Cultural homophily and Collaboration in Global Teams

Gábor Békés* and Gianmarco I.P. Ottaviano**

*CEU, KRTK KTI and CEPR
** Bocconi University, Baffi-CAREFIN, IGIER, CEP and CEPR

CEU MRSS Seminar // February 2022

Introduction: Globalization and homophily in teams

- Globalization - mixing the best of global expertise in multinational teams
  - Diversity benefits: learning, innovation
  - Hurdles: communication, trust
- Interaction between people of different cultural background key to understand function of teams
- Homophily is association of similar people (shared cultural background)
- What we know most: how teams are formed, tie-formation, friendship networks
- Our focus is collaboration (work for a common purpose)

Introduction: Globalization and homophily in teams

- How does cultural **homophily** – barriers related to **culture** – affect **collaboration** in multinational **teams**?

Introduction: Measuring *choice* homophily

- Homophily = Opportunity (**induced**) + Preference (**choice**)
  - Opportunity: mechanically induced - distributions across categories define the probability they choose similar others
- Challenge: partial out induced homophily to measure choice homophily in a setup with external validity to modern workplaces
  - Option A: With experiment / random team formation
  - Option B: With observational data / modeling baseline.

Introduction: Using football data to capture collaboration

- Teams: use professional football - top European leagues
- Collaboration: measured by pass rate between a pair of players
  - Team (squad) composition is exogenous to players
  - Collaboration is individual choice

- Ideal setting:
  - Observe collaboration repeatedly in great detail
  - Collaborative with well defined objectives and roles.
  - Rules are simple and known - allow calculating baseline
  - Global workplace: several countries, players from 130 countries

Homophily: induced and choice

Measuring cultural homophily

- Cultural background (culture') – set of cultural traits
- Language, norms, values and attitudes learnt at home
- Proxy through nationality and colonial (empire) legacy.

- Team members of same culture collaborate more than team members of different culture.
  - 'Border effect' hamper collaboration

How we measure choice homophily: Data and model

- Exhaustive dataset recording passing events from professional football
  - Spain, Germany, France, Italy and England, 8 seasons
  - 7000 Players from 132 countries, characteristics
  - 10.7 million passes (passer, receiver: ID, location)
- Baseline – discrete choice model of players' passing behavior
  - Pass rate for a pair of players is pinned down by their characteristics and opportunities during the matches
  - Estimate directly in model.

Related literature

- Diversity in teams Lazear (1999) Lang (1986)
- Diverse environments, performance in cities, plants: Ottaviano and Peri (2006, 2005) Buchholz (2021)
- Diversity in teams: team performance and composition
  - Hockey: Kahane et al. (2013), Football: Nüesch and Haas (2013), Tovar (2020).
  - Ethnic conflict: Hjort (2014), Laurentsyeva (2019),
  - Team formation process Calder-Wang et al. (2021)
- Homophily in finding partners in scientific publishing: Freeman and Huang (2015), AlShebli et al. (2018)
- Homophily network friendships Currarini et al. (2009, 2010)
- Review psychology to management (Lawrence and Shah, 2020; Ertug et al., 2021).

Contributions

1. Focus on everyday workplace collaboration - high skilled, lowly charged context in a developed area with no real conflicts.
2. Well defined measure of collaboration at individual level through time (not rare pair formation)
3. Careful model of baseline, both theory and empirics (go beyond randomization)
4. Rich and precise measures of individual characteristics
5. Very large, global sample - external validity

Preview of results

- Baseline difference is 6.6%
- Once partialling out induced homophily, player pairs of same culture have a 2.5% higher pass rate (*choice homophily bias*)
- Culture: both same nationality and colonial (empire) background matters.
- Homophily bias is higher for ...
  - deep collaboration – intensive passing (vs one shot passes)
  - young players (vs veterans)
  - players playing in a foreign country (vs domestic)

Data collection and definitions

Data: scope

- 5 top leagues (France, Germany, Spain, Italy, England),
- 8 seasons (2011/12-2018/19) every teams play with every other twice
  - 20 (18) teams per league, 14,608 games in total
  - 800 passes/game/team
- teams, 154 different team, each with 25-30 strong squad, regular churning (twice a year)
- events: 10.7 million passes in total
- players: 7,000, from 132 countries

Raw data 1: events

- Event data as xml file
- Structured text, events with features, qualifiers
  - separately recorded with a timestamp
  - A pass between players, type of pass, success
- Webscraping
  - originally created by OPTA (cameras+algorithms + humans)
  - Scraped from a sports website
  - Pass events separated

Data as text: people (players)

- Webscraping
- Source: Transfermarkt, a player information database
- Player characteristics:
  - nationalities,
  - position in team
  - age, height
  - player valuations – over time

Data as text: names

- Player characteristics:
- All common aspects of working with names

- Entity resolutions (accents, middle names, nick-names)
- Combing datasets: entity resolution / coreference
  - Matching algorithm by motifs

- Entity resolution II: country names

Data: defining culture

- Culture ...
- Same nationality (citizenship)
- Shared history - colonial (empire) legacy
  - Argentina-Spain, England-Egypt (ruler and colony)
  - Uruguay-Argentina (colony siblings),
  - Ukraine-Russia, Northern Ireland-Ireland (same country once)
- Same language not colony
  - Switzerland and Germany, Congo and France

Homophily: induced and choice 2 . colony siblings

Data: Same nationality definition

- 26% of players have two or three nationalities
  - Born in a country and moved to another as minor and got nationality with family (Argentina and Spain)
  - Parents have multiple nationalities (French and Algerian)
- Same nationality definition = two players have **a** common nationality

Data: Same nationality definition

Man City players Ryad Mahrez(Alg, Fr) and Aymeric Laporte(Fr, Es) in 2020

Data: aggregation

- From a choice model, aggregate to relative frequencies
- Aggregate to half-seasons (16-20 games) – compromise
  - Squads are large, only 11 players at field at once, lot of variation across games, selection major issue for a single game.
  - Noise is high / randomness of games
  - There is churning in mid-season
  - Player quality stable in half-season / vary long run
- Object of interest = pass count for player pairs
- compared to total passes by passer when both on pitch + total passes at team.

Model

Model: setup

- Football team $N = 11$ players, two players indexed o, d.
- The passer's decision = problem of passing the ball to the receiver who generates the highest expected benefit for the *team*.

- Game = series of short units of time (t) up to T ('periods').
- $T^{o,d}$ is subset of passing episodes: both players are on the pitch, player o has ball possession

- A 'pass' (o, d, t) = player o ('passer') to teammate d ('receiver'). Started by o in t, received by d in $t + 1$

The modeling task

- Discrete choice dynamic programming model
- to disentangle choice from opportunity in an internally consistent way

- controlling for
  - observable player characteristics (such as team, position, valuation, citizenship)
  - pass features (such as average distance)

Model: The passer's decision

Value of having a ball = two components: what the player may do + option value of a pass.

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

Model: The passer's decision

Value of having a ball = two components: what the player may do + option value of a pass.

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

- $\ln u_t^d$ = benefit due to player d's characteristics
- z_t^d = random part (shock'), Gumbel (κ)
- $\widetilde{c}^{o,d}$ = challenges – 'passing cost'
- φ^d = probability of successful pass to receiver d
- β = relative importance the team attaches to passing in general (style)

Model: The player's decision 2

- The probability that player o with ball in t successfully passes to teammate d depends on
  - the (relative) value of both players
  - ability to pass/receive a pass successfully
  - cost of the pass

- Aggregation to T - probabilities to relative frequency
- Probability = the average share of successful passes that player o makes to player d per episode over a half-season
  - in the subset of time (passing episodes) $T^{o,d}$ when both o and d are fielded and player o has ball possession

Probability of a successful pass

$$\pi_t^{o,d} = \left(V_{t+1}^d\right)^{\kappa\beta\varphi^d} \left(c^{o,d}\right)^{-\kappa} \left(\Lambda_{t+1}^o\right)^{-\kappa} \tag{2}$$

$$\text{with } \Lambda_{t+1}^o \equiv \left[\sum_{s=1}^{N} \left(V_{t+1}^s\right)^{\kappa\beta\varphi^s} \left(c^{o,s}\right)^{-\kappa}\right]^{\frac{1}{\kappa}} \tag{3}$$

- V_{t+1}^d: player d controlling the ball in period $t+1$ ($V_t^o \equiv \exp E[U_{t+1}^o]$)
- φ^d his probability of taking control of the ball
- $c^{o,d}$ the difficulty of passing the ball to him
- Λ_{t+1}^o: 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 ($c^{o,s}$).

Model: Passing cost

Model passing cost

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

- $g^{o,d}$ captures all distance-related frictions
- $I^{o,d}$ captures all non-distance-related frictions.
  - effect of same/different cultural traits
  - also: experience in playing together, physical difference
  - May be high if o and d find it hard to collaborate
- In model, assume separability (true empirically)

Model: pass rate over a half-season

Define

- P: total number of team passing episodes
- $P^{o,d}$ number of passing episodes involving a pass from o to d
- $T^{o,d}$: the number of passing episodes when both o and d are fielded and player o has ball possession.

Half-season 'pass rate' $p^{o,d} = T^{o,d}\pi_t^{o,d}/P$, and thus

$$\log p^{o,d} = \log T^{o,d} - \log P - \kappa \log \Lambda_{t+1}^{o} + \kappa\beta\varphi^{d} \log V_{t+1}^{d} \quad (5)$$
$$-\kappa\gamma \log g^{o,d} - \kappa\lambda \log I^{o,d}$$

Model: pass rate over a half-season

The pass rate in a half-season depends on

- P: total number of team passing episodes
- $T^{o,d}$: the number of passing episodes when both o and d are fielded and player o has ball possession.
- Passer characteristics
- Receiver characteristics
- Passing frictions
- Homophily

Estimation model and results

Model Estimation

- Homophily: $SameCult_{o,d}$ as the same culture indicator (0/1).
- Model pass rate = f(player characteristics, position, passing friction, homophily)
- Poisson model for count of passes conditional on total passes
- Poisson (PPML-FE) with many fixed effects has several advantages over ln *count* (Fally, 2015; Santos-Silva and Tenreyro, 2021)
  - Result is robust to OLS with ln *count*

Estimated model: Frictions

We capture distance-related frictions that make it difficult to pass the ball from o to d

$$PassFric_{o,d,t} = \gamma_1 PassDist_{o,d,t} + \gamma_2 Forwardness_{o,d,t} + \eta Position_o Position_d \quad (6)$$

- a half-season $PassDist_{o,d,t}$ is the average distance of passes between the two players
- $Forwardness_{o,d,t}$ is the share of passes between the two players with a forward direction
- $Position_o Position_d$ is a dummy variable capturing the two players positions

Estimated models: Poisson with player characteristics / FE

(1) Poisson model with passer and receiver characteristics

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + PassFric_{o,d,t} + \ln T_{o,d,t} + \sum_{j=o}^{d}(\eta_j value_{j,t} + \theta_j playerchar_{j,t})) \quad (7)$$

(2) Poisson model with passer*half-season + receiver*half-season FE

$$E(pass\_count_{o,d,t}|...) = exp(\delta SameCult_{o,d} + PassFric_{o,d,t} + \ln T_{o,d,t} + \upsilon_{o,t} + \upsilon_{d,t}) \quad (8)$$

Estimated models: Poisson with player characteristics / FE

Arsenal: Bellerin and Cazorla and Ozil and Cazorla

Estimated models: Poisson with culture traits

- Poisson model with passer*half-season and receiver*half-season FE,
- Same culture indicator $SameCult_{o,d}$ separately in three traits:
  - $SameNat_{o,d}$ Same nationality
  - $SameCol_{o,d}$ Same colonial legacy but not same nationality
  - $SameLan_{o,d}$ Same language but no colonial legacy, not same nationality

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

Results: Baseline models

Dep.var: Pass count	(No ctrls)	(Player chars)	(Player FE)
Same culture (any) (0/1)	**0.0655*****	0.0204***	**0.0250*****
	(0.0091)	(0.0038)	(0.0042)
Average length of passes (ln)		-0.6759***	-0.7944***
		(0.0077)	(0.0094)
Average forwardness Ind (0-1)		0.0066	0.0143
		(0.0077)	(0.0099)
Receiver valuation (ln)		0.0103***	
		(0.0015)	
Team FE	yes	yes	yes
Passer, Receiver-position FE		yes	yes
Passer, Receiver characteristics		yes	
Passer-half_season, Receiver-half_season FE			yes
Pseudo R^2	0.078	0.741	0.759

Poisson regression model. N= 668,108. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Player features: position, value, total passes, citizenship. Exposure=ln shared passes together

Result discussion

- (Unconditional) Same culture players tend to pass 6.5% more compared to different culture players
- Partialing out baseline homophily: it is around 2.5%
  - Robust to a variety of specifications I.- confounders/selection
    - More variables on passes (direction)
    - Physical differences
    - Assortative matching
    - Experience with club
  - Robust to a variety of specifications II.- functional form
    - Allowing minutes coefficient to vary (1.07)
    - log count as dependent variable.

Results: Cultural traits

Dep.var: Pass count	(No ctrls)	(Player chars)	(Player FE)
Same nationality (0/1)	0.0799***	0.0238***	**0.0301*** **
	(0.0102)	(0.0042)	(0.0048)
Same colonial legacy (0/1)	0.0140	0.0227***	**0.0234*** **
	(0.0149)	(0.0058)	(0.0066)
Same language only (0/1)	0.0501**	0.0008	**0.0066**
	(0.0213)	(0.0089)	(0.0095)
Team FE	yes	yes	yes
Passer, Receiver-position FE		yes	yes
Passer, Receiver characteristics		yes	
Passer-half_season, Receiver-half_season FE			yes
Pass direction, distance	yes	yes	yes
Pseudo R^2	0.07835	0.74164	0.75930

Poisson regression model. N= 668,108. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Player features: position, value, total passes, citizenship. Exposure=ln shared passes together.

Deeper collaboration

- Passing is collaboration
- Deeper collaboration = pass sequences (like ABAB)
- For deeper collaboration, trust/understanding/taste should be more important
- Instead of pass count: count of pass sequences

Results: deeper collaboration – pass sequence

	pass seq count sequences (1)	complex pass seq count (2)
Same culture (any) (0/1)	0.0209***	**0.0501***
	(0.0039)	(0.0071)
Pseudo R^2	0.74590	0.55971

Poisson regression model. N= 668,108. Standard errors, clustered at player 1 level, are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. **Both columns**: Player1*half-season FE, Player2*half-season FE+ friction. Exposure=ln shared passes together

Deeper collaboration - result discussion

- Deeper collaboration = pass sequences (like ABAB)
- Same culture premium is substantially higher for more complicated passes. (5% vs 2%)
- For deeper collaboration, trust/understanding/taste is more important: same cultural background pairs collaborate substantially more.

Heterogeneity of homophily premium - by passer

- Heterogeneity from the passer's point of view
- Various features - linked to theory in management or labour economics
  - Experience (age)
  - Being in minority or majority (play at home or abroad)
  - Status (valuation)

Heterogeneity of homophily premium

Heterogeneity source		Freq	Coeff %	Sig
Nationality = league	Home national	57%	**1.70**	
	Foreign national	43%	**3.40**	yes
Player values (eur)	Low (below 3.5m, avg=1m)	27.4%	2.17	
	Medium (1-16m, avg=4.3m)	48.4%	2.60	no
	High (16m+, avg=22m)	22.2%	2.62	no
Age category (ys)	Veteran (29.3+, avg=31.9)	25%	**1.99**	
	Experienced (23-29, avg=26)	50%	2.44	no
	Young (below 23ys, avg=21)	25%	**3.38**	yes
With same club (day)	Low (below 164, mean: 75)	25%	2.29	
	Medium (165-959, mean 484)	50%	2.70	no
	High (960+, mean: 1850)	50%	2.37	no

Baseline Poisson FE regression model, defined by the passing player characteristic. Base is first line. Sig: difference statistically significant at 5%.

Result discussion: Heterogeneity

Heterogeneity from the passer's point of view

- Age: younger players have twice the bias
- Minority: players who play abroad have also twice the bias
- Status: The bias does not depend on the quality of players (valuation)

Results summary

- Evidence of homophily: player pairs of same culture pass more
- Same nationality / same historical (colonial) legacy similar.
- More likely engaged in deeper collaboration

Summary

- We isolated choice homophily to learn about any possible premium of cultural homophily for workplace collaboration
- Culture: same nationality and legacy of shared history via colonial rule or past in a union.

- Homophily is pervasive even in teams of
  - very high skill individuals
  - with clear common objectives and aligned incentives
  - and involved in well defined tasks
  - activities not particularly language-intensive.