the same as in the previous tables as the control group now excludes players of different nationality who speak the same language.

Results reported in 7 show that shared language has about half the impact of shared nationality.

Our next was to consider shared colonial links as a way to ascertain common historical ties. Following Head and Mayer (2014), we used colonial links data from CEPII. We used topcoding, as colonial past almost always implies a shared language, as well. This aspect may be most relevant for us in terms of South American and Spanish, Brazilian and Portuguese and French and North African players. Shared language without colonial ties would be Belgium and France and Switzerland. In our data (at season-half, origin-destination player level), 37.8% has common nationality, 8% has common colonial past, 4.1% has shared language but not colonial past.

In column 3 of 7, we find that it is actually colonial ties rather than language without such ties that may matter. Against a baseline of player pairs with no shared language or common colonial past, we find that language alone hardly matters, while as the coefficient of colonial ties (2.56%) is close to the shared nationality one (3.02%).

6.7. Additional results: Moderating effects, heterogeneity

There are several potential 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 . For all these comparisons, we consider the passing (origin) player's characteristics. Table ?? presents results, along with some basic information about subgroups.

Table 8: Heterogeneity

Heterogeneity source		Freq	Coeff (%)	Diff?
Nationality same as league	Home national	57%	2.09	
	Foreign national	43%	2.81	no
Player values (euros)	Low (below 3.5m, avg=1m)	27.4%	1.99	
	Medium (1-16m, avg=4.3m)	48.4%	2.62	no
	High (16m+, avg=22m)	22.2%	2.77	no
Age category (ys)	Veteran (29.3+, avg=31.9)	25%	1.39	
	Experienced (23-29.3, avg=26.2)	50%	2.77	yes
	Young (below 23ys, avg=21)	25%	3.61	yes
Experience club (days)	Low (below 164, mean: 75)	25%	2.73	
	Medium (165-959, mean 484)	50%	2.73	no
	High (960+, mean: 1850)	50%	2.22	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[21]. In our setup, the majority group is players playing at their home league; Spanish players in Spain or Germans in Germany, while minority group is non-locals, such as Spanish

[21] Lit HERE

players working in Germany. To test this, we interacted same nationality variable with a binary dummy of being a home vs foreign national. Results show a small difference only, that may no be told apart from zero at the 5% significance level.

Second, we look at low and high status, where evidence suggested that shared cultural background matters more for low status individualsLIT HERE. In our setup, we used player valuations to distinguish among three groups: low (bottom 25%), medium and high (top 25%) valued players. In this case, better players may need to rely less on information or trust coming from the same nationality channel. Once again, we found a small, not statistically significant difference (where the point estimate is actually higher for high status players).

Third, experience may be expected to lower the role of shared cultural backgroundLit HERE. In our setup we devised two metrics. The most obvious is age: we created three groups by age, young (bottom 25%, below 23 years), experienced and veterans (top 25%, above 29.3ys). Here we actually found a clear pattern, the coefficient is highest for young people, and lowest for veterans, and differences are actually different from zero. As players become more experienced, the same nationality bias is lower, on average. Age is an important mediator. As an alternative, we also checked experience - days spent with the actual club. There, we saw no meaningful difference.

7. Conclusions

We have investigated how homophily related to nationality and language affect collaboration in multinational teams. In doing so, we have 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.

We have measured collaboration as the average number of passes conditioning on total passes by the passer between a pair of players in a half-season.

To separately identify excess collaboration within nationality or language due to preferences ('choice homophily') from collaboration due to opportunities ('induced homophily'), we have used a discrete choice model of players' passing behavior as a baseline.

First we showed that in teams, same nationality pairs pass 7.5% more frequently. This measure combines both the choice and the induced homophily aspect.

Estimating the model with a rich set of player and play characteristics as well as player fixed effects, we have found strong evidence of homophily. Conditioning on players' and teams' characteristics, player pairs of same nationality exhibit an average number of relative pass frequency that is 2.5 percent higher than player pairs of different nationality. Same nationality is about as likely to lead to more passes as doubling the player pair's valuation, which is a consensus measure of players' skills. Shared language has about half the impact of same nationality, but having a colonial tie is close to the same nationality premium. Pairs of same nationality are also more likely to engage in deeper collaboration, disproportionately participating in more complex pass sequences.

These findings show that homophily based on nationality and language is pervasive even in 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.

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 season-halves 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 dataset, 89% of observations have a passing player who is considered to be unrestricted in the European Union.

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.

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

[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

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, 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

[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.

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 origin and destination player is the same.

The analysis 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 season-half. In our sample, we observed 954 occasions when players played for two teams within the same season-half (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 season-half. This is possible because of the transfer window - players may play in Team A for a few games before moving to another team in

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

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 results*

E. *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 (season-halves). 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 season-halves 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-season-half 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 .9 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*season-half fixed effects. Conditioning on league specific aggregate trends, in season-halves 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.

F. *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

Table .9: 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.

by (i) sharing the aggregated data with individual identifiers purged the dataset and (ii) sharing all codes, including data cleaning.

Analysis are carried out in R, replication codes are available at the project github page (https://github.com/gbekes/football-diversity), while the data is available from OSF.io ().

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 R. It is available on request, and we are currently building a data infrastructure to ease the process.

References

AlShebli, B. K., Rahwan, T. and Woon, W. L. (2018), 'The preeminence of ethnic diversity in scientific collaboration', *Nature communications* **9**, 1–10. 6

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. 6

Berge, L. (2018), Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm, Working Paper 13. 14

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. 5

Buchholz, M. (2021), 'Immigrant diversity, integration and worker productivity: uncovering the mechanisms behind 'diversity spillover' effects', *Journal of Economic Geography* **21**(2), 261–285. 4

Calder-Wang, S., Gompers, P. A. and Huang, K. (2021), Diversity and performance in entrepreneurial teams, Working Paper 28684, National Bureau of Economic Research. 5, 6

Currarini, S., Jackson, M. O. and Pin, P. (2009), 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica* **77**(4), 1003–1045. 6

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. 6

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. 4

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* **0**(ja), null. 3, 26

Fally, T. (2015), 'Structural gravity and fixed effects', *Journal of International Economics* **97**(1), 76–85. 14

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

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, 25

Hinz, J., Stammann, A. and Wanner, J. (2021), State Dependence and Unobserved Heterogeneity in the Extensive Margin of Trade, CEPA Discussion Papers 36, Center for Economic Policy Analysis.
URL: *https://ideas.repec.org/p/pot/cepadp/36.html* 15

Hjort, J. (2014), 'Ethnic divisions and production in firms', *The Quarterly Journal of Economics* **129**(4), 1899–1946. 5

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. 2

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, 4

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. 2, 5

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. 5

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

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

Melitz, J. and Toubal, F. (2014), 'Native language, spoken language, translation and trade', *Journal of International Economics* **93**(2), 351–363. 24

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

Ottaviano, G. I. and Peri, G. (2005), 'Cities and cultures', *Journal of Urban Economics* **58**(2), 304–337. 4

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. 4

Santos-Silva, J. and Tenreyro, S. (2021), The log of gravity at 15, Discussion Paper 1, School of Economics, University of Surrey. 14, 15

Terenzini, P. T., Cabrera, A. F., Colbeck, C. L., Bjorklund, S. A. and Parente, J. M. (2001), 'Racial and ethnic diversity in the classroom', *Journal Higher Education* **72**(5), 509–531. 4

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

Weidner, M. and Zylkin, T. (2021), 'Bias and consistency in three-way gravity models', *Journal of International Economics* **132**, 103513. 15