Centre for Economic Performance

Discussion Paper ISSN 2042-2695

No. 1873
October 2022

Cultural homophily and collaboration in superstar teams

Gábor Békés
Gianmarco I.P. Ottaviano

LSE The London School of Economics and Political Science

UKRI Economic and Social Research Council

Abstract
One may reasonably think that cultural preferences affect collaboration in multinational teams in general, but not in superstar teams of professionals at the top of their industry. We reject this hypothesis by creating and analyzing an exhaustive dataset recording all 10.7 million passes by 7 thousand professional European football players from 138 countries fielded by all 154 teams competing in the top 5 men leagues over 8 sporting seasons, together with full information on players' and teams' characteristics. We use a discrete choice model of players' passing behavior as a baseline to separately identify collaboration due to cultural preferences (`choice homophily') from collaboration due to opportunities (`induced homophily'). The outcome we focus on is the `pass rate', defined as the count of passes from a passer to a receiver relative to the passer's total passes when both players are fielded together in a half-season. We find strong evidence of choice homophily. Relative to the baseline, player pairs of same culture have a 2.42 percent higher pass rate due to choice, compared with a 6.16 percent higher pass rate due to both choice and opportunity. This shows that choice homophily based on culture is pervasive and persistent even in teams of very high skill individuals with clear common objectives and aligned incentives, who are involved in interactive tasks that are well defined, readily monitored and not particularly language intensive.

Key words: organizations, teams, culture, homophily, diversity, language, globalization, big data, panel data, sport

This paper was produced as part of the Centre's Trade Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

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

Gábor Békés, Central European University, KRTK and CEPR. Gianmarco I.P. Ottaviano, Bocconi University, Baffi-CAREFIN, IGIER, CEPR and Centre for Economic Performance, London School of Economics.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1. Introduction

To compete in the global economy, companies are increasingly calling on a multinational workforce. As discussed, for instance, by Neeley (2015), this has pros and cons (Lazear, 1999*b*,*a*). On the one hand, a multinational workforce allows companies to build teams that feature the best talent from around the world, and draw on the benefits of international diversity by bringing together people from many cultures with varied work experiences and perspectives. On the other hand, teams like these also face several hurdles. When team members have different cultural background, communication can rapidly deteriorate, misunderstanding can ensue, cooperation can degenerate into distrust, and collaboration can be reconfigured in ways that favor similar others to the detriment of team performance. Homophily, defined as the tendency to associate with similar others, may limit the gains from diversity.

While the cons of multiculturality can be mitigated and possibly eliminated by careful team selection, training and tasking, we show that homophily is persistent and 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 culture intensive. To tap the best global talents, companies have to accept that some homophily may still distort collaboration.

We define 'collaboration' as the situation of two or more people working together to create or achieve the same thing (dictionary.cambridge.org), and study teams that are not geographically dispersed, as is often the case in multinational companies, since dispersion may *per se* inhibit collaboration (Joshi et al., 2002). We characterize the cultural background (henceforth, simply 'culture') of team members in terms of a set of cultural traits (Spolaore and Wacziarg, 2016; Desmet and Ortuño-Ortín, 2017). These include norms, values and attitudes that are transmitted intergenerationally, which we proxy through nationality, colonial legacy (past membership of a colonial empire), federal legacy (past membership of a political union) and native language. We show that a 'border effect' between team members of different culture may indeed hamper collaboration, pretty much as different language and colonial past hamper international trade in goods and services between the regions of different countries (Head and Mayer, 2014). Partialling out all other player characteristics, team members of same culture collaborate more than team members of different culture. This implies that collaboration may indeed inefficiently favor interactions with similar others beyond their context specific value for the team.

We base our investigation on the unique features of a newly assembled dataset recording all passes made by professional European football players in the top five men leagues (Premier League in England, Ligue 1 in France, Bundesliga in Germany, Serie A in Italy, La Liga in Spain) over eight sporting seasons (2011-12 to 2018-19) together with information on key players' and teams' characteristics.[1] There are 10.7 million

[1]With 'European football', or simply 'football' henceforth, we refer to 'association football', which

passes made in 14,608 games by 6,998 players from 138 countries fielded by 154 teams. Pass data are aggregated to obtain the sum of passes by passer-receiver player pairs in a half-season, a time period composed of 16-20 games (depending on the country) in which football clubs have stable squads. Aggregated pass data are then matched with detailed information on player characteristics. The analysis is carried out on the resulting three-dimensional (unbalanced) panel of 669 thousand passer-receiver pairs over 16 half-seasons.

We measure collaboration as the average number of passes per minute between a pair of players in a half-season (which we call their 'pass rate'), and study how it is affected by the pairs' nationalities, languages, colonial and federal legacies. Passes are the essential building blocks of football. They represent how players work together for the common objective of scoring or preventing the opponent from scoring a goal. Importantly, passes are also positively correlated with winning more league points and achieving higher league standings.[2]

This type of sports data has several advantages. First, the European football industry is very globalized: fans are spread around the world, and multinational teams are the rule in the top five leagues.[3] Second, these leagues represent the pinnacle of the industry with superstar companies that can be expected to act as such with regard to team selection, training and tasking. Third, 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. Fourth, in the top five leagues players are very diverse in terms of origin as they come from over a hundred countries. At the same time, they are all very high skilled (and well-paid) workers hardly facing obstacles with integration outside the workplace. Moreover, while language may matter for collaboration, the role of language as a sheer means of communication rather than a broader cultural trait is unlikely to dominate as football tasks are not particularly language intensive (Nüesch and Haas, 2013). Fifth, 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. Sixth, while team composition is exogenous to players' decisions, collaboration with other team members is mostly up to their individual choices. Seventh, the 'rules of the game' are codified, and crystal clear to all players and teams, ruling out the possibility that players of any specific culture may collaborate more with one another only because

is commonly known as 'football' in Europe and 'soccer' in the United States (Tovar, 2020) .

[2]In our dataset, regressing points per game (in levels) on log total passes, team and league by half-season fixed effects and conditioning on league specific aggregate trends shows that, in a half-season when a team passes 10% more than its average pass frequency across half-seasons, it wins 0.044 point (or 3.5%) more than its average points across half-seasons. Over a typical league's season of 38 games, this sums up to almost 2 points (compared to an average of 50 points per team in a season). This difference is equivalent to one position difference in final standings. See Appendix D for details.

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

they happen to have a better grasp of those rules than other players. Last, unlike most other team ball games, the rules always leave a player with a wide range of options on which teammate to pass to.

All these features allow us to directly investigate collaboration in competitive global teams of high skilled workers with precise common objectives, leveraging a big dataset on interactions in an actual workplace rather than in an artificial experimental lab (Jackson et al., 2003), while also exploiting an extremely rich set of team and worker controls. Moreover, the fact that all players are men allows us to analyze how cultural barriers affect collaboration in multinational teams separately from issues of gender diversity.[4]

We are not the first to exploit team sports data to analyze the potential gains and losses from employing culturally diverse work teams. In the case of the top North American ice hockey league (National Hockey League), (Kahane et al., 2013) find that the presence of European players (with Europe being the typical origin of foreign players) does increase firm-level performance: teams that employ a higher proportion of European players perform better. However, their results also indicate that teams perform better when their European players come from the same country rather than being spread across many countries. When teams have players from a wide array of European countries, integration costs associated with language and cultural differences may start to override any gains from diversity. Parallel evidence based on European football leads to mixed conclusions. In the top German league multinational teams have been found to perform worse than teams with less national diversity (Nüesch and Haas, 2013), whereas the opposite has been found in top continental tournament (Ingersoll et al., 2017). Studying the top leagues of England and Spain, Tovar (2020) suggests that conflicting results may derive from a hump-shaped relation between team performance and predominant nationality. This echoes (Kahane et al., 2013) in that an optimal degree of diversity may exist. What distinguishes our analysis from these and related works is that we zoom in on collaboration and we can measure it accurately through the pass

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

data[5].

The key methodological challenge that our investigation faces has been highlighted in several studies on homophily (Lawrence and Shah, 2020). That team members of same culture collaborate more than team members of different culture is a statement about homophily. It highlights common cultural traits as the antecedents of homophily, that is, the specific attributes that serve as its basis, while singling out collaboration as the targeted consequence of homophily (Ertug et al., 2021). In this respect, in studying homophilic 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'.[6] Hence, to be of any interest at all, the statement that team members of same culture collaborate more has to be based on choice homophily after controlling for induced homophily.

Defining the baseline is quite straightforward in the previous example. It is much less so when individuals may or may not differ along several potential attributes that could confound the roles of the targeted antecedents of homophily, making it harder to ascertain whether individuals are mechanically induced to choose similar others. We address these identification issues by designing the baseline in terms of a discrete choice model of players' passing behavior. The model determines how the pass rate for a pair

[5]Beyond diversity, team sports data are increasingly used to study various issues in labor and public economics. For example, Parsons et al. (2011) exploit data from the top North American baseball league (Major League Baseball) on the way umpires judge throws by pitchers of different race/ethnicity to study discrimination and its impact on discriminated groups' behavior. Kleven et al. (2013) rely on data for professional football players in Europe to shed light on the international mobility responses of workers to tax rates and their impact on local labor markets. Arcidiacono et al. (2017) use data on the top North American basketball league (National Basketball Association) to assess whether worker compensation is influenced by productivity spillovers to coworkers. Gauriot and Page (2019) use European football data for the five top leagues to understand how workers' valuations may be affected by luck.

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

of players is pinned down by their characteristics and opportunities during the matches they play together in a given time period. It is implemented empirically by a Poisson regression with player characteristics as well as player-time fixed effects as controls. Results are then corroborated by a rich set of robustness checks.

We find strong evidence of choice homophily: players have a preference to pass more to players of their same culture than to other players. Specifically, the outcome we focus on is the 'pass rate' defined as the count of passes from a passer to a receiver relative the passer's total passes when both players are fielded together during a half-season. In a regression with passer by half-season fixed effects as well as receiver by half-season fixed effects, conditioning on pass features (such as length) shows that player pairs of same culture have a pass rate 2.42 percent higher than player pairs of different culture. Accordingly, sharing the same cultural background is even more likely to lead to more passes than doubling the player pair's valuation (a consensus measure of their skills).

To put the estimated choice homophily into context, we also estimate how homophily affects passes without distinguishing between its choice and induced aspects. Removing all controls except team by half-season fixed effects, we find that player pairs of same culture have a pass rate 6.16 percent higher than player pairs of different culture. Keeping all controls except players' time spent together on pitch, we find that player pairs of same culture have a pass rate 3.79 percent higher than player pairs of different culture. These findings suggest that 2.36 percent of the overall homophily premium of 6.16 percent vanishes when controlling for player and pass characteristics, and a further 1.38 percent disappears when one considers players' time spent together on pitch. The latter reduction reveals the role of managers' decisions in allowing the team to internalize the effects of homophily.

As for the different cultural traits, passes between players of same nationality and same colonial legacy are associated respectively with a pass rate that is 2.85 and 2.83 higher than the pass rate between players without shared language, shared federal legacy, shared colonial legacy and shared nationality. Same language alone or shared federal legacy are not correlated with pass rate.

Choice homophily is more pronounced for complex pass sequences in which collaboration is more intense as the ball goes back and forth between a given player pair. For these sequences, the pass rate for pairs of same culture is 4.81 percent higher than for pairs of different culture, compared with 2.04 percent for single passes. Shared experience does not affect these results: once individual experience is controlled for, shared experience is irrelevant. Accordingly, choice homophily does not seem to be about prejudice, limited familiarity with diverse environments or lack of professional experience. If this were the case, one would expect choice homophily to eventually evaporate along a player's career.

These findings confirm that choice homophily based on cultural traits is indeed pervasive and persistent even in carefully selected and carefully trained teams of very high skill individuals with clear common objectives and aligned incentives, and involved in interactive tasks that are well-defined and not particularly language intensive.

The rest of the paper is organized as follows. Section 2 offers a selective overview of the related literature beyond works already referenced in this introduction. Section 3 describes data collection and our dataset. Section 4 introduces the discrete choice model of passing behavior. Section 5 presents the model estimation, whose results are then discussed in Section 6. Section 7 concludes.

2. Related literature

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

Four streams are particularly relevant to what we do. The first is concerned with 'diversity spillovers', which improve team performance in a diverse environment, but not necessarily in a team that is itself diverse (Ottaviano and Peri, 2006, 2005). This stream highlights four main mechanisms (Buchholz, 2021). Diversity increases productivity: (i) when people from different countries work on problems together, in turn identifying better solutions by combining their knowledge ('interactive problem-solving'), (ii) through increasing the specialization, variety of skills and approaches to tasks within an occupation, though without necessarily requiring interaction between people from different countries of birth ('complementary task specialization'), (iii) when people from the same country of birth cluster in particular occupations and this clustering facilitates stronger knowledge exchanges ('niching effects'), (iv) when simply through exposure to a diverse range of knowledge and approaches to problems workers learn and become more productive ('exposure effects'). The evidence on US Metropolitan Statistical Area reported in Buchholz (2021) supports exposure effects as the main mechanism, but also interactive problem-solving and complementary task specialization seem to play an important role.

This first stream does not leverage information on diversity and collaboration within teams, which is what we do. In this respect, our investigation is more closely related to a second research stream that studies how individuals of different ethnicities may complement each other in production, but workers of the same ethnic background may collaborate more effectively (Lazear, 1999*b*,*a*; Lang, 1986). Specifically related to our investigation are works highlighting how distortions due to ethnic diversity and discriminatory worker attitudes affect firms and their organization of production. These studies face stiff data challenges. To systematically examine the effects of culture and language within a firm, one needs a host of detailed data: the nationalities of all workers must be identifiable, each worker's skills and output, as well as the collective output of the firm, must be measurable, and all other factors of production should be held constant (Kahane et al., 2013). That is why works on firms are typically based on field experiments (Bertrand and Duflo, 2017). For instance, Hjort (2014) studies team production at a plant in Kenya, where an upstream worker supplies and

distributes flowers to two downstream workers, who assemble them into bunches. He finds that upstream workers undersupply non-coethnic downstream workers (vertical discrimination) and shift flowers from non-coethnic to coethnic downstream workers (horizontal discrimination), at the cost of lower own pay and total output. Team pay, whereby the two downstream workers are remunerated for their combined output, is shown to mitigate discrimination and its allocative distortions[7].

In Hjort (2014), the upstream worker's decision on distributing flowers to the downstream workers resembles the choice a football player faces on passing the ball to his teammates. The context is, however, quite different. Whereas a Kenyan plant is a low skilled, highly charged context in a developing country with ethnic conflicts, a European football team is a high skilled, lowly charged context in a developed area with no real conflicts during our period of observation. Moreover, the flower plant and the football team setups have different pros and cons. The former can exploit an essentially random rotation process to assign workers to positions for identification, but its external validity may be limited. In the latter setup rotation is arguably not random as it depends on the manager's choices, but the richness of information from which to obtain all sorts of individual and team controls makes the case for external validity stronger. Be it as it may, non-random rotation due to endogenous team formation leads to known biases. Calder-Wang et al. (2021) exploit a dataset of MBA students who participated in a required course to propose and start a real micro-business that allows them to examine horizontal diversity (i.e., within the team) as well as vertical diversity (i.e., team to faculty advisor) and their effect on performance. The course was run in multiple cohorts in otherwise identical formats except for the team formation mechanism used. In several cohorts, students were allowed to choose their teams among students in their section. In other cohorts, students were randomly assigned to teams based on a computer algorithm. In the cohorts that were allowed to choose, Calder-Wang et al. (2021) find strong selection based on shared attributes. Among the randomly-assigned teams, greater diversity along the intersection of gender and race/ethnicity significantly reduced performance. However, the negative effect of this diversity is alleviated in cohorts in which teams are endogenously formed. In this respect, as long as the manager of a football team acts as mediator allowing the team to internalize the effects of diversity, the negative impact of diversity on collaboration we find can be seen as a lower bound estimate with respect to what would be found in randomly composed teams.

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 coauthor together more frequently than predicted by their proportion among authors;

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

and that greater homophily is associated with publication in lower impact journals and with fewer citations, even holding fixed the authors' previous publishing performance. By contrast, diversity in inputs by author ethnicity, location, and references leads to greater contributions to science as measured by impact factors and citations. In the same vein, AlShebli et al. (2018) study the relationship between research impact and five classes of diversity: ethnicity, discipline, gender, affiliation, and academic age. Using randomized baseline models, they establish the presence of homophily in ethnicity, gender and affiliation. However, ethnic diversity has the strongest correlation with scientific impact. To further isolate the effects of ethnic diversity, they use randomized baseline models and again find a clear link between diversity and impact. Differently from these studies, we use a discrete choice model rather than randomized models to separate choice homophily from induced homophily.

Finally, the fourth research stream is concerned with the formation of social networks (Jackson, 2008). In particular, Currarini et al. (2009) and Currarini et al. (2010) study friendship formation in US schools when students have types (ethnicities) and may see type-dependent benefits from friendships. They show that any baseline matching process such that types are matched in frequencies in proportion to their relative stocks cannot replicate the homophily they observe in their data. On the contrary, a static model with both type-sensitive preferences ('choosing friends') and a matching bias ('meeting friends through friends') generates the observed patterns of homophily. Differently from these studies, our baseline is derived from a discrete choice model in which forward-looking behavior allows us to highlight the role of biased preferences ('passing to teammates') after netting out also the implications of biased meeting rates ('passing to teammates through teammates') through the model's structure and the dataset's richness.

3. Data

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

The raw data used in the research had been collected from different sources. The passing dataset as well as game level information have been collected from whoscored.com, a sports media company's website. The information on players and their characteristics is compiled by drawing from a set of websites including transfermarkt.com, whoscored.com, or Wikipedia (such as the List of Arsenal F.C. players)[8].

[8]For reproducibility of our results, the combined dataset, codes producing it, as well as all analytical codes will be made available.

3.1. Scope

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

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

A season is composed of two halves: the Fall half-season runs from mid-August till the end of December, the Winter-Spring runs till mid-May. The Premier League, La Liga, Serie A and Ligue 1 are all composed of 20 teams (playing 20×19=380 games), while there are 18 teams (18×17=306 games) in the Bundesliga. In any given season, there are 98 teams in our sample, and we have 98x16=1568 team by half-season units in our dataset. Due to relegation and promotion, we have a total of 154 teams in the sample. Overall, our dataset covers a total of 8×(380×4+306)=14,608 games.[9]

3.2. Player dataset

For every player, data include his country of birth, single or multiple citizenship information, country of birth, date of birth, height, and participation in a national team. These are all time-invariant in our dataset.

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

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

Table 1: Most frequent nationalities

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

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

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

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

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

Fourth, for language we rely on CEPII data as in Head and Mayer (2014) to ascertain

whether or not two countries share one or more common languages. We assume that a player speaks (as mother tongues) the official and widely spoken languages of his country of citizenship at the beginning of his career. We consider some languages that are very close, even if not identical as one language (See Appendix B.1 for details). The fact that our language variable refers essentially to mother tongue implies that it should be seen more as a cultural marker than as a means of communication.

For many players (such as Argentinean and Spanish, Brazilian and Portuguese, or French and Senegalese players) same colonial legacy subsumes same language. For other players (such as Croatian and Serbian, Czech Slovakian or Irish and British players) same federal legacy subsumes same language. As a result, there is a small residual group of players that share the same language but neither colonial nor federal legacies.

Based on nationality, language, colonial legacy and federal legacy, we define the following categories:

1. Same nationality (e.g. two Argentinian players)
2. Same colonial legacy: different nationality, but same colonial legacy (e.g. Argentina and Spain, England and Egypt)
3. Same federal legacy: different nationality, but same federal legacy (e.g. Croatia and Serbia)
4. Same language: different nationality, different colonial legacy, different federal legacy, but same language (e.g. Belgium and France)
5. No shared culture: different nationality, different colonial legacy, different federal legacy and different language (e.g. Argentina and France).

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

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

Finally, we have a proxy measure of player quality. Based on data from a popular player value collection website, our data also includes a player's estimated transfer market value, that is, the "expected value of a player in a free market" as determined by a group of experts. This estimate is based on how much a player may contribute to the team's success, how well he plays, how valuable he may be to another team. As such, a player's transfer value is considered a consensus measure of the quality of his football skills. Transfer values are estimated twice a year in correspondence with the transfer windows.

3.3. Pass dataset

The pass dataset contains aggregate information about passes between any two players at the half-season level.

A pass includes any movement of the ball from one player to another. There are about 365 successful passes on average per game, which for two teams implies 730 passes per game, or about 8.1 passes per minute. In a game most players pass to each other, but not all of them do. For instance, a goalkeeper may not make any pass to a striker. We have 10.73 million passes in total.

Passes are aggregated to average out match contingencies as prescribed by the model.[10] Aggregation is at the level of half-seasons. The partition in half-seasons is determined by the timing of the transfer windows, which are located between seasons (summer transfer window) and at the beginning of the calendar year (winter transfer window). It also splits the number of games during a season into two approximately equal parts: the number of games per team in a half-season ranges between 16 and 20 compared with the exact equal split of 17 for the German Bundesliga and 19 for the other top five leagues.

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

An alternative to half-seasons would be considering full seasons. However, half-seasons have advantages compared to seasons. The presence of the winter transfer window implies that during a season a team's squad may change composition. Our assumption of unchanged player quality makes more sense in a half-season than in a season, especially as younger players may evolve. The fact that half-seasons are separated by transfer windows allows us to cleanly map players' careers as they change teams, thereby combining player and pass information in a consistent way. Finally, considering a half-season allows us to investigate the role of common experience as players who spend more time together on the pitch may learn to pass more to each other.

3.4. Combined dataset

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

[10] See Section 4, equation (2)

Table 2: Variable types - based on level of aggregation

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

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

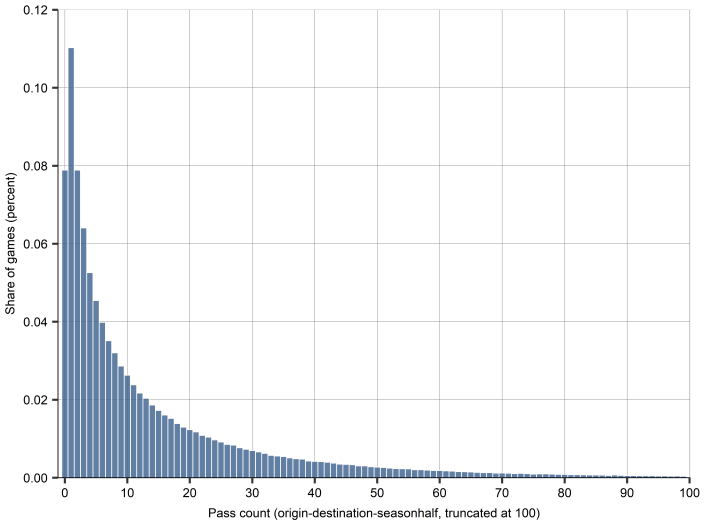
information.

To match player and pass data, we had to identify players in both datasets and create a unique identifier for players. This process has proved to be a difficult task. First, there are players who are recorded differently across datasets - especially when their names have diacritical marks (such as 'é'), are translated from a non-Latin alphabet, or include many middle names. Second, different players may have the same name, especially in the case of frequent family names. To solve this issues, we developed a matching algorithm based on player names and additional information.[11] Variables are aggregated at different levels, as shown in Table 2.

The resulting estimation dataset is a directional pass dataset that, keeping track of who is the passer and who is the receiver, consists of 669,022 observations at the passer, receiver and half-season level. In a half-season, a player makes a total of 294 passes on average (ranging between 2 and 2166, with median equal to 228). On average, he passes to 18.07 receivers (ranging from 2 to 35 with median equal to 19). The average pass count from passer to receiver is 15.92 (ranging from 0 to 488 with median equal to 8). The distribution looks highly skewed to the right as shown in Figure 1, where its support is truncated at 100 (98.63% of observations) for better visibility.

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

Figure 1: Distribution of passes



4. A Discrete Choice Model of Passing Behavior

A crucial challenge in assessing how common culture affects collaboration through passes arises from the conflation of choice and opportunity. As discussed in the introduction, individuals may collaborate more with similar others because they choose to do so ('choice homophily') or because collaboration with similar others is forced on them by circumstances ('induced homophily'). In this section we develop a discrete choice dynamic programming 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).[12]

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 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

[12]See Keane and Wolpin (2009), Todd and Wolpin (2010) and Keane et al. (2011) for surveys of applications of dynamic programming models of discrete choice in labor economics and other applied microeconomic fields

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

to any of the other nine teammates.

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

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

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

where the team's benefit U_t^o of controlling the ball in period t is split into two components: the benefit of player o currently controlling the ball (e.g. in the current period the player could try to score a goal; or he could decide to kick the ball out of play to allow the team to reorganize) and the option value of player o passing (or keeping control of) the ball at the beginning of the future period. These two components correspond to $\ln u_t^o$ and $\max_{\{d\}_{d=0}^N} \left\{ \beta \varphi^d E\left[U_{j+1}^d\right] - \widetilde{c}^{o,d} + z_t^d \right\}$ respectively, with expectation $E\left[U_{t+1}^d\right]$ taken over the realizations z_t^d of the shock. The parameter $\beta \in [0,1]$ measures the relative importance the team attaches to passing in general, independently from the specific passing episode. This is an important characteristic of the team's style of play. For example, low β would be associated with teams that try to score goals by quickly moving the ball into scoring range by long passes, through balls or long air balls, whereas high β would refer to teams that prefer to play less quickly, using many short passes (also sideways or backwards) to find a weakness in the opposing team's tactics.

We assume that the random variable Z follows the Gumbel distribution (Type-I

Extreme Value distribution)

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

with mode 0 and concentration around the mode inversely related to $\kappa > 0$. Zero mode implies that there is no systematic deviation from the deterministic part of the team's benefit across players' assessments of match contingencies. As all players share the same κ, this is a team characteristic: players are trained to assess match contingencies in a common way. Smaller κ then implies more intense training to reduce variation in their individual assessments. The Gumbel assumption leads to a simple expression for the probability of player o passing to teammate d in period t. Specifically, after having taken expectations on both sides of (1), defining $V_t^o \equiv \exp E[U_{t+1}^o]$ and $c^{o,d} = \exp \widetilde{c}^{o,d}$ allows one to express the *ex ante* probability that player o in possession of the ball in period t successfully passes to teammate d at the beginning of period $t+1$ as

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

which *ex post* becomes (approximately) the average share of successful passes that player o makes to player d per episode over a half-season in the subset of passing episodes $T^{o,d}$ when both o and d are fielded and player o has ball possession.[14] The probability that player o successfully passes the ball to player d in period t is thus determined by the team's expected benefit from player d controlling the ball in period $t+1$ (V_{t+1}^d), his probability of taking control of the ball (φ^d) and the difficulty of passing the ball to him ($c^{o,d}$), relative to the team's average benefit from its players $s = 1, ..., N$ controlling the ball in period $t+1$ (V_{t+1}^s), weighted by their probability of taking control of the ball (φ^s) and the difficulty of passing the ball to them ($c^{o,s}$).

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

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

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

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

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

In (4) $g^{o,d}$ is the physical distance between the two players' positions so that $\left(g^{o,d}\right)^{\gamma}$ captures all distance-related frictions that make it difficult to pass the ball from passer o to receiver d independently of their identities. The term $\left(l^{o,d}\right)^{\lambda}$ captures, instead, all non-distance-related frictions that make it difficult to pass the ball from passer o to receiver d independently of their positions. These may include, for instance, limited experience in playing together but, crucially, also different cultural traits.

Substituting (4) into (3) gives

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

which we will use in the next sections as the theoretical basis to empirically investigate the relation between the pass rate $p^{o,d}$ and the cultural dimensions of $l^{o,d}$. Before proceeding, three remarks are in order. First, equation (5) distinguishes the role of biased preferences ('passing to teammates'), which work through $\log l^{o,d}$, from the implications of biased meeting rates ('passing to teammates through teammates'), which work through the forward looking terms $\log \Lambda^{o}_{t+1}$ and $\log V^{d}_{t+1}$. Second, such distinction allows us to argue that in (5) the cultural dimensions of $l^{o,d}$ determine choice homophily (i.e. biased preferences), while induced homophily is determined by all other terms on the right hand side of (5).[15] Third, equation (5) resembles what is called in international economics a 'gravity equation', which explains exports from a country of origin to a country of destination in terms of the countries' characteristics as well as distance- and non-distance-related trade frictions. In this respect, in (5) cultural differences hamper collaboration between teammates pretty much as a 'border effect' hampers international trade in goods and services between the regions of different countries (Head and Mayer, 2014).

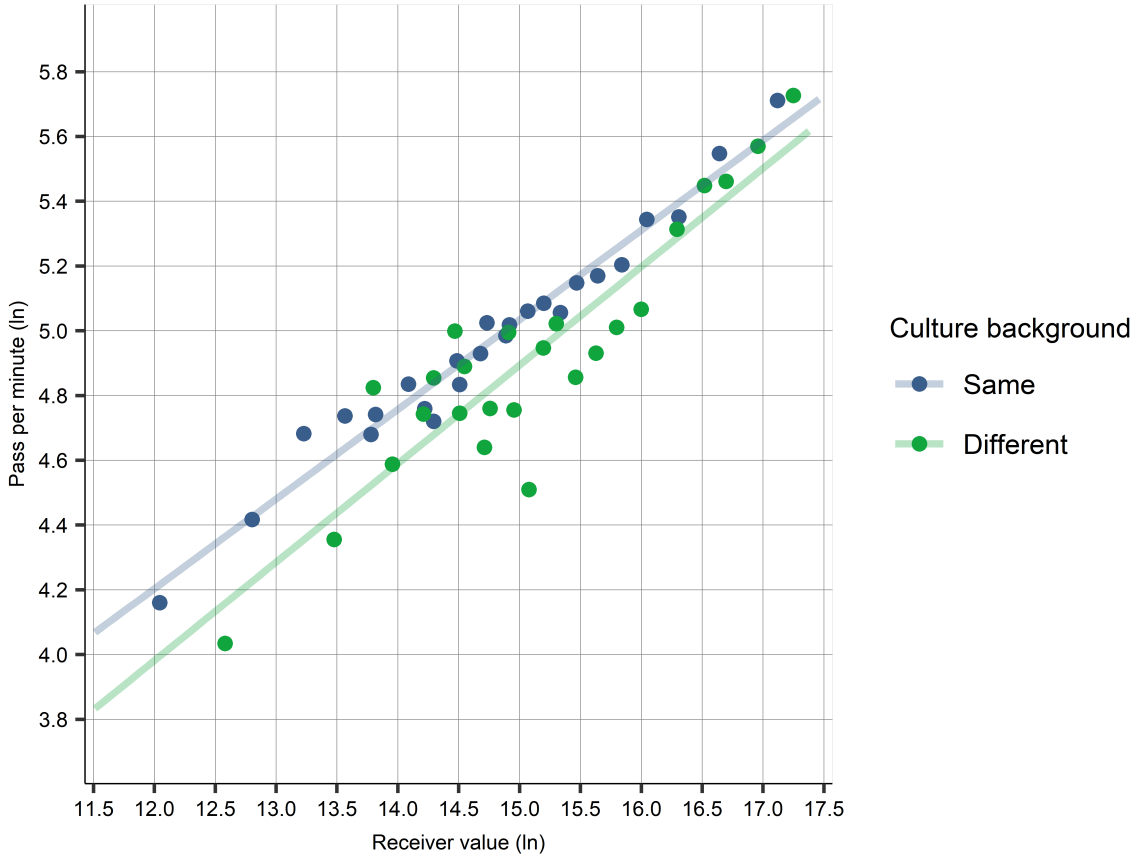
The empirical implications of the theoretical model can be visualized in an intuitive way through a simple concrete example involving 24,299 Spanish midfielders of the Spanish league and pooling their passes over the period of observation. We pick midfielders as they have the greatest freedom of choice in passing and are approximately equally distanced from teammates during a game. We look at the relation between a midfielder's passes per minute and the value of each receiving teammate, controlling for the time they are both on the pitch and the midfielder has the ball. All this allows us to isolate the relation between $\log p^{o,d}$ on the left hand side of (5) and $\log V^{d}_{t+1}$ with $\log l^{o,d}$ on its right hand side, holding all the rest constant. We then proxy V^{d}_{t+1} with receivers' value as a measure of footballing quality and split their sample in two groups depending on whether they have the same culture as the passer or not.

The result is Figure 2, which depicts a binscatter of (log) pass rate against (log) receiver value with fitted linear regression lines for the two groups of receivers. The

[15] The pass rate in equation (5) is conditional on players being in the team, and homophily may play a role in team composition. As long as this affects induced homophily, our model allows us to net it out.

bins are constructed to simplify the scatter plot. There are 25 of them for each receiver group with a bin containing about 500 players. The figure reveals a clear positive relation between pass rate and passer quality for both groups. However, the higher fitted line for same culture receivers shows that, for given receiver quality, the pass rate is higher to same than to different culture receivers. The gap between the two fitted lines visualizes choice homophily.

Figure 2: Pass rate, receiver value and choice homophily



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

5. Empirical Implementation of the Passing Model

The empirical implementation of our theoretical model requires a generalized linear estimator with a log link function. In such setup there is a large literature on the

benefits of using a Poisson model rather than a log(count) model with a large number of fixed effects.[16] This approach is in line with best practices in the estimation of gravity equations through fixed effects Poisson Pseudo Maximum Likelihood estimation (FE-PPML) in international trade (see e.g. Fally (2015) and Santos-Silva and Tenreyro (2021)). Nonetheless, we also provide robustness checks with a log(count) model.

We map our theoretical model into a Poisson model as follows. We use the number of passes from player o to player d im period t as dependent variable. We call it $pass\_count_{o,d,t}$, which corresponds to $P^{o,d}$ in the theoretical model. We then introduce an exposure variable for the total number of passes by o when both o and d are on the pitch, which we call $T_{o,d,t}$ and corresponds to $T^{o,d}$ in the theoretical model. The log of this exposure variable is handled as an offset variable by forcing its coefficient to be equal to 1. The total number of passes per team in a half-season (P in the theoretical model) is absorbed through team by half-season fixed effects. As the parameter β may depend on team and player characteristics that change over time (such as the position and style of players as well as team tactics), we rely on player characteristics and team fixed effects, or alternatively player fixed effects, to absorb its variation.

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

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

where, on average in a half-season, $PassDist_{o,d,t}$ is the distance of passes between the two players, $Forwardness_{o,d,t}$ is the share of passes between the two players with a forward direction, and $Position_o Position_d$ is a dummy variable capturing the two players positions (such as defender, midfielder and forward). As we acknowledge that $PassDist_{o,d,t}$ and $Forwardness_{o,d,t}$ may actually be a mechanism rather than a confounder, we will show results with and without them. In all models we will have a set of dummies for each pair of passer and receiver broad positions (such as forward to defender) ($position_o position_d$).

As for the cultural aspect of non-distance-related frictions that make it difficult to pass the ball from passer o to receiver d independently of their positions, we measure cultural affinity through the time-invariant variable $SameCult_{o,d}$, which combines the categories described in Section 3.2: $SameNat_{o,d} = 1$ if o and d have the same nationality, and $SameNat_{o,d} = 0$ if they have different nationality; $SameCol_{o,d} = 1$ if o and d have the same colonial legacy but different nationality, and $SameCol_{o,d} = 0$ if they

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

have different colonial legacy and different nationality; $SameFed_{o,d} = 1$ if o and d have the same federal legacy but different nationality, and $SameFed_{o,d} = 0$ if they have different federal legacy and different nationality; $SameLan_{o,d} = 1$ if o and d have the same language but different colonial legacy, different federal legacy and different nationality; $SameLan_{o,d} = 0$ if they have different language as well as different colonial, different federal legacies and different nationality. The benchmark therefore consists of pairs with different nationality, different colonial legacy, different federal legacy and different language.

We estimate three versions of the Poisson model. The first version features a variety of player characteristics as controls:

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + PassFric_{o,d,t} + \ln T_{o,d,t} + \sum_{j=o}^{d} (\eta_j value_{j,t} + \theta_j playerchar_{j,t})) \quad (6)$$

where $SameCult_{o,d}$ is the same culture indicator. An estimated δ larger than zero would reveal the presence of choice homophily as it would imply that, all the rest given, players with same culture pass more to each other than to players of different culture. Potential confounding factors associated with the two players $j = \{o, d\}$ are captured by player valuation ($value_{j,t}$) and other player characteristics ($playerchar_{j,t}$). These include age, height, time (in days) elapsed since joining the team, and a binary indicator for being on loan.[17] We also include position by half-season dummies, nationality by half-season dummies, and team by half-season dummies. The second version of the Poisson model differs from the first in that it captures the potential confounding factors associated with the two players through additional fixed effects rather than player valuation and other player characteristics:

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

where $\upsilon_{o,t}$ and $\upsilon_{d,t}$ are player by half-season fixed effects.[18]

The first version will be used to benchmark the same culture coefficient, but the second is generally preferred, not only because it is directly derived from the theoretical model but also because the player fixed effects absorb the additional constraints imposed on the passer by team composition through Λ^o_{t+1} and V^d_{t+1}. Neglecting these constraints may lead to biased estimation.[19] For this reason, the third version of the Poisson model unbundles the four components of the same culture indicator $SameCult_{o,d}$ by extending

[17] For additional details on loans see in Appendix, Section A.2.

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

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

(7):

$$E(pass\_count_{o,d,t}|...) = exp(\delta_1 SameNat_{o,d} + \delta_2 SameCol_{o,d} + \delta_3 SameFed_{o,d} \quad (8)$$
$$+\delta_4 SameLan_{o,d} + PassFric_{o,d,t} + \ln T_{o,o,t} + \upsilon_{o,t} + \upsilon_{d,t})$$

where estimated δ's larger than zero would again reveal the presence of choice homophily based on the corresponding cultural aspects. In all estimations standard errors are clustered at the passer level.[20]

6. Homophily in Collaboration

The presentation of our empirical findings is organized as follows. First, we highlight our main results. Second, we discuss some extensions and robustness checks. Third, we investigate whether homophily is particularly relevant for complex collaboration. Fourth and last, we look at experience as a moderator.

6.1. Main results

Table 3 presents our main results. Columns (1), (2) and (3) report the results for the estimation of equations (6), (7) and (8) respectively. In all three columns, by forcing the coefficient of $lnT_{o,d,t}$ to be equal to one, the effect of culture is estimated for the number of passes from the passer to the receiver relative to the former's total number of passes when both players are fielded together. Team by half-season fixed effects absorb the total number of team passing episodes (P in the theoretical model). They also absorb team and team by half-season characteristics such as history or current management.

Column (1) shows that, after conditioning on player characteristics and pass features, player pairs of the same culture have a pass rate of 1.98% higher than player pairs of different culture. All player characteristics vary over time, valuations change every half-season, and hence team, position, citizenship fixed effects are interacted with time period fixed effects. Receiver valuation and age are positively related to the pass rate, whereas the passer's valuation is negatively related. Also pass distance is negatively related to the pass rate (concurring with the gravity logic of our theoretical model). Comparing the estimated coefficients of same culture and players' valuations reveals that they are of comparable magnitudes: sharing the same cultural background is even more likely to lead to more passes as doubling the players' valuation (conditioning on other player characteristics). This suggests that choice homophily has a substantial effect on collaboration.

Column (2) replaces observable player characteristics with time-varying passer by half-season and receiver by half-season fixed effects. This still allows player character-

[20] A common alternative in light of the gravity literature is clustering at passer-receiver level. That would lead to somewhat smaller standard errors.

istics to change over time.[21] The estimated coefficient of same culture is 2.42%. To interpret this coefficient, consider the passes made by a player in a half-season, conditioning on constant and time-varying receiver characteristics, passer-receiver position pair and other pass features. This player is expected to pass 2.42% more to teammates of same culture than to teammates of different culture. This is our preferred estimate of the 'homophily premium'.

Column (3) shows that different cultural components have different effects on homophily. We find a homophily premium of 2.85% for same nationality and 2.83% for same colonial legacy without same nationality. We find no homophily premium for player pairs from countries with shared federal background or shared language but without same colonial legacy or same nationality. Whereas same colonial legacy is almost as consequential as same nationality, there is no homophily in player pairs from formerly federated countries including Ireland, Northern Ireland and Great Britain.[22].

As discussed in Section 2, endogenous team formation may mitigate the effects of homophily on collaboration. For example, as the manager observes his players in training, he may decide to field same culture players in a game because he sees them collaborating more. In this case, the manager acts as a mediator allowing the team to internalize the effects of homophily. Hence, the effects of homophily on collaboration we estimate can be seen as a lower bound estimate with respect to what would be found in a randomly composed team.

6.2. Total and choice homophily

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

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + \nu_{team,t}) \tag{9}$$

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

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

[21] It also implies that the estimated coefficient of same culture is close to what would be the average of coefficients if estimated one by one for teams and time periods.

[22] See Appendix B.1 for details

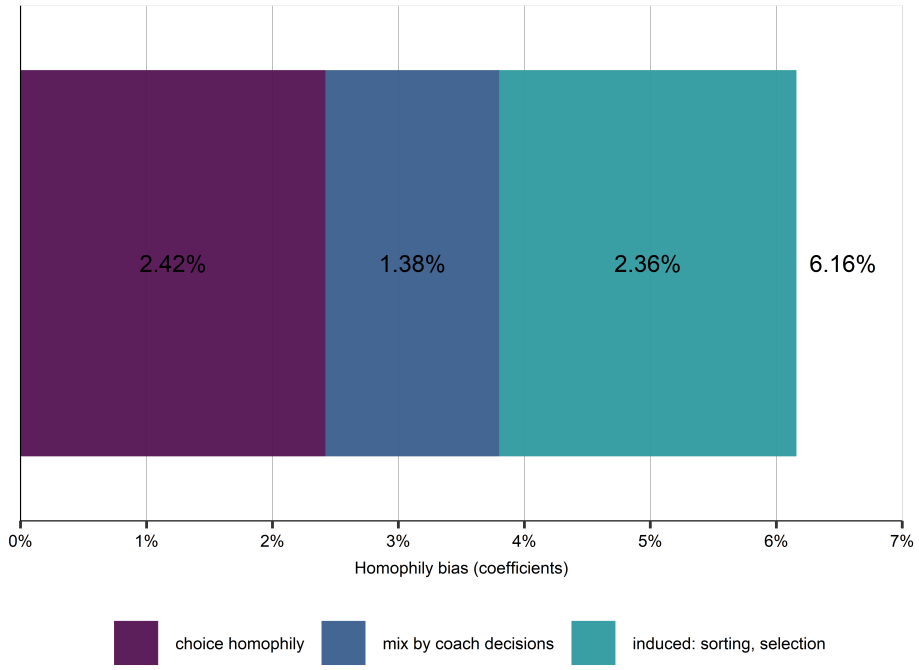
Table 3: Baseline results

	pass_count		
	(1)	(2)	(3)
Same culture (any) (0/1)	0.0198***	0.0242***	
	(0.0037)	(0.0041)	
Same nationality (0/1)			0.0285***
			(0.0048)
Same colonial legacy (0/1)			0.0283***
			(0.0065)
Same federal legacy (0/1)			-0.0227
			(0.0151)
Same language (0/1)			-0.0047
			(0.0123)
Passer valuation (ln)	-0.0087***		
	(0.0008)		
Receiver valuation (ln)	0.0106***		
	(0.0015)		
Average length of passes (ln)	-0.6753***	-0.7944***	-0.7944***
	(0.0077)	(0.0094)	(0.0094)
Average forwardness Ind (0-1)	0.0080	0.0143	0.0142
	(0.0077)	(0.0099)	(0.0099)
Observations	668,982	668,105	668,105
Pseudo R^2	0.74184	0.75929	0.75931
team-half_season fixed effects	✓		
passer_position2-receiver_position2 D	✓	✓	✓
passer-half_season fixed effects		✓	✓
receiver-half_season fixed effects		✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Same cultural background is equality of either cultural aspect (language, colonial, federal legacy or nationality). Player values are start of the half-season In column 1, additional controls are (for both players): height, age, time since with club (in days), binary if on loan. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team *half-season fixed effects.

Comparing the three columns of Table 4 suggests that 2.36% of the overall homophily premium of 6.16% vanishes when controlling for player and pass characteristics, and a further 1.38% disappears when one considers players' time spent together on pitch. This further reduction reveals the role of managers' decisions in allowing the team to internalize the effects of homophily.

Figure 3: Dissecting total homophily bias



Note: Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Coefficient values from fixed effects regression from Table 4.

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

6.3. Extensions and robustness

Table 5 reports the results for some extensions and robustness checks concerning our baseline specification (7) as reported in Column (2) of Table 3.[23]

In Column (1), as the average length and the average forwardness of passes may be a mechanism rather than a confounder, we exclude them from the regression. This

[23]Parallel results for the same set of extensions and robustness checks when keeping nationality, colonial legacy and language distinct can be be found in Table .9 in Appendix C

Table 4: From total homophily to choice homophily

	pass_count		
	(1)	(2)	(3)
Same culture (any) (0/1)	0.0616***	0.0380***	0.0242***
	(0.0090)	(0.0052)	(0.0041)
Observations	669,022	668,105	668,105
Pseudo R^2	0.07813	0.67154	0.75929
minutes shared together			✓
team-half_season fixed effects	✓		
passer-half_season fixed effects		✓	✓
receiver-half_season fixed effects		✓	✓
pass features		✓	✓
passer_position2-receiver_position2 D		✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Same cultural background is equality of either cultural aspect (language, colonial, federal legacy or nationality). Column (2) and (3) includes ln(average pass distance) and forwardness index. Total team pass count is captured via team *half-season fixed effects.

Table 5: Results on Robustness

	(1) Poisson	pass_count (2) Poisson	(3) Poisson	ln_pass_permin (4) OLS	pass_count (5) Poisson
Same culture (any) (0/1)	0.0238***	0.0208***	0.0225***	0.0287***	0.0237***
	(0.0047)	(0.0041)	(0.0040)	(0.0044)	(0.0043)
Shared experience, 1sh+ (0/1)		0.0105*			
		(0.0056)			
Height difference in cm		-0.0126***			
		(0.0004)			
Players value difference, d(ln)		-0.0008***			
		(0.0002)			
Both treated as EU player (0/1)		0.0102			
		(0.0116)			
Passer total passes when together			0.1403***		
			(0.0048)		
Average length of passes (ln)		-0.7824***	-0.7861***	-0.3859***	-0.8143***
		(0.0094)	(0.0091)	(0.0077)	(0.0114)
Average forwardness Ind (0-1)		0.0113	-0.0026	0.3490***	0.2868***
		(0.0098)	(0.0099)	(0.0057)	(0.0115)
Observations	668,105	668,105	668,105	666,230	432,125
Pseudo R^2	0.74289	0.76073	0.76038	0.23492	0.71358
passer-half_season fixed effects	✓	✓	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓	✓	✓
passer_position2-receiver_position2 D	✓	✓	✓	✓	✓

Column 1,2,3,3,5 Poisson, column 4 OLS regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Same cultural background is equality of either cultural aspect (language, colonial, federal legacy or nationality). Player values are start of the half-season. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team *half-season fixed effects. Both players EU+ reflect national regulations to play, see Appendix A.3.

exclusion has, however, only a marginal effect on the homophily premium (2.38% vs 2.42%).

In Column (2), we include additional potential confounding variables. First, we add a variable measuring the difference in the (log) of players' values. As a small number captures similar player quality, positive assortativity would imply a negative coefficient. If quality were correlated with nationality, this could confound our estimated homophily. Second, in the same vein, there may be some physical attribute that is somehow typical of players of a certain country but not of others. As we have data on the height of all players, we control for the absolute height difference (in cm) between passer and receiver. Third, we condition on a regulatory aspect that restricts the fielding of players from non-EU countries with league-specific exemptions. Combing through these, we add a binary variable equal to one if passer and receiver are from the EU or other exempted countries, and zero otherwise.[24] Finally, we add a binary variable that captures shared club experience as shared experience could be hidden behind homophily. This variable equals one if the passer and the receiver spent at least a half-season at a club together, and zero otherwise.[25] While most of these additional variables are actually correlated with passes, they alter the baseline estimates of the homophily premium only marginally (2.08% vs 2.42%).

In Column (3), we add the total number of passes when both players are on the pitch, that is, we do not restrict the exposure coefficient of $lnT_{o,d,t}$ to unity. The fact that the resulting estimate is positive reveals a more-than-proportional effect of total passes on the bilateral pass count. Nonetheless, the coefficient estimate for the homophily premium hardly changes (2.25% vs 2.42%).

In Column (4) we estimate the same core model, but with fixed effect OLS rather than Poisson estimation, and get a very similar coefficient.

Finally, in Column (5), to check whether our results are driven by peculiar cases, we exclude observations when a player passes to a teammate but is never reciprocated in a half-season, when two players spend less than 45 minutes together in a half-season, and when either the passer or the receiver is a goalkeeper.[26] While the number of observations is reduced by 36%, the point estimate for the homophily premium is essentially unchanged (2.37% vs 2.42%).

6.4. Complex collaboration

If same culture is helpful for collaboration in general, one would expect it to be even more so when collaboration is more 'complex'. In our setup, this implies that same culture should matter more for more complex passing patterns. To investigate whether this is indeed the case, we distinguish between single passes (in which player o passes to player d and the ball does not come back) and complex pass sequences

[24] For details, see Appendix A: 20% of player pairs have at least one restricted player.
[25] Results are robust to using the log number of days spent together instead.
[26] Goalkeepers can pass, but often their choice set is more limited.

Table 6: All pass sequences and complex pass sequences

Dep var: pass sequences	all (1)	complex (2)	all (3)	complex (4)
Same culture (any) (0/1)	0.0204***	0.0481***		
	(0.0039)	(0.0069)		
Same nationality (0/1)			0.0239***	0.0546***
			(0.0045)	(0.0083)
Same colonial legacy (0/1)			0.0252***	0.0476***
			(0.0064)	(0.0105)
Same federal legacy (0/1)			-0.0229	-0.0328
			(0.0149)	(0.0239)
Same language (0/1)			-0.0051	0.0130
			(0.0119)	(0.0177)
Observations	668,105	644,539	668,105	644,539
Pseudo R^2	0.74590	0.55970	0.74591	0.55971
passer-half_season fixed effects	✓	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓	✓
passer_position2-receiver_position2 D	✓	✓	✓	✓

Poisson regression model. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Top 5 soccer leagues, 8 seasons: 2011-2019. Half-season: 16-20 games before and after 1 January. Same cultural background is equality of either cultural aspect (language, colonial, federal legacy or nationality). Includes ln(average pass distance) and forwardness index. 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. Includes ln(average pass distance) and forwardness index. Total team pass count is captured via team *half-season fixed effects.

(in which the ball goes back and forth between the two players at least once). On average, as already mentioned, player pairs make 15.98 passes per half-season. A vast majority of them (87%) consists of single passes, but 13% are complex pass sequences. These complex pass sequences feature 3.54 passes on average and 50% of player pairs are involved in at least a complex pass sequence in our sample. Conditional on having joined at least a complex pass sequence in a half-season, on average player pairs are involved in 3.88 complex pass sequences in that half-season.

In Table 6 we present the results for the number of pass sequences instead of the number of passes. For cleaner comparison, in Column (1) we re-estimate our baseline model using as dependent variable the number of pass sequences (simple and complex) between two given players. In contrast, in Column (2) we re-estimate the baseline model using as dependent variable the number of complex pass sequences in which the two

players are involved. Comparing the two columns reveals that the homophily premium is more than twice as large for complex pass sequences: 4.81% in Column (2) compared to 2.04% in Column (1). Repeating the same exercise for the unbundled culture traits in Columns (3) and (4) reveals similar differences for the significant cultural traits.

This confirms that homophily is especially important for more complex collaboration.

6.5. Experience as a moderator

The tendency of a player to collaborate more with similar others may be due to prejudice, limited familiarity with diverse environments or lack of professional experience. If this were the case, one would expect homophily to eventually evaporate along the player's career.

We check whether this is the case by introducing three additional moderator variables in our baseline regression: the passer's age, the number of days he has spent in the current club, and the number of days he has spent there together with the receiver. For each variable, we compare the estimated homophily premia in three split samples corresponding to the first quartile (bottom 25%), the two middle quartiles (middle 50%) and the fourth quartile (top 75%) of the corresponding distribution. Once again, estimation is based on specification (7).

The estimated homophily premia are reported in Table 7, with the moderator variables on the rows and the corresponding quartile bins on the columns. In the case of age, the first row shows that the homophily premia for passers in the bottom quartile (aged 23.2 or less), middle quartiles (aged 23.2-29.3) and top quartile (aged 29.3 or more) are 2.60%, 2.38% and 1.81% respectively. While in this case the point estimate of the homophily premium decreases indeed with age, contrasting patterns emerge for the other variables. Whereas passer time spent in the current club exhibits a hump-shaped trajectory (1.87%, 2.52% and 2.36%), passer time spent there together with the receiver exhibits an increasing trajectory: 2.02%, 2.70% and 2.73%. These patterns, however, neither show a statistically different gap, nor are robust to plausible alternative sample splits.

Overall, the only robust finding is that, in most alternative sample splits, the homophily premium remains positive, and not far from the baseline obtained for the pooled sample, and this holds also for the highest bins. Accordingly, homophily does not seem to be really about prejudice, limited familiarity with diverse environments or lack of professional experience. As such, it remains a persistent feature of football players.

Table 7: Experience as moderator? Split sample regressions

Variable	Description (measurement)	Low Bottom 25%	Low Middle 50%	High Top 25%
Age	Age (ys) at half-season start	0.0260*** (0.0063)	0.0238*** (0.0050)	0.0181* (0.0078)
With club	Number of days since with club	0.0187*** (0.0050)	0.0252*** (0.0044)	0.0236** (0.0087)
Together	Number of days together	0.0202*** (0.0053)	0.0270*** (0.0043)	0.0273** (0.0085)

Each cell reports the same culture coefficient from a split sample regression for three mediator variable by passing player: age, days with team, days playing with receiver. The regressions are baseline Poisson fixed effect regression models, see Table 3. Standard errors, clustered at passer level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The first column is the first quartile by the mediator variable (lest experienced), the second column is the middle second and third quartile, and the last column is for the fourth quartile (most experienced).

7. Conclusions

We have investigated how homophily based on cultural traits affects collaboration in superstar multinational teams. In doing so, we have collected and exploited a newly assembled exhaustive dataset recording all passes by professional European football players in all teams competing in the top five men leagues over eight sporting seasons, together with full information on players' and teams' characteristics.

The outcome we have chosen as our measure of collaboration is the 'pass rate', defined as the count of passes from a passer to a receiver relative the passer's total passes when both players are fielded together in a half-season. The cultural traits we have focused on are nationality, colonial legacy, federal legacy and language, and we have measured 'culture' through their combination.

We have used a dynamic discrete choice model of players' passing behavior as a baseline to separately identify collaboration due to cultural preferences ('choice homophily') from collaboration due to opportunities ('induced homophily'). Induced homophily includes domain-specific aspects of collaboration such as the specialization of players with given cultural traits in certain tasks and positions, the correlation between quality and culture or behavioral aspects like patience. Our methodology has also allowed us to exclude mechanisms such as assortative matching of players of similar quality or shared experience.

We have found strong evidence of choice homophily. Relative to the baseline, player pairs of same culture have a 2.42 percent higher pass rate. Same culture is about as likely to lead to more passes as doubling the player pair's valuation, which is a consensus

measure of players' skills. As for the different traits, passes between players of same nationality and between players with same colonial legacy but without same nationality are associated with 2.85 and 2.83 percent higher pass rates respectively. Same language and same federal legacy with neither same nationality nor same colonial legacy play no role.

We have also found that sharing a common cultural background is more important for more complex sequences (involving the same player pair repeatedly) than for single passes: the former's homophily premium is about twice as large. Players' experience does not change the results. When time spent together on the pitch (a component of induced homophily) is not controlled for, measured homophily increases: same culture players are selected to play together at 1.38 percent higher frequency than players with different cultural backgrounds. This reveals the managers' role in making their teams internalize players' homophilic preferences.

The fact that players' experience does not change the results suggests that choice homophily is not due to prejudice, limited familiarity with diverse environments or lack of professional experience. This shows that choice homophily based on culture is pervasive and persistent even in teams of very high skill individuals with clear common objectives and aligned incentives, who are involved in interactive tasks that are well defined, readily monitored and not particularly language intensive. Cultural preferences affect collaboration also in superstar teams of professionals at the top of their industry.

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.[27] In this section, we review some relevant aspects of the game.

In a league, all teams play all other teams twice: in a home and in an away game. A team gets 3 points for winning, 1 for drawing and 0 for losing. There is churning season by season: every year the worst few (2 or 3) teams are relegated, while a few are promoted from the lower division to replace them.[28]

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

In a game, there are twenty-two (2x11) players on the pitch. All decisions on who plays are down to the head coach, who is sometimes also the manager of the team. For each game, the coach nominates a "starting 11" - 11 players who start the game. In the period of observation there are up to 3 substitutions per team/game (these tend to occur in the last third of the game). Substitutions may happen because of an injury or due to some tactical decision. At any time during a game, there are 11 players on the pitch unless a player gets a red card and is sent out (permanently). However, this rarely happens - about once every five games.

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

A.2. Teams and transfers

Football teams are organizations with 25-30 players (also called the squad). Churning in the squad composition is high from one season to another, typically 20-40% of

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

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

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

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

A.3. Nationality rules in leagues

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

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

France, Italy and Spain have restrictions about the number of 'foreign' players, defined as players coming from non-EU countries. In France their number is capped at 4, in Spain at 3, and in Italy at 2. Among these three countries, the definition of non-EU varies only marginally, but in all cases they include a set of African, Caribbean and Pacific countries (such as Nigeria, Ivory Coast, Guyana) with which the European Union has eased labor laws under the "Cotonou" agreement.[29]

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

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

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

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

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

European Union. In a robustness check, we condition on these regulations and find them having no effect on our results.

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

B. Additional information on data and cleaning

In this subsection, we describe the major decisions.

B.1. Defining player nationality, colonial legacy, federal legacy and language

We kept nationalities as defined by FIFA, the international football's governing body.[30] In practice, a nationality is defined as being a polity with a national football team. In most cases a country would form a nationality. However, there are some exceptions: the United Kingdom has four national teams (Wales, Scotland, Northern Ireland, and England), and small nations like the Faroe Islands (a constituent country of Denmark) or Jersey (a British Crown Dependency) are also treated as separate nations. Apart from the UK, which is treated separately in the paper, all others have only a handful of players.

In the period of observation some players changed nationality as their countries, such as Czechoslovakia, Yugoslavia and the USSR, dissolved and gave birth to new countries. Hence, for country of birth we had to occasionally make edits to match the current list of countries.

Regarding colony and official language definitions, we followed CEPII data.[31] Similar (but not the same) languages like Danish and Norwegian were considered as different. For the definition of same culture, we had to consider pairs with multiple nationalities. For instance, it is possible that two players have both the same nationality and the same colonial legacy (for example, P1 is Moroccan and French and P2 is French). In such cases, we adopted a '*top-coding*' approach, and considered them to be of the same nationality. This is actually a large set of player pairs: 54% of those who share a colonial past also share a citizenship. As for colonial legacy, the majority of the same colonial legacy category comes from a link between a ruler and a former colony. Some links are derived from having the same colonial ruler. It is possible for two players to have a ruler-colony legacy and a colonial sibling legacy (for example, if P1 is a citizen of Ivory Coast and P2 is a citizen of Senegal and France). In 86% of the cases, the same colonial

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

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

history means the same language as well. However, this is not true for all country pairs (e.g. England and Egypt or Russia and Georgia). Some countries had multiple colonial rulers (such as Cameroon with France and England, so linked to both).

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

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

Data on football players come from two different sources: passing data and player information. In each datasets, players are identified via their names. To combine them, we developed an entity coreference algorithm to match players based on variations on their names, and some background information.[32]

Our method improves upon a standard fuzzy matching algorithm. First, even for ten thousand players as in our sample, it takes a lot of computing power to calculate all possible similarities and find the best ones. Second, simply matching the players by themselves is not precise enough, and thus we must use additional information, such as their teams or first nationality. However, even player features (such as team names) are also not precise and unique.[33] Third, data quality problems also mean that in one dataset some players might have two or more different records. Fourth, we added an algorithmic checkup, because re-examining and correcting the possible matches for over ten thousand players by hand is simply not feasible.

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

The algorithmic matching is not perfect as players may use different names, and accents may be incorrectly used as well. When the matching score was low, we checked the match by hand and corrected player names if deemed necessary - reaching about 1% of total player names.

B.3. Detailed cleaning steps and decisions

One important aspect is the possibility of zero passes. Due to aggregation, all passer-receiver pairs in the pass data have non-zero passes. However, there are 52,092

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

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

player-pairs*time (7.8%) where only one direction of the pass is recorded. As clearly a pass was possible, we added zeroes for these pairs for the opposing direction.

There are several additional steps of data wrangling:

- We dropped observations (N=340), when a player had only a single partner in half-season.
- Player age for every season was defined as the number of days to the 1 of September in the current year. When a player age was missing, we created sample means by teams and seasons and replaced the missing with that mean.
- When player position was missing, we replaced it with 'Central Midfielder'.
- Player ID was missing in 0.1% of cases and in 62 cases the passer and receiver were the same. We dropped these observations.
- When a player value for one season was missing, we imputed his average valuation over time. When player valuations were missing, we imputed 100,000 euros. This happened almost entirely for young and new players.
- There were 6 player pairs who moved together to a different club within a time period. We dropped them.
- As noted earlier, games are not suspended during the transfer windows. Hence players moving within a window may end up playing for more than a team in a half-season. In our sample, we observed 954 events when players played for two teams within the same half-season.[34] We kept them only once, in the team where they had the longer spell.[35]

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

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

C. Additional Tables and Results

Table .8: Selection into play: culture detailed

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

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

Table .9: Results on Robustness

	pass count			pass count (ln)	pass count
	(1)	(2)	(3)	(4)	(5)
	Poisson	Poisson	Poisson	OLS	Poisson
Same nationality (0/1)	0.0269***	0.0249***	0.0265***	0.0238***	0.0298***
	(0.0055)	(0.0048)	(0.0047)	(0.0048)	(0.0051)
Same colonial legacy (0/1)	0.0314***	0.0264***	0.0266***	0.0231***	0.0272***
	(0.0075)	(0.0065)	(0.0064)	(0.0063)	(0.0069)
Same federal legacy (0/1)	-0.0304*	-0.0307**	-0.0237	-0.0093	-0.0294*
	(0.0176)	(0.0150)	(0.0147)	(0.0161)	(0.0154)
Same language (0/1)	-0.0038	-0.0087	-0.0052	0.0152	-0.0147
	(0.0145)	(0.0121)	(0.0120)	(0.0111)	(0.0124)
Shared experience, 1sh+ (0/1)		0.0105*			
		(0.0056)			
Height difference in cm		-0.0126***			
		(0.0004)			
Players value difference, d(ln)		-0.0008***			
		(0.0002)			
Both treated as EU player (0/1)		0.0064			
		(0.0117)			
Passer total passes when together			1.140***	0.2838***	
			(0.0048)	(0.0036)	
Average length of passes (ln)		-0.7824***	-0.7861***	-0.3918***	-0.8141***
		(0.0094)	(0.0091)	(0.0071)	(0.0114)
Average forwardness Ind (0-1)		0.0112	-0.0027	0.3089***	0.2866***
		(0.0098)	(0.0099)	(0.0052)	(0.0115)
Observations	668,105	668,105	668,105	666,230	432,125
Pseudo R^2	0.74290	0.76074	0.76039	0.26121	0.71361
passer-half_season fixed effects	✓	✓	✓	✓	✓
receiver-half_season fixed effects	✓	✓	✓	✓	✓
passer_position2-receiver_position2 D	✓	✓	✓	✓	✓

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

D. Team level evidence: passes and winning

In this appendix, we show the correlation between passing intensity and team performance. For this purpose, we aggregated our estimation dataset to the level of teams and half-seasons. We have N=1,568 observations (16 time periods, 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_apc* defined as log(average pass count per game).

First, we only include league dummies, and show a cross-section correlation for a single half-season (2015-16, H1).

$$Average\_points_{team} = \beta \ln apc_{team} + \eta_{league} \tag{.1}$$

Then, we estimate a panel fixed effects model adding league-half-seasons and team fixed effects: $team$ and t for half-seasons:

$$Average\_points_{team,t} = \beta \ln apc_{team,t} + \eta_{team} + \theta_t \tag{.2}$$

Columns (1) and (2) have points per game, Column (3) has log(points per game) as the dependent variable for easier interpretation. Table .10 presents the results.

Table .10: Team level performance and passes

	points_per_game (1)	points_per_game (2)	ln_points_per_game (3)
Pass count per game (ln)	1.196***	0.4434***	0.3487***
	(0.1900)	(0.0707)	(0.0616)
Constant	-5.568***		
	(1.099)		
Observations	98	1,568	1,568
Pseudo R^2	0.39971	0.62163	0.81025
league_season-season_half fixed effects		✓	✓
teamid fixed effects		✓	✓

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

Column (1) 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 still relevant relationship.

We find evidence that when teams pass more, they also tend to win more. In Column (2), we regress points per game (in levels) on log total passes, team and league by half-seasons fixed effects. Conditioning on league specific aggregate trends, in half-seasons when a team passes 10% more than its average pass frequency, it tends to win a 0.044 point (or 3.5%) more than average. Over 38 games, this is almost 2 points (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.

References

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

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

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

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

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

Arcidiacono, P., Kinsler, J. and Price, J. (2017), 'Productivity spillovers in team production: Evidence from professional basketball', *Journal of Labor Econonomics* **35**(1), 191–225. 5

Berge, L. (2018), Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm, Working Paper 13. 20

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. 7

Buchholz, M. (2021), 'Immigrant diversity, integration and worker productivity: uncovering the mechanisms behind 'diversity spillover' effects', *Journal of Economic Geography* **21**(2), 261–285. 7

Calder-Wang, S., Gompers, P. A. and Huang, K. (2021), Diversity and performance in entrepreneurial teams, Working Paper 28684, National Bureau of Economic Research. 8

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

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

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

Desmet, K. and Ortuño-Ortín, I.and Wacziarg, R. (2017), 'Culture, ethnicity, and diversity', *American Economic Review* **107**(9), 2479–2513. 2

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. 7

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* . 5

Fally, T. (2015), 'Structural gravity and fixed effects', *Journal of International Economics* **97**(1), 76–85. 20, 21

FIFA (2021), *Commentary on the Rules Governing Eligibility to Play for Representative Teams*, FIFA: International Federation of Association Football. 35

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. 8

Gauriot, R. and Page, L. (2019), 'Fooled by performance randomness: Overrewarding luck', *The Review of Economics and Statistics* **101**(4), 658–666. 5

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. 2, 11, 18

Hinz, J., Stammann, A. and Wanner, J. (2021), State Dependence and Unobserved Heterogeneity in the Extensive Margin of Trade, CEPA DP 36, Center for Economic Policy Analysis. 20

Hjort, J. (2014), 'Ethnic divisions and production in firms', *The Quarterly Journal of Economics* **129**(4), 1899–1946. 7, 8

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. 4

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. 4, 7

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. 2

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. 4, 7

Keane, M. P., Todd, P. E. and Wolpin, K. I. (2011), The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications, *in* O. Ashenfelter and D. Card, eds, 'Handbook of Labor Economics', Vol. 4, Elsevier, pp. 331–461. 15

Keane, M. and Wolpin, K. I. (2009), 'Empirical applications of discrete choice dynamic programming models', *Review of Economic Dynamics* **12**(1), 1–22. 15

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

Lang, K. (1986), 'A language theory of discrimination', *Quarterly Journal of Economics* **101**(2), 363–382. 7

Laurentsyeva, N. (2019), From friends to foes: National identity and collaboration in diverse teams, Working Paper 226. 8

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

Lazear, E. (1999*a*), 'Language and culture', *Journal of Political Economy* **107**(6), S95–S126. 2, 7

Lazear, E. P. (1999*b*), 'Globalisation and the market for team-mates', *The Economic Journal* **109**(454), 15–40. 2, 7

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

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

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

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. 5

Neeley, T. (2015), 'Global teams that work', *Harvard Business Review* . 2

Nüesch, S. and Haas, H. (2013), 'Are multinational teams more successful?', *International Journal of Human Resource Management* **23**(15), 3105–3115. 3, 4

Ottaviano, G. I. and Peri, G. (2005), 'Cities and cultures', *Journal of Urban Economics* **58**(2), 304–337. 7

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. 7

Parsons, C. A., Sulaeman, J., Yates, M. C. and Hamermesh, D. S. (2011), 'Strike three: Discrimination, incentives, and evaluation', *American Economic Review* **101**(4), 1410–1435. 5

Santos-Silva, J. and Tenreyro, S. (2021), The log of gravity at 15, Discussion Paper 1, School of Economics, University of Surrey. 20

Spolaore, E. and Wacziarg, R. (2016), Ancestry, language and culture, *in* 'The Palgrave Handbook of Economics and Language', Springer, pp. 174–211. 2

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. 7

Todd, P. and Wolpin, K. I. (2010), 'Structural estimation and policy evaluation in developing countries', *Annual Review of Economics* **2**, 21–50. 15

Tovar, J. (2020), 'Performance, Diversity And National Identity Evidence From Association Football', *Economic Inquiry* **58**(2), 897–916. 3, 4

Weidner, M. and Zylkin, T. (2021), 'Bias and consistency in three-way gravity models', *Journal of International Economics* **132**, 103513. 20

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1872	Ihsaan Bassier	Firms and inequality when unemployment is high
1871	Francesco Manaresi Alessandro Palma Luca Salvatici Vincenzo Scrutinio	Managerial input and firm performance. Evidence from a policy experiment
1870	Alberto Prati	The well-being cost of inflation inequalities
1869	Giulia Giupponi Stephen Machin	Company wage policy in a low-wage labor market
1868	Andrew Seltzer Jonathan Wadsworth	The impact of public transportation and commuting on urban labour markets: evidence from the new survey of London life and labour, 1929-32
1867	Emanuel Ornelas John L. Turner	The costs and benefits of rules of origin in modern free trade agreements
1866	Giammario Impullitti Syed Kazmi	Globalization and market power
1865	Nicolás González-Pampillón Gonzalo Nunez-Chaim Henry G. Overman	The economic impacts of the UK's eat out to help out scheme
1864	Rocco Macchiavello Ameet Morjaria	Acquisitions, management and efficiency in Rwanda's coffee industry
1863	Andrew E. Clark Conchita D'Ambrosio Jan-Emmanuel De Neve Niccolò Gentile Caspar Kaiser Ekaterina Oparina Alexandre Tkatchenko	Human wellbeing and machine learning

1862	Fabrizio Leone Rocco Macchiavello Tristan Reed	Market size, markups and international price dispersion in the cement industry
1861	Elias Einio Josh Feng Xavier Jaravel	Social push and the direction of innovation
1860	Xiao Chen Hanwei Huang Jiandong Ju Ruoyan Sun Jialiang Zhang	Endogenous cross-region human mobility and pandemics
1859	Xavier Jaravel Danial Lashkari	Nonparametric measurement of long-run growth in consumer welfare
1858	Leonardo Bursztyn Jonathan Kolstad Aakaash Rao Pietro Tebaldi Noam Yuchtman	Political adverse selection
1857	Attila Lindner Balázs Muraközy Balázs Reizer Ragnhild Schreiner	Firm-level technological change and skill demand
1856	Jeremiah Dittmar Ralf R. Meisenzahl	The university, invention and industry: evidence from German history
1855	Donna Brown Jonathan Wadsworth	Accidents will happen: (de)regulation of health and safety legislation, workplace accidents and self employment
1854	Fabrizio Leone	Foreign ownership and robot adoption