Diversity and Group Performance:

Evidence from the World's Top Soccer League

Online Appendix

Appendix	Pages
A1: Average Tournament Scores by Team (2003-2012)	A
A2: Bivariate Correlations between Aggregate Team Statistics (2003-2012)	B
A3: Relationship between Transfer Value and Talent (Player-Level Analysis)	C
A4: Club versus National Team Performance	D-F
A5: Coaches' Impact	G-H
A6: Alternative Approach (Instrumental Variables Analysis)	J-N

Appendix 1: Average Tournament Data by Team (2003-2012)

Team	League	Average Goal Differential	Winning Percentage	Linguistic Distance	Country Fractionalization	Total Transfer Value	Average Transfer Value	Elo Rating	Number of UEFA Cups	Quota
Barcelona	La Liga	1.08	57.48	43.67	0.30	400.56	13.39	1971.89	3.00	9.00
Bayern M.	Bundesliga	0.82	54.70	64.77	0.25	228.75	8.47	1859.89	1.11	9.00
Chelsea	Premier	0.78	48.75	71.64	0.14	358.89	11.57	1909.80	0.00	10.00
Sevilla	La Liga	0.74	57.50	61.21	0.16	182.50	5.21	1855.50	3.00	2.00
Malaga	La Liga	0.70	40.00	42.47	0.18	110.00	3.44	1713.00	3.00	1.00
R.Madrid	La Liga	0.69	52.50	51.51	0.26	365.56	11.98	1928.70	3.00	10.00
Man Utd	Premier	0.66	51.29	61.49	0.25	312.78	9.08	1928.20	0.00	10.00
Liverpool	Premier	0.60	49.37	71.47	0.14	232.50	7.27	1866.50	0.00	6.00
Arsenal	Premier	0.60	49.07	72.58	0.12	254.44	6.98	1873.70	0.00	10.00
Lyon	Ligue 1	0.51	49.17	46.23	0.38	166.88	5.28	1807.78	4.00	9.00
Monaco	Ligue 1	0.50	48.08	59.55	0.16	90.00	3.10	1749.00	4.00	2.00
Juventus	Serie A	0.44	48.47	53.54	0.34	252.00	7.94	1826.67	1.50	6.00
Schalke	Bundesliga	0.38	39.79	74.60	0.17	120.63	3.92	1739.75	1.25	4.00
Napoli	Serie A	0.38	50.00	53.34	0.21	180.00	6.00	1679.00	1.00	1.00
Milan	Serie A	0.34	42.47	59.78	0.29	233.13	7.13	1829.44	1.78	9.00
Valencia	La Liga	0.31	35.14	38.43	0.38	172.50	5.93	1826.33	3.00	6.00
Tottenham	Premier	0.30	40.00	65.94	0.20	270.00	7.11	1766.00	0.00	1.00
Fiorentina	Serie A	0.27	45.83	55.97	0.29	172.50	5.17	1733.50	1.00	2.00
Marseille	Ligue 1	0.23	31.67	49.23	0.23	129.00	4.12	1738.50	4.00	6.00
Inter	Serie A	0.20	46.65	61.90	0.16	255.63	7.68	1838.78	1.78	9.00
PSG	Ligue 1	0.18	38.33	57.32	0.28	170.25	5.42	1751.50	4.00	2.00
Bordeaux	Ligue 1	-0.09	48.89	49.00	0.39	103.67	3.73	1748.33	4.00	3.00
Dortmund	Bundesliga	-0.12	35.26	60.85	0.38	210.00	7.36	1877.00	0.00	2.00
Man City	Premier	-0.17	25.00	72.37	0.17	435.00	12.13	1849.00	0.00	2.00
Roma	Serie A	-0.21	40.00	50.64	0.35	192.00	6.31	1794.80	2.20	5.00
Celta Vgo	La Liga	-0.25	25.00	39.52	.	.	.	1789.00	3.00	1.00
R.Sociedad	La Liga	-0.25	25.00	55.29	.	.	.	1841.00	3.00	1.00
Stuttgart	Bundesliga	-0.32	30.56	64.17	0.29	102.75	3.02	1734.00	0.00	3.00
Udinese	Serie A	-0.33	33.33	47.70	0.37	74.00	2.31	1699.00	3.00	1.00
W Bremen	Bundesliga	-0.35	34.03	61.35	0.28	120.08	3.97	1778.83	1.67	6.00
Leverkusen	Bundesliga	-0.38	37.50	52.97	0.39	100.00	3.39	1773.00	2.50	2.00
Lille	Ligue 1	-0.39	18.75	52.06	0.28	96.38	3.33	1751.50	4.00	4.00
A.Madrid	La Liga	-0.44	18.75	40.16	0.28	182.50	6.79	1807.50	3.00	2.00
Lazio	Serie A	-0.58	16.67	56.14	0.30	94.50	2.95	1766.00	1.50	2.00
Villarreal	La Liga	-0.64	18.33	21.61	0.33	112.33	3.73	1812.67	3.00	3.00
Deportivo La Coruna	La Liga	-0.67	25.00	18.54	0.53	125.00	4.81	1865.00	3.00	2.00
Real Betis	La Liga	-0.67	33.33	23.68	0.59	69.50	2.78	1782.00	3.00	1.00
Montpelr	Ligue 1	-1.00	0.00	43.25	0.43	69.00	2.16	1727.00	4.00	1.00
Hamburg	Bundesliga	-1.33	16.67	69.38	0.15	89.00	2.70	1691.00	0.00	1.00
Auxerre	Ligue 1	-1.50	16.67	48.14	0.33	58.50	1.46	1724.00	4.00	1.00
Median		**0.19**	**39.06**	**54.41**	**0.28**	**171.38**	**5.25**	**1791.90**	**2.75**	**2.50**

Appendix 2: Bivariate Correlations between Aggregate Team Statistics in (2003-2012)

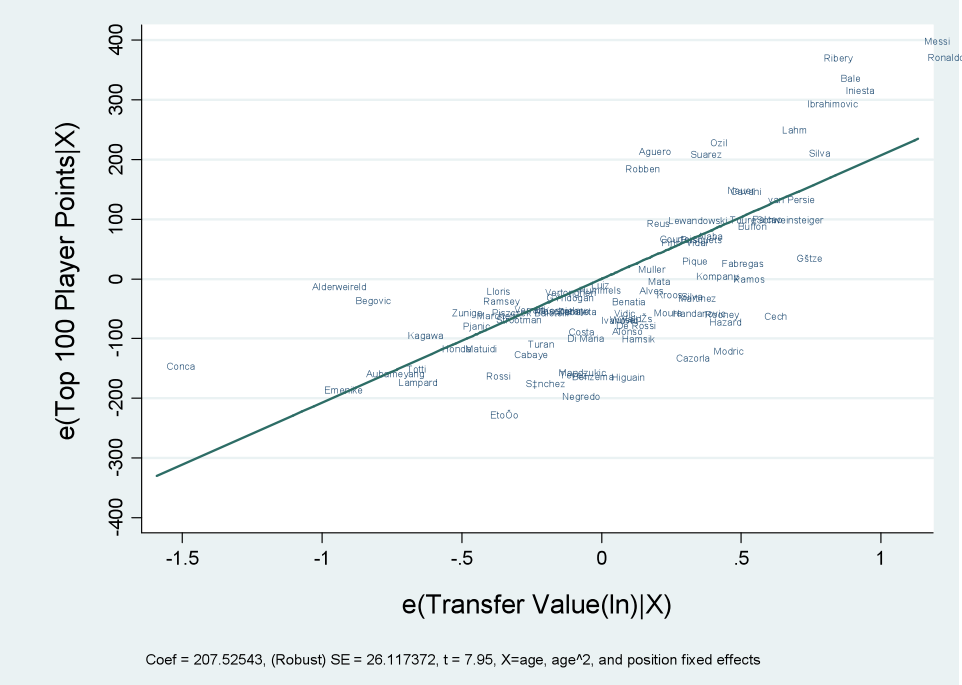
Variables	1	2	3	4	5	6	7	8	9
1. Average Goal Differential	1								
2. Winning Percentage	0.8670*	1							
3. Linguistic Distance	0.3359*	0.3055	1						
4. Country Fractionalization	-0.4478*	-0.2986	-0.7783*	1					
5. Total Transfer Value	0.6127*	0.5172*	0.3283*	-0.3592*	1				
6. Average Transfer Value	0.6268*	0.5351*	0.2495	-0.2869	0.9815*	1			
7. Elo Rating	0.4999*	0.4387*	0.0579	-0.0609	0.7714*	0.8111*	1		
8. Quota	-0.2	-0.2071	-0.6684*	0.4708*	-0.4244*	-0.3641*	-0.2221	1	
9 Number of UEFA Cups	0.5989*	0.5735*	0.2111	-0.2045	0.6094*	0.6355*	0.6650*	-0.1664	1

* Significant at the 0.05 level.

Appendix 3: Relationship between Transfer Value and Talent
(Player-Level Analysis)

Dependent Variable: Points awared to play er in Guardian Top 100 Index	(1)	(2)	(3)
AverageTransfer Value (ln)	94.819***	221.026***	207.525***
	(28.721)	(27.124)	(26.117)
Player Age		-173.932***	-160.334***
		(37.631)	(35.109)
Age Squared		3.789***	3.516***
		(0.720)	(0.672)
Defender			-16.955
			(37.010)
Forward			56.520
			(40.462)
Midfield			19.930
			(38.274)
Constant	-171.249*	1,338.211***	1,195.170***
	(88.400)	(438.333)	(408.973)
Observations	100	100	100
R-squared	0.197	0.574	0.601
RMSE	139.1	102.3	100.6

Robust standard errors, clustered at team level in parentheses, *** p<0.01, ** p<0.05, * p<0.1.



Coef = 207.52543, (Robust) SE = 26.117372, t = 7.95, X=age, age^2, and position fixed effects

Appendix 4: Club versus National Team Performance

An interesting implication of our theory and findings is that, on average, the *UEFA Champions League's* club teams should be able to outperform their countries' national teams if such head-to-head competition was possible. Paine (2014) uses the individual plus/minus player ratings that underlie *ESPN's Soccer Power Index* (SPI) rankings to compare the 2014 Brazil World Cup national teams with the wealthiest club teams in the "Big Five" European leagues. By aggregating the SPI ratings for an entire roster (weighting by the number of minutes each player was on the pitch), he derives an indicator to describe the amount of talent on any team, whether it plays for club or country.[1]

Paine (2014) finds that the three best club teams in the world in 2014 (Real Madrid, FC Barcelona and Bayern Munich) were a cut above the four most talented national teams at the onset of the World Cup (Brazil, Argentina, Germany and Spain). FIFA's eligibility rules determine which national team a player may represent in international soccer. Any player who is a citizen, or a naturalized citizen of a country, is eligible to play for a national team of that country. Still, according to FIFA, players must be able to demonstrate a "clear connection" to a country that they had not been born in but wished to represent.[2] As such, national teams are both nationally and linguistically much more homogeneous than their club-level counterparts. For example, the average share of foreign-born players in the 12 club teams analyzed by Paine was

[1] As Paine (2014) notes, SPI calculates the talent level of every team based on the skill of its component players. To that end, each player is assigned a per-minute effect rating determined both by his individual statistics and by how his team performs with him in the lineup (compared to its performance without him).

[2] The FIFA ruling explicitly states that, in such scenarios, the player must have at least one parent or grandparent who was born in that country, or the player must have been resident in that country for at least two years.

61%, compared to an average share of 13% in the case of the 32 teams that participated in the 2014 World Cup.

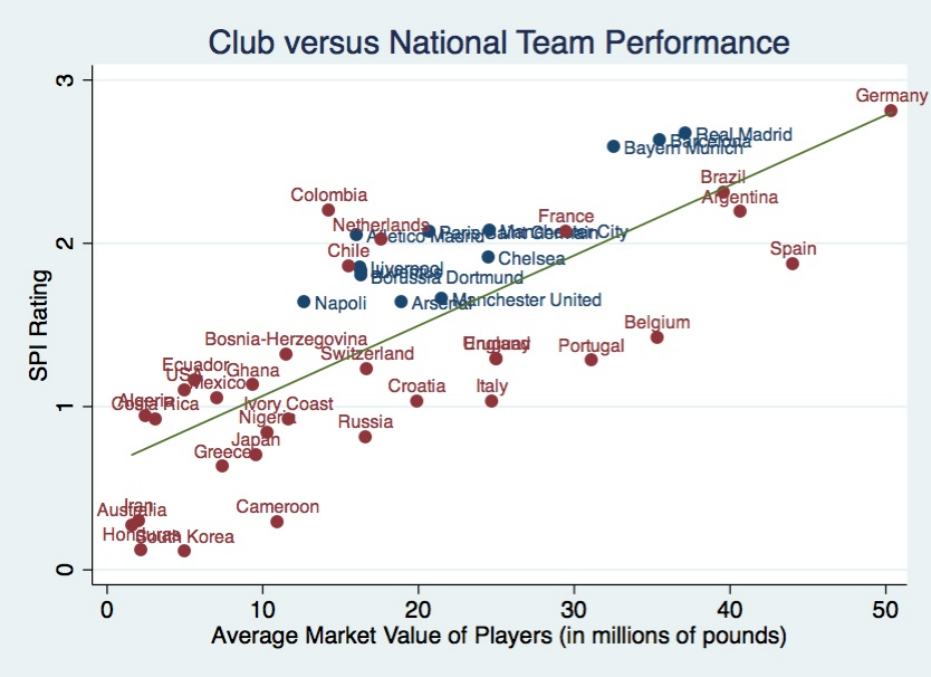


Figure A4: Scatterplot comparing the performance of national and club teams.
Label colors differentiate the type of team (Blue=Club, Red=National).

Of course, given their financial wherewithal, some of the richest club teams have the ability to recruit more talent than some of the top-tier national teams. The analysis presented in Table A4 below indicates that, after controlling for the average market value of each team's players (using *Transfermarkt* valuations), there are significant advantages associated with a more diverse roster. Holding their players' average value constant, the SPI rating of the typical national team is roughly one-standard deviation lower than that of a typical club team. To illustrate this result, consider the comparison between Chelsea and England's national team. Chelsea's estimated SPI rating is 1.91, whereas England's is only 1.29. On average, their players are equally worth (14.38 and 12.77 million pounds, respectively). But, while close to 80% of

Chelsea's players were born outside of England, only single player (our of the list of 23) in England's national team was born elsewhere. (Liverpool's Raheem Shaquille Sterling, born in Kingston, Jamaica). Our results imply that if Chelsea were to become as "homogeneous" as England's national team, its predicted SPI score would drop to from 19.1 to 1.47 (which would put the team below Arsenal and Manchester United, and slightly ahead than Belgium's national team).

Table A4: Club versus National Team Performance (OLS Regression)

ESPN SPI Rating	**National v. Club** (1)	**% Foreign** (2)
Average Transfer Value (ln)	0.038***	0.043***
	(0.004)	(0.005)
National Team=1; Club Team=0	-0.624***	
	(0.092)	
% Foreign Players		0.008***
		(0.002)
Constant	1.164***	0.415***
	(0.107)	(0.129)
Observations	45	45
R-squared	0.72	0.67
RMSE	0.385	0.42

Robust standard errors, clustered at team, level in parentheses, *** p<0.01, ** p<0.05, * p<0.1.
The dependent variable is the ESPN SPI rating for 2014.

Paine, Neil. 2014. "Could the World Cup Champion Beat the Best Club Team in the World?" <http://fivethirtyeight.com/features/could-the-world-cup-champion-beat-the-best-club-team-in-the-world/>

Appendix 5: Club versus National Team Performance (OLS Regression)

Dependent variable: Goal Differential	**Diversity**	**Average**	**Foreign**	**Experienced**
	(1)	(2)	(3)	(4)
Linguistic Distance	0.023*	0.023**	0.024**	0.022**
	(0.012)	(0.009)	(0.009)	(0.009)
Average Value (ln)		0.832***	0.844***	0.627**
		(0.269)	(270)	(0.237)
Foreign Coach			-0.425	
			(0.324)	
Constant	-1.799**	-3.051***	-3.097***	-2.726**
	(0.691)	(0.667)	(0.676)	(0.615)
Year FE	Yes	Yes	Yes	Yes
League FE	Yes	Yes	Yes	No
Coach FE	Yes	Yes	Yes	Yes
Observations	168	152	152	80
R-squared	0.668	0.744	0.751	0.629
RMSE	0.603	0.556	0.552	0.465

Robust standard errors, clustered at team level in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the average goal differential of a team in the UEFA Champion's league in a given year. Model 1 is the bivariate correlation. Model 2 controls for average player value. Model 4 only includes teams whose coaches participated at least 4 times in the competition.

Appendix 6: Alternative Approach (Instrumental Variables Analysis)

As an alternative analysis, we replicate our analysis using data from every team in each of the top-five European soccer leagues, regardless of whether they qualified for the Champions League or not. The dataset, collected by Miguel et. al (2009) contains information about every player in each team for the 2004/2005 and 2005/2006 seasons. We treat teams as the unit of analysis by aggregating the player-level data. In each season, teams have a single observation of their seasonal performance. In total, our sample contains 170 observations of 97 unique teams. For each team, we possess information on its performance, its financial prowess, the demographic characteristics of its locality, and the diversity of its composition (see Table A7).

Table A7: Descriptive Statistics of Key Variables - National Leagues

Variable	n	Mean	SD	Min.	Max.
Dependent Variable					
Per-Game Goal Differential	170	2.99	19.13	-43.00	57.00
Key Causal Variables					
Linguistic Distance	170	43.07	18.24	0.00	77.23
Control Variable					
Average Player Pay in 100,000s of USD (Log	170	0.66	0.56	-0.47	2.54
Average Wage City (Logged)	170	10.00	0.22	9.61	10.72
City Population (Logged)	170	13.210	1.150	10.327	15.506
Instruments					
Germany's Bundesliga	170	0.18	0.38	0.00	1.00
Italian Calcio	170	0.21	0.41	0.00	1.00
English Premier League	170	0.22	0.42	0.00	1.00
France's Ligue 1	170	0.19	0.39	0.00	1.00
Spain's La Liga	170	0.20	0.40	0.00	1.00
Percent Foreign City	170	0.12	0.14	0.01	0.80

Exogenous Sources of Variation in Linguistic Distance

To identify selection, we examine the characteristics of the city in which the selected team is located. Empirical analyses of immigration have long demonstrated an enduring association between previous immigrants from the sending nation and the settlement choices of

new immigrants, leading to ethnic enclaves (Bartel 1998, Funkauser 2000, Chiswick and Miller 2004, Chiswick et al. 2002). More recent research has suggested that minority owned entrepreneurs (McDaniel and Drever 2009) and international business talent may also make locational choices based on the presence of ethnic compatriots or a historical track record for safe treatment of immigrants (Papademetrieu 2008).

For our analysis, this implies that we can identify a portion of the selection process that is exogenous to team performance, as players select teams with a larger presence of compatriot migrants. For instance, a talented Korean star may select a London-based team over a similarly strong Manchester squad because a larger population of Korean immigrants in London is more comforting. Even if the player does not already have friends and family in London, local Korean speakers will help ease his transition into society, helping him find places to live, obtain a driver's license, and learn about the best Korean supermarkets and restaurants. Locations with larger foreign populations may have Korean-language print media and television, as well as greater possibility of familiar social interactions.

Building on this insight, we use the presence of foreign population in a team's location as an instrument in a two-stage regression model. In the first stage, we regress our measure of linguistic distance on the share of foreign population in a city. This allows us to predict the linguistic distance for each team that results solely from differences in existing stocks of foreigners in the cities.

Theoretically, the instrument satisfies the exclusion criterion, because there is little reason to suspect that share of foreign population in a city has direct effect on team performance. Indeed, the bivariate correlation between foreign population in a team's city is 0.38, compared to only 0.20 for the relationship between foreign population and goal differential. Because foreign

population may be proxying for the wealth or market size of a team's location, we control for these variables in both stages of the regression model, allowing us to shut off the most likely alternative causal pathways.

To this end, we collected data for the new analysis from each country's national statistics agency of the country where the team is located. As demographics are assumed to be slow moving, the same data are used for each team in both soccer seasons. Across countries, the administrative borders at which local statistics are available vary in area and aggregation. Furthermore, within countries major metropolitan areas are treated differently than medium and small metropolitan areas in the national statistics data. For instance, data at the level of boroughs are available in London, while the only data for Norwich is at the level of Norfolk County. In this exercise, the local demographic data included are the data at the lowest available level. This preserves variation between teams that may be located near one another.

The cost of including local demographic data pooled from varying zone sizes is that it allows for a potential bias in analysis. A city's percentage of foreigners will likely be higher if the measurement is taken at the city borders than at the county border. Therefore, teams in countries where local data is only available at a higher level may look as if they are located in less diverse areas. Likewise, teams in metropolitan areas where data is at the neighborhood level may seem to be in more diverse areas relative to teams in smaller cities where data is at the county level.

For the purpose of this study, the measurement error is likely to be stochastic, making it more difficult to identify a statistically significant result. The demographic data are used as a metric for what brings players to certain places to play soccer rather than other places. What the data represent is the local area around a team that is important to a player in decision-making. In

this interpretation, the local area around a team in a large city is the city borough and the local area around a team in a small city is the county.

Instrumental Variable Results

Table A8 presents the results of the IV-2SLS exercise. Model 1 reports the results when we regress our measure of team performance (goal differential) on linguistic distance using Ordinary Least Squares (OLS). Because every team in our sample plays the same number of games in a given season, we do not normalize the goal differential by the number of games played.

Model 2 provides the first stage regression results using cities' foreign population as the instrument. Controlling for the same team and new city-level covariates, we find that foreign population has a significant and sizable effect on linguistic distance. The coefficient on the foreign population is 56.91, indicating that a ten percentage points increase in the share of foreign citizens in a city would increase the linguistic distance the locale's team by 5.7 units (about 1/3 of one standard deviation). Interestingly, this is roughly the same change in linguistic distance that Manchester United Experienced between 2004 & 2005, when it added Edwin Van de Sar (Netherlands) and Park Ji-Sung (Korea) to its line-up (See Figure 1)

Diagnostics of the instrument demonstrate that the first stage is identified. The Angrist and Pishke multivariate F-test of excluded instruments indicates that we can reject the null hypothesis that the endogenous regressor (linguistic distance) is unidentified. Worries that weak identification may bias the results also can be set aside according to the Kleibergen-Paap test.

Models 3 presents the second-stage estimates for each of the instruments. Taking into account team selection effects, the core associations remain between linguistic distance and team

performance. This lends strong credence to the theory that diversity can enhance team performance.

One concern, however, is that the size of the coefficients in the IV-2SLS models are larger than in the OLS regressions (Model 1). For instance, the coefficient on linguistic distance increases from .255 to .329 and .478 in the two-stage setting. In addition, the R-squared of the model remains roughly the same, indicating that 40% of variation in goal differential is explained by the set of covariates. Both of these results are surprising, as instrumental variables regression should reduce reverse causality, which should lead the size of the effect and explained variance to decline.

Given the lack of weak instrument bias, the most likely explanation is that the model identifies a Local Average Treatment Effect (LATE) caused by the fact that the pull of ethnic enclaves only plays a significant role for a subset of teams and international players. Extremely wealthy and prestigious teams can afford to compensate outstanding players for the discomfort for locating far from ethnic compatriots. The benefits of foreign populations are most likely felt in the middle to low levels of national standings. Correspondingly, the benefits of attracting international talent are larger.

Table A8: Two-Stage Least Squares Regression 2004-2006

	OLS	First	Second
Model	*(1)*	*(2)*	*(3)*
Linguistic Distance	0.255***		0.478***
	(0.064)		(0.157)
Player Pay ($100,000 ln)	14.342***	9.453***	12.040***
	(2.408)	(2.325)	(2.636)
Average Wage City (ln)	10.259*	-26.953***	13.194**
	(5.259)	(6.758)	(5.545)
Population (ln)	4.221***	-1.407	4.936***
	(1.006)	(1.108)	(1.082)
Foreign Residents/City Population		54.906***	
		(12.982)	
Constant	-175.794***	318.9901***	-222.706***
	(57.99)	(69.576)	(65.458)
Year FE	Yes	Yes	Yes
Observations	170	170	170
R-squared	0.41	0.234	0.370
RMSE	14.903	16.21	15.13
Angrist-Pischke F-Test (Weak ID)		17.89***	
Kleibergen-Paap LM Statistic (Under ID)		32.73***	
Stock-Yogo 10% Maximal IV Size		16.38	

Robust standard errors, clustered at league level, in parentheses; *** p<0.01, ** p<0.05, * p<0.

N