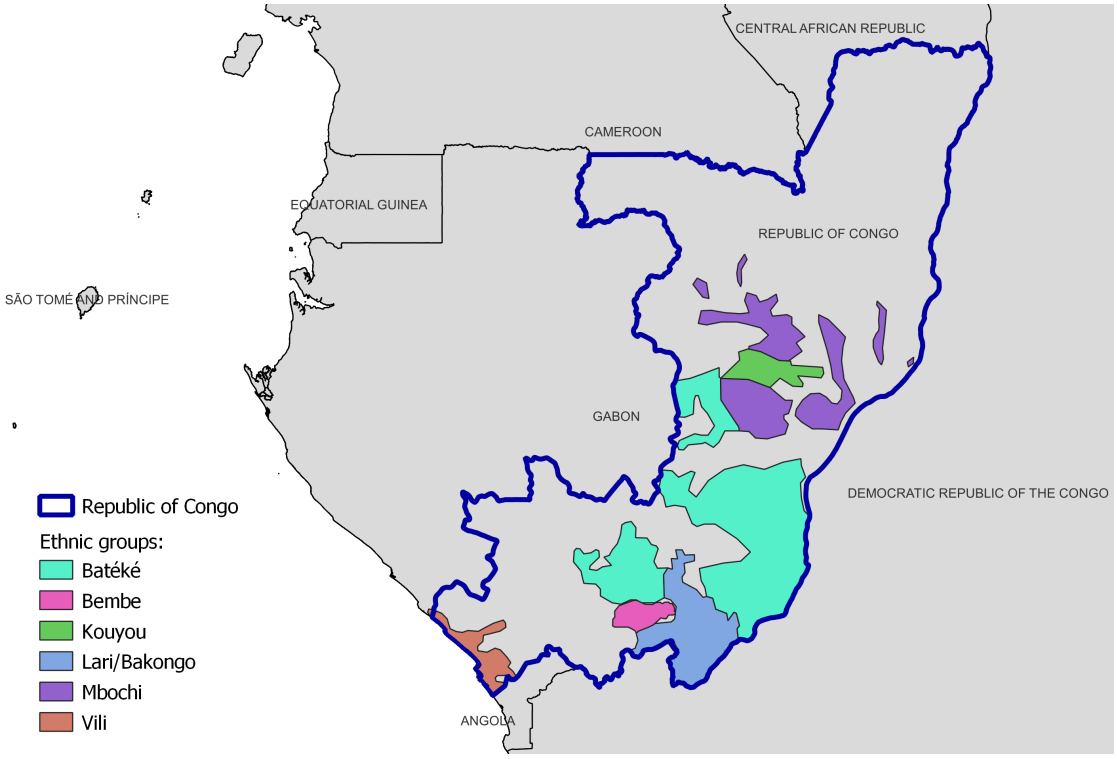
Appendix (Intended for Online Publication)

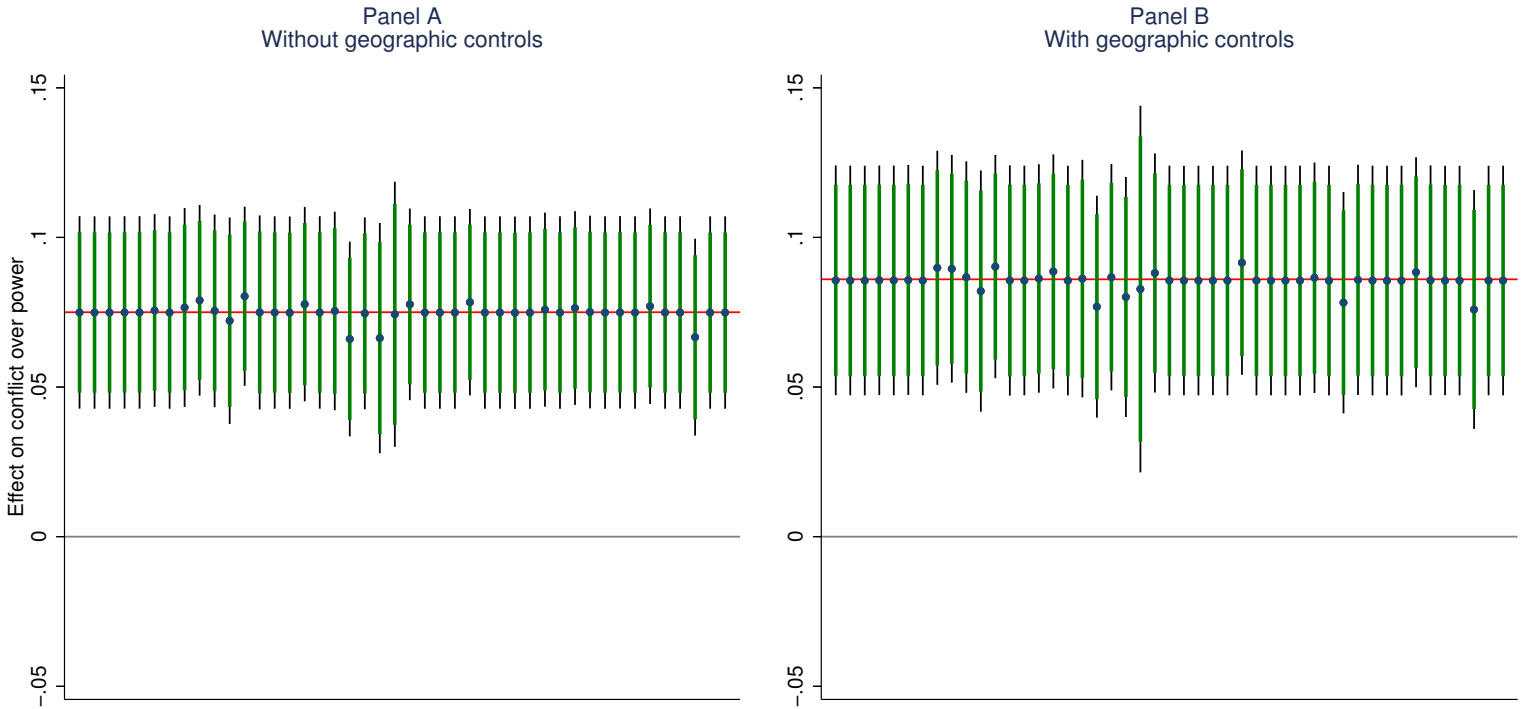
Appendix A: Additional Figures and Tables

FIGURE A-1: *Ethnic Groups and Ethnic Settlements in Congo*



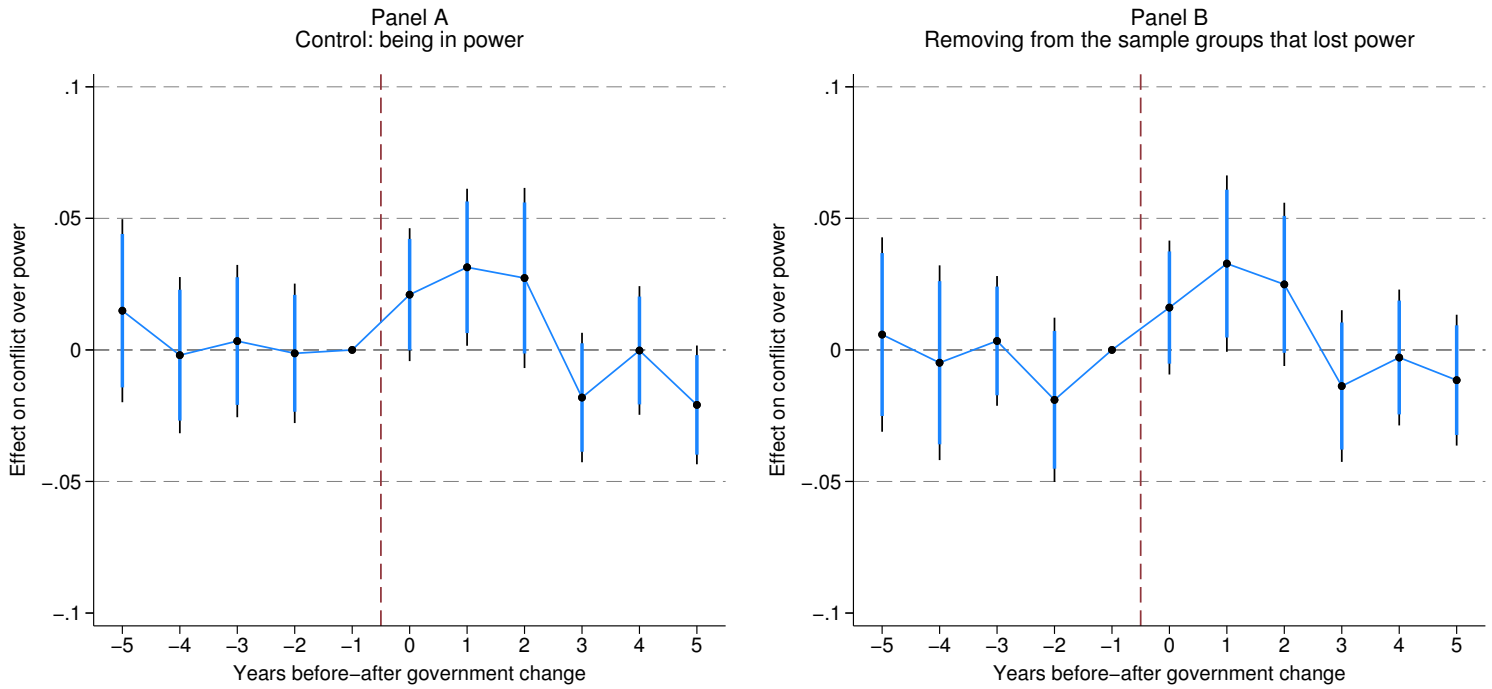
Notes: The figure illustrates the six politically relevant ethnic groups (Lari/Bakongo, Kouyou, Mbochi (proper). Bateke, Bemba, Vili) in Congo-Brazzaville and their settlements. Source: Ethnic Power Relations (EPR) Core (Wucherpfennig et al., 2012) and GEO-EPR (Wucherpfennig et al., 2011).

FIGURE A-2: *Checking for Outliers: Dropping one Country at a Time*



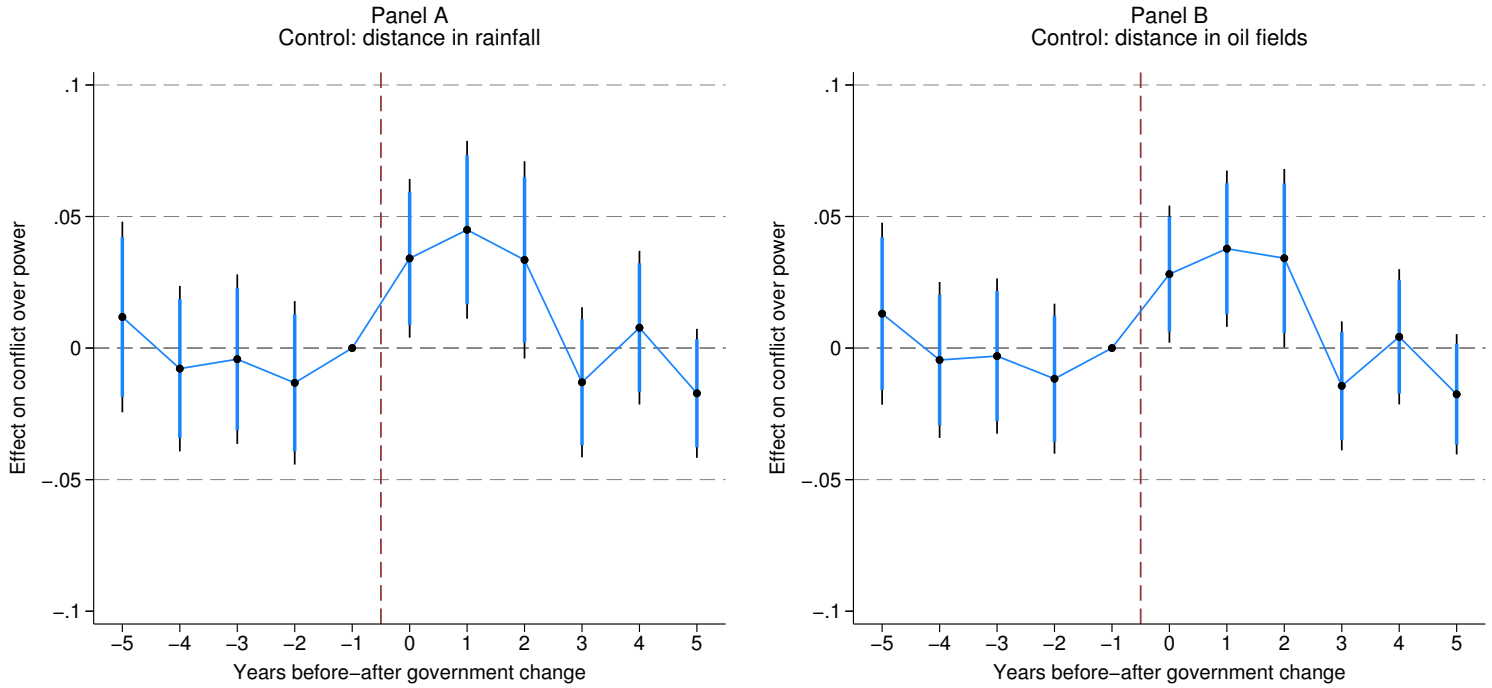
Notes: The figures show the stability of the estimates when dropping one country at a time from the sample. The red horizontal line indicates the baseline coefficient; green and black vertical lines indicate 90- and 95-percent confidence intervals, respectively. All specifications include country, year, country-year, and ethnicity fixed effects as well as ethnicity-specific time trends. Panel B includes geographic controls.

FIGURE A-3: *Effect of Cultural Distance Net of the Effect of Political Representation*



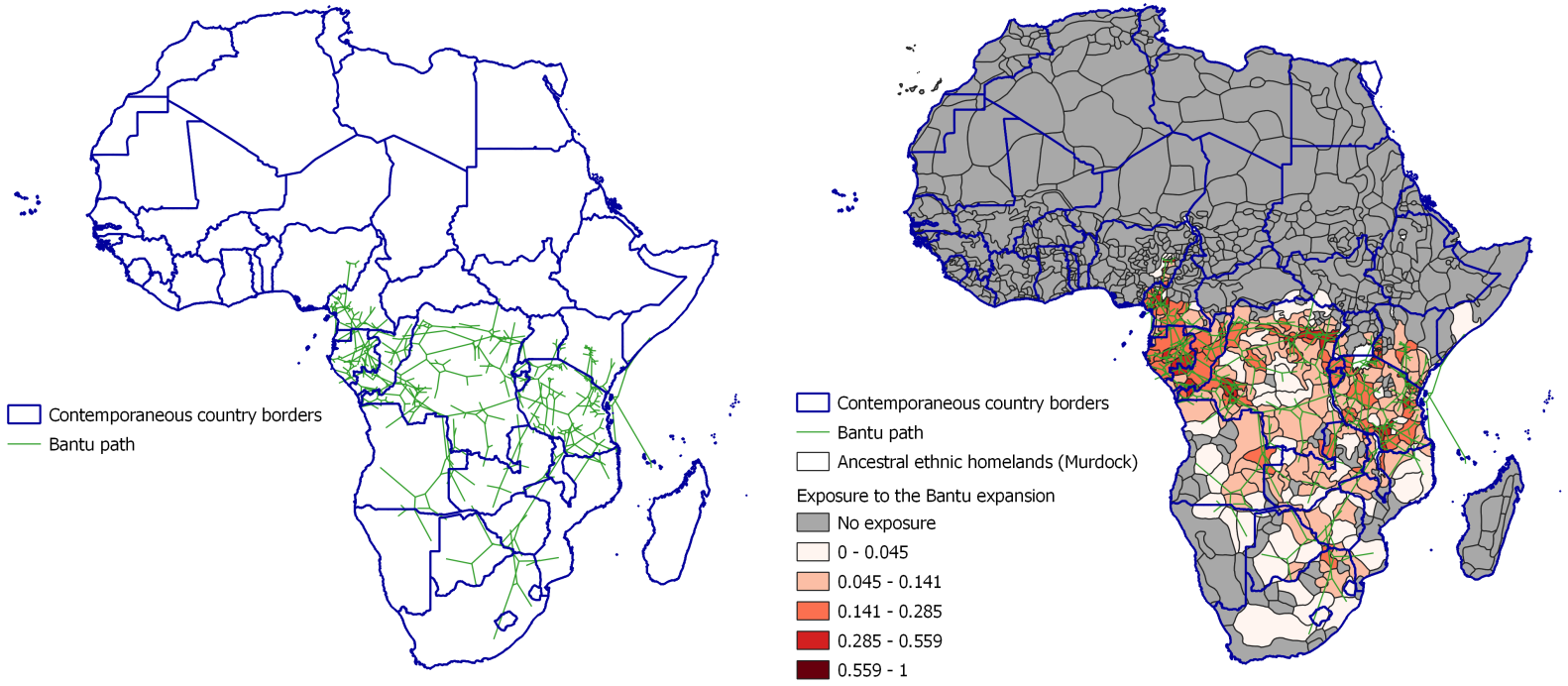
Notes: Event studies as the one in Figure 1: the graph on the left controls for whether or not an ethnic group is part of the government coalition, a variable that equals 1 if an ethnicity is represented in the government coalition in a certain year and 0 otherwise (sample size: 7,077); the graph on the right removes from each transition ethnic groups that have lost power as a result of that specific transition (sample size: 6,214). Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

FIGURE A-4: *Effect of Cultural Distance Net of the Effect of Income Distance*



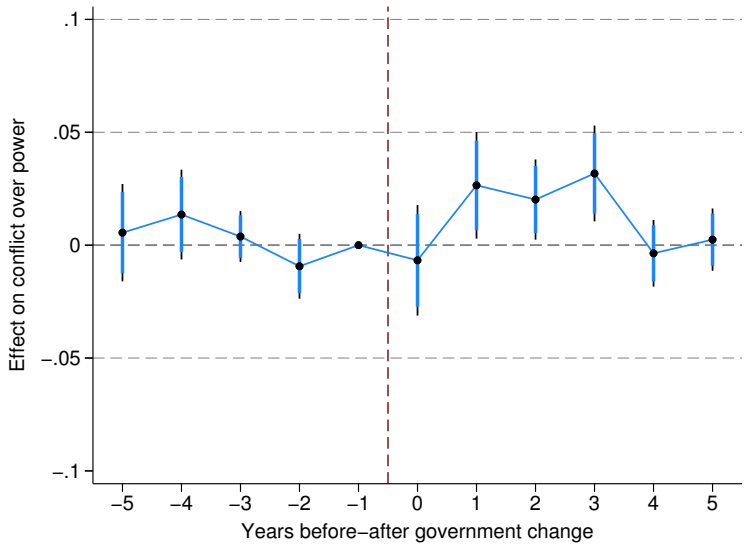
Notes: Event studies as the one in Figure 1 adding controls for income distances between potential rebels and governments as in Table 3. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

FIGURE A-5: *The Path of the Bantu Expansion and the Murdock Map*



Notes: The figure on the left shows the historical migration route of the Bantu tribe reconstructed by Grollemund et al. (2015b) and contemporaneous country borders (Ruggles et al., 2024). See Appendix B for additional details on the route reconstruction. The figure on the right overlays the historical migration route with the map of pre-colonial ethnic settlements constructed by Murdock. Different colors denote the intensity of exposure of each ethnic group to the Bantu expansion, computed by dividing the length of the path crossing an ethnic homeland by the size of the ethnic settlement, all normalized to range between 0 and 1.

FIGURE A-6: *Distance in Preferences over Public Goods and Conflict over Power*



Notes: Event study as the one described in equation 4, with the difference that MoreDist_{rce} is now an indicator that equals 1 if the distance in preferences over public goods between ethnic group r and the government of country c increases after the transition of interest e. Data on distance in preferences over public goods comes from the Afrobarometer survey. Section 7 describes this variable in detail. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

TABLE A-1: *Data Extract and Illustration of the Identifying Variation*

country	year	potential rebel	Government: ethnic group 1 [*status*]	ethnic group 2 [*status*]	ethnic group 3 [*status*]	ethnic group 4 [*status*]	ethnic group 5 [*status*]	linguistic distance	conflict over power
Congo	1995	Lari/Bakongo	Bembe [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]	Vili [*Junior partner*]	Lari/Bakongo [*Junior partner*]	0.136	0
Congo	1996	Lari/Bakongo	Bembe [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]	Vili [*Junior partner*]	Lari/Bakongo [*Junior partner*]	0.136	0
Congo	1997	Lari/Bakongo	Bembe [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]	Vili [*Junior partner*]	Lari/Bakongo [*Junior partner*]	0.136	0
Congo	1998	Lari/Bakongo	Mbochi (proper) [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]			0.202	1
Congo	1999	Lari/Bakongo	Mbochi (proper) [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]			0.202	1
Congo	2000	Lari/Bakongo	Mbochi (proper) [*Senior partner*]	Bateke [*Junior partner*]	Kouyou [*Junior partner*]			0.202	0

Notes: The table illustrates a data extract for the potential rebel Lari/Bakongo in Congo between 1995 and 2000. Within-ethnicity changes in cultural distance result from government transitions. In this case, the change in the government coalition occurred in 1998. The Lari/Bakongo group experienced an increase in cultural distance and started engaging in conflict against the government.

TABLE A-2: *Effect of Cultural Distance Net of the Effect of Political Representation: Including Dominant Groups and Monopolists as Potential Rebels*

	Dependent variable: Conflict over power					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	0.093***	0.061***	0.086***	0.066***	0.092***	0.088***
	(0.022)	(0.022)	(0.022)	(0.024)	(0.022)	(0.022)
In power		-0.107**				
		(0.041)				
Lost power			0.095**			
			(0.042)			
Status in government				-0.060***		
				(0.019)		
Lost senior power					0.023	
					(0.034)	
Downgraded						0.050**
						(0.023)
Observations	11,233	11,233	11,233	11,233	11,233	11,233
Adjusted R-squared	0.509	0.514	0.512	0.514	0.509	0.510

Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. *In power* is equal to 1 if an ethnicity is represented in the government coalition in a certain year and 0 otherwise. *Lost power* is equal to 1 if an ethnicity has lost government power in the previous 3 periods and 0 otherwise. *Status in government* equals 0 if an ethnicity does not hold any position in government, 1 if the ethnicity is a junior partner, 2 if the ethnicity is a senior partner, 3 if an ethnicity is dominant or monopolist. *Lost senior power* is equal to 1 if an ethnicity has lost senior government power in the previous 3 periods and 0 otherwise. *Downgraded* equals 1 if an ethnicity has been downgraded in the power scale over the past 3 periods. The sample includes 44 African countries, 57 years (1961-2017), and 284 distinct ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-3: *Cultural Distance and Ethnic Civil Conflict: Robustness Tests*

	Dependent variable: Ethnic conflict over power										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Linguistic distance	0.043**	0.037**	0.086***	0.086***	0.086**	0.086**	0.080***	0.092***	0.084***	0.058*	0.043**
	(0.021)	(0.017)	(0.018)	(0.021)	(0.039)	(0.036)	(0.024)	(0.020)	(0.019)	(0.031)	(0.019)
Mean of dep. var.	0.042	0.042	0.048	0.048	0.048	0.048	0.049	0.070	0.051	0.042	0.048
Robustness test	*Groups split across countries*		*Cluster by country*	*Cluster by country-year*	*Cluster by ethnicity*	*Cl. two-way ethn. & year*	*Climatic controls*	*At least 1 conflict*	*At least 1 gov. change*	*Balanced panel*	$\lambda = 0.05$
Distance type	w	uw	w	w	w	w	w	w	w	w	w
Country-year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Ethn.-country FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Ethn.-year FE	yes	yes									
Ethn. year trend			yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Climatic controls							yes				
Observations	7,434	7,434	9,990	9,990	9,990	9,990	9,196	6,792	8,370	8,866	9,990
Adjusted R-squared	0.836	0.836	0.507	0.507	0.507	0.507	0.517	0.503	0.486	0.539	0.504

Notes: For a description of all explanatory variables, refer to notes in Table 1. In columns (1) and (2), the unit of observation is a Murdock ethnic group in a country and year. The sample includes 76 Murdock groups partitioned across 33 countries during the Scramble for Africa (166 distinct Murdock groups by country). Because the definition of ethnicity in the Murdock map is finer than in the EPR database, a given EPR group can be matched to multiple Murdock groups. This keeps the resulting estimation sample still large (N=7,434), even though almost 60 percent of the original EPR groups included in the baseline analysis are dropped due to not having been split across countries. See Appendix Section B-5 for additional details on the data construction and the sample adopted to conduct this analysis. In columns (1) and (2), standard errors are clustered at the ethnic group by country level. In the remaining columns, two-way clustered standard errors by year and country are reported in parenthesis, unless otherwise indicated in columns 3, 4, 5, and 6. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-4: *Cultural Distance and Ethnic Civil Conflict: Lagged Conflict and Anticipatory Effects*

	Ethnic conflict over power				
	(1)	(2)	(3)	(4)	(5)
Linguistic distance$_{t-1}$ (lag)				0.045**	
				(0.020)	
Linguistic distance$_t$	0.086***	0.046***	0.036**		0.074**
	(0.019)	(0.013)	(0.017)		(0.029)
Linguistic distance$_{t+1}$ (lead)					0.011
					(0.032)
Conflict over power$_{t-1}$		0.572***	0.649***		
		(0.082)	(0.123)		
Mean of dep. var.	0.048	0.048	0.048	0.048	0.048
Country-year fixed effects	yes	yes	yes	yes	yes
Ethnicity fixed effects	yes	yes	yes	yes	yes
Ethnicity year trend	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes
Observations	9,990	9,676	9,384	9,676	9,676
Adjusted R-squared	0.507	0.676	0.319	0.515	0.515

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power and 0 otherwise. *Linguistic distance*$_t$ captures the weighted linguistic distance between a potential rebel and the ethnic groups in power. Conflict$_{t-1}$ is a binary variable that is equal to 1 if an ethnic group was involved in a conflict over power in the previous year. Since one caveat with the specification in column 2 is that the inclusion of a lagged dependent variable in a fixed effects model might generate Nickell bias (Nickell, 1981), in column 3 I instrument the first lag of the dependent variable (Conflict$_{t-1}$) with the second lag (Conflict$_{t-2}$). Linguistic distance$_{t-1}$ and Linguistic distance$_{t+1}$ denote a lagged and lead measure of linguistic distance, respectively. For a description of all explanatory variables, refer to notes in Table 1. The sample includes 44 African countries, 57 years (1961-2017), and 265 ethnic groups. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-5: *IV: Inference using Lee et al. (2022) tF Procedure*

	Conflict over government power	
	(1)	(2)
Linguistic distance	0.203***	0.151***
	(0.036)	(0.029)
tF 0.05 se	[0.056]	[0.036]
Mean of dep. var	0.047	0.047
Distance type	w	uw
Country-year FE	yes	yes
Ethnicity FE & trends	yes	yes
Controls	yes	yes
Observations	4,027	4,027

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting over power. *Linguistic distance* captures linguistic distance between each potential rebel and the ethnic groups at the government. For a description of geographic controls, refer to the notes in Table 1. The sample includes 18 African countries in the Bantu region and 57 years (1961-2017). Two-way clustered standard errors by year and country, adjusted for the low number of clusters using the number of countries, are reported in parenthesis and standard errors further re-adjusted using Lee et al. (2022) *tF* procedure are reported in square brackets. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-6: *IV: Alternative Clustering*

	Linguistic distance	Conflict over government power	
	First stage	*Reduced form*	*2SLS*
	(1)	(2)	(3)
Bantu distance	0.332***		
	(0.023)		
Linguistic distance		0.067**	0.203**
		(0.028)	(0.085)
Kleibergen-Paap F-statistic			156.3
Mean of dep. var	0	0.047	0.047
Country-year FE	yes	yes	yes
Ethnicity FE & trends	yes	yes	yes
Controls	yes	yes	yes
Observations	4,027	4,027	4,027

Notes: The unit of observation is an ethnic group-country-year. *Linguistic distance* and *Bantu distance* are standardized. The sample includes 18 African countries in the Bantu region and 57 years (1961-2017). Standard errors are clustered by country and year. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-7: *Discontent with Government Performance Index: Principal Component Analysis Loadings*

How well or badly would you say the current government is handling the following matters:	Loadings						
	Round 1 1999-2001	Round 2 2002-2003	Round 3 2005-2006	Round 4 2008-2009	Round 5 2011-2013	Round 6 2014-2015	Round 7 2016-2017
Managing the economy		0.297	0.327	0.260	0.273	0.291	0.223
Improving living standards of the poor				0.278	0.277	0.302	0.234
Creating jobs	0.581	0.281	0.310	0.266	0.260	0.294	0.235
Keeping prices stable	0.587	0.282	0.301	0.231	0.225	0.276	0.217
Narrowing the income gap	0.563	0.278	0.306	0.260	0.254	0.284	0.227
Reducing crime		0.288	0.292	0.242	0.244	0.257	0.221
Improving basic health services		0.291	0.310	0.261	0.266	0.278	0.245
Addressing educational needs		0.290	0.303	0.243	0.255	0.271	0.239
Improving water and sanitation services		0.221	0.276	0.249	0.245	0.270	0.223
Ensuring enough to eat		0.268	0.317	0.268	0.262	0.283	0.231
Fighting corruption		0.295	0.312	0.254	0.261	0.279	0.226
Reducing conflict		0.290			0.251		
Combating HIV		0.253	0.249	0.198	0.208		
Combating malaria		0.264					
Maintaining roads and bridges				0.247	0.243	0.254	0.204
Providing reliable electric supply				0.251	0.237	0.260	0.205
Protecting rivers and forests				0.254			
Promoting equal rights and opportunities for women				0.228	0.231		0.227
Preventing election violence							0.230
Preventing or resolving violent community conflict							0.242
Countering violence from armed extremists							0.231
Addressing needs of youth							0.246
Protecting rights, promoting opportunities for disabled							0.238
Common variance explained by first principal component	62%	37%	39%	37%	36%	44%	38%
N (respondents)	12,666	15,281	19,661	19,462	36,904	46,134	13,361

Notes: The table illustrates the loadings of each variable from the Afrobarometer survey used for the principal component analysis. Each variable is binary and is equal to 1 if the respondent thinks the government is performing badly or fairly badly in handling each policy goal. Sample sizes denote the set of respondent answering each question for which information on ethnicity is available.

TABLE A-8: *Cultural Distance and Disagreement by Type of Public Good*

	Panel A:							
	jobs	prices	income gap	economy	crime	health	education	water
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Linguistic distance	0.153***	0.069**	0.112***	0.153***	0.040	0.104**	0.075***	0.089***
	(0.025)	(0.035)	(0.033)	(0.046)	(0.038)	(0.044)	(0.027)	(0.028)
Mean dep. var.	0.710	0.733	0.756	0.577	0.509	0.425	0.405	0.549
Observations	117,062	117,062	117,062	105,295	105,295	105,295	105,295	105,295
Adjusted R-squared	0.094	0.122	0.076	0.153	0.107	0.096	0.115	0.083
	Panel B:							
	eat	corruption	HIV	poor	roads	electricity	women	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Linguistic distance	0.096***	0.070*	-0.049	0.079*	0.097*	0.177***	0.121	
	(0.029)	(0.038)	(0.034)	(0.041)	(0.054)	(0.052)	(0.078)	
Mean dep. var.	0.673	0.633	0.252	0.690	0.512	0.591	0.341	
Observations	105,295	105,295	59,521	89,352	89,352	89,352	54,243	
Adjusted R-squared	0.107	0.118	0.105	0.098	0.104	0.143	0.102	

Notes: The dependent variable varies by column and captures the extent to which the person thinks the government is performing very badly or fairly badly in handling a specific policy matter included in at least three rounds of the survey. *Linguistic distance* captures the weighted linguistic distance between a respondent and the ethnic groups in power and is standardized. Individual controls include age, age squared, a dummy for female and a dummy for residence in an urban area. Standard errors clustered at the ethnicity-country level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-9: *Preferences over Policy Index: Principal Component Analysis Loadings*

What are the most important problems facing this country that government should address?	Loadings						
	Round 1 1999-2001	Round 2 2002-2003	Round 3 2005-2006	Round 4 2008-2009	Round 5 2011-2013	Round 6 2014-2015	Round 7 2016-2017
Health	0.138	0.101	0.239	0.350	0.244	0.367	0.326
Infrastructure	0.195	0.208	0.240	0.149	0.344	0.186	0.229
Agriculture/Food	0.445	-0.132	0.137	-0.026	-0.222	-0.127	-0.042
Governance	-0.104	0.621	0.427	0.387	0.186	0.337	0.243
The economy	-0.695	-0.722	-0.771	-0.737	-0.667	-0.663	-0.711
Services	0.413	0.121	0.243	0.387	0.535	0.501	0.522
Other	0.128	-0.060	-0.106	0.079	-0.068	-0.061	-0.005
No pressing problems	0.253	0.062	0.140	0.070	0.051	0.088	0.047
Common variance explained by first principal component	21%	21%	20%	21%	21%	23%	21%
N (ethnic groups)	144	232	292	373	585	525	557

Notes: The table illustrates the loadings of each variable from the Afrobarometer survey employed for the principal component analysis to construct an index capturing preferences over policy at the ethnic group level. Each variable captures the fraction of respondents within each ethnic group stating that a given domain should be the one that the government should address. The options that respondents can provide are more fine-grained than the above-listed domains. For example, under the "health" domain, respondents can list health, AIDS, or sickness/disease. I follow the grouping in the Afrobarometer questionnaires (e.g., see https://www.afrobarometer.org/wp-content/uploads/2022/02/alg_r6_questionnaire.pdf) to aggregate up the various options into fewer domains.

TABLE A-10: *Cultural Distance and Distance in Preferences by Type of Public Good*

	Distance in the percentage of respondents who think priority should go to:					
	health	infrastructure	governance	food	services	economy
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	0.000	-0.003***	0.005	0.001	0.002	0.018***
	(0.001)	(0.001)	(0.004)	(0.002)	(0.003)	(0.006)
Mean of dep. var.	0.090	0.092	0.110	0.124	0.149	0.221
Observations	6,308	6,308	6,308	6,308	6,308	6,308
Adjusted R-squared	0.792	0.783	0.775	0.679	0.617	0.605

Notes: The unit of observation is a pair of ethnic groups in a country and year. The sample includes 401 distinct ethnic groups as defined in the Afrobarometer survey. The dependent variable in each column is the absolute difference between ethnicity i and ethnicity j in the ratio of respondents thinking that priority should go to health, infrastructure, governance, food, services or the economy. These variables range between 0 and 1. *Linguistic distance* is standardized. Standard errors clustered two-way at the ethnicity i-country and ethnicity j-country level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE A-11: *Preferences over Policy Index (time-invariant): Principal Component Analysis Loadings*

What are the most important problems facing this country that government should address?	Loadings
Health	0.2345
Infrastructure	0.2876
Governance	-0.1297
Agriculture/Food	0.3347
Economy	-0.7134
Services	0.4606
Other	-0.0113
No pressing problems	0.1107

Notes: The table illustrates the loadings of each variable from the Afrobarometer survey used for the principal component analysis used to construct a time-invariant index capturing preferences over policy at the ethnic group level. Each variable captures the fraction of respondents within each ethnic group stating that a given domain should be the one that the government should address. The options that respondents can provide are more fine-grained than the above-listed domains. For example, under the "health" domain, respondents can list health, AIDS, or sickness/disease. I follow the grouping in the Afrobarometer questionnaires (e.g., see https://www.afrobarometer.org/wp-content/uploads/2022/02/alg_r6_questionnaire.pdf) to aggregate up the various options into fewer domains.

Appendix B: Additional Analyses

B-1 Definition of Politically Relevant Groups and Robustness to the Inclusion of Politically Irrelevant Groups

In this section, I provide a more detailed description of how EPR defines ethnic groups, with a particular focus on EPR's inclusion criteria based on a group's political relevance. I then discuss why focusing on politically-relevant groups is an appropriate choice for the purpose of this paper. Finally, I replicate my findings employing a broader sample of ethnicities that includes politically irrelevant groups.

Selection of ethnicities and definitions in the EPR dataset

The unit of analysis in my empirical specification is an ethnic group observed in a country and year. As such, an important prerequisite for the analysis is to define a list of ethnic groups to be included. There is no clear consensus in the literature on what constitutes an ethnic group (Fearon, 2003). Defining ethnicity is a challenge, as the notion of ethnicity may vary across contexts and over time. To define my sample of potential rebels, I rely on the definition of ethnicity provided by the EPR database. In the EPR database, ethnicity is defined as any "subjectively experienced sense of commonality based on a belief in common ancestry and shared culture" (Wucherpfennig et al., 2011, p. 427) and, as a result, EPR ethnic groups are characterized by a shared language, religion, or phenotypical feature. The EPR database includes "politically relevant" ethnic groups, defined as follows:

> "*An ethnic group is politically relevant if either at least one significant political actor claims to represent the interests of that group in the national political arena or if group members are systematically and intentionally discriminated against in the domain of public politics.*" (Wucherpfennig et al., 2011, p. 427).

As I elaborate in the next subsection, politically relevant groups are a natural

set of groups to consider when studying the relationship between cultural distance and conflict over government power, due to these groups' active participation in the national political arena. Three things are worth noticing about EPR's definition of ethnicity and its inclusion criterion. First, to be politically relevant, a group only needs to be represented in the national arena by at least one actor. This condition does not require an ethnic group to be part of the government coalition: for instance, being represented in the legislature, or even being represented by a party running in national elections without winning a seat in parliament would be enough for a group to be classified as politically relevant. Moreover, Wucherpfennig et al. (2011) emphasize that a "significant" political actor may refer to any political organization, and not necessarily a party, which makes this inclusion criterion even less restrictive.

Second, groups that are excluded from executive government power are not necessarily in conflict with or discriminated against by the central government. In fact, groups that EPR defines as being "discriminated" account for only 13.65% of the observations in my estimation sample, as shown in Table B-1 below; the most common category among groups that are not in the government coalition is the "powerless" category, which accounts for 32.18% of the observations. Powerless groups are not openly discriminated by the government, as indicated by the EPR dataset's definition: "*Powerless: Elite representatives hold no political power (or do not have influence on decision making) at the national level of executive power - although without being explicitly discriminated against.*" (EPR codebook, pages 5-6). Moreover, powerless groups are not necessarily involved in conflict: 30 percent of my estimation sample is accounted for by powerless ethnic groups not involved in conflict against the government.

Third, while conflict is common among discriminated and powerless groups (as expected), Table B-1 shows that conflict can be initiated also by rebels that share the same ethnic identity of members of the executive.

TABLE B-1: *Distribution of Potential Rebels and Conflict Prevalence by Status*

Status	N	Freq.	Conflict
Discriminated	1,364	13.65%	0.122
Powerless	3,215	32.18%	0.067
Self-exclusion	21	0.21%	0
Junior Partner	3,778	37.82%	0.014
Senior Partner	1,612	16.14%	0.024
Total	9,990	100.00%	

Notes: This table reports the distribution of observations in the main estimation sample by status. The last column reports the prevalence of conflict in each category.

Rationale behind the focus on politically relevant groups in the analysis

At its core, the analysis in this paper aims at uncovering when and why an ethnic group decides to rebel against the central government. I call ethnic groups that face the decision of whether or not to rebel at a given point in time *potential rebels*. Consider a potential rebel's decision in a cost-benefit framework. On the one hand, fighting is costly, as it requires a group to mobilize citizens, gather weapons, and potentially undergo losses of civilians and soldiers. On the other hand, conflict may come with benefits conditional on a group winning the war and gaining government power, as the group will be in the position to provide the type of public good that is most aligned with its members' preferences. The main argument of the paper is that, the larger an ethnic group's cultural distance to the government, the greater the potential benefit of war and the more likely the occurrence of conflict.

Crucially, some groups will *never* find fighting against the government profitable and I argue that this makes them unfit to enter my "potential rebels" sample. Non-politically relevant groups are a case in point: given that they do not make any political claims in the national arena, it is unlikely that these groups would perceive any potential benefit from entering conflict against the government.[1] For

[1]Of course, these groups may engage in conflict with their neighbour, or experience intra-group conflict, but these types of disputes are not the object of the current paper.

instance, these groups may not be interested in the policies and public goods provided at the national level because they rely on their own local (or ethnic) institutions. Since the potential benefit of engaging in conflict over national government power is arguably low (or non-existent) for these groups (irrespective of who is in power), they should not be considered to be "potential rebels" in the first place.[2]

For all these reasons, I argue that politically relevant ethnic groups are the "target population" of interest for the study of the relationship between cultural distance to the government and the decision to engage in civil conflict.

Replication using alternative datasets

Keeping in mind the arguments proposed above in favour of focusing on politically relevant groups in the main analysis, I nevertheless test whether my results hold when considering a broader sample of potential rebels, i.e., including those groups that we should not expect to react to cultural distance and rebel. In this subsection, I find that the main results hold when broadening the sample of ethnic groups and including politically irrelevant groups as potential rebels.

I build a dataset that includes as potential rebels all the ethnic groups listed in the All Minority at Risk (AMAR) dataset (Birnir et al., 2018). AMAR provides a sample frame that includes all contemporary *socially relevant* ethnic groups. This sample is thus not defined by any political criterion such as being at risk or being politically relevant.[3] Based on Fearon's (2008) definition of "socially relevant," AMAR focuses on groups where members "notice and condition their actions on ethnic distinction in everyday life." The social relevance of such identity depends on the salience of ethnic identities in guiding people's actions both in public and

[2]By the EPR definition of politically relevant, groups that are politically irrelevant are indeed never involved in conflict against the government.

[3]AMAR is not to be confused with the original Minorities at Risk (MAR) dataset, which was created to identify culturally-defined groups "at risk" of human rights violations and/or of protracted conflict behavior. The AMAR project was introduced to address the selection problems associated with MAR's restricted sample of groups.

in private lives.[4]

AMAR includes a global sample of ethnic groups, in turn divided into subgroups. I consider both groups and subgroups as potential rebels. I construct a panel that follows each African group between 1961 and 2017 and match potential rebels to their spoken languages using the LEDA package.[5] I construct linguistic distance to the government using information on ethnic groups forming the government from the EPR dataset, as in my baseline dataset.[6] To establish whether an AMAR potential rebel is rebelling or not, I proceed as follows. First, I match EPR ethnic groups to AMAR ethnic groups using the LEDA package;[7] this allows me to merge AMAR ethnic groups to ACD2EPR, which contains information on the ethnic identity of rebel forces and to UCDP, which includes information on rebels' conflict involvement; finally, I assign a value of 1 to AMAR groups if they are part of a rebel organization fighting against the government in a given country and year, and zero otherwise. An important downside of AMAR is that it does not include information on the location or settlements of the ethnic groups. To assign geographic characteristics (latitude and longitude, CSI, elevation, and ruggedness), I need to leverage the Ethnologue's polygons assigned to languages spoken by each AMAR group.[8] The data sources I adopt to construct geographic controls are the same I used for the dataset in the baseline analysis.

[4]Based on this concept of "social relevance," the criteria for inclusion in AMAR are the following (see AMAR website): "membership in the group is determined primarily by descent by both members and non-members; the group may be a caste if membership is determined by descent and precludes public social mobility; membership in the group is recognized and viewed as important by members and/or non-members; the importance may be psychological, normative, and/or strategic; members share some distinguishing cultural features, such as common language, religion, occupational niche, and customs; one or more of these cultural features are either practiced by a majority of the group or preserved and studied by a set of members who are broadly respected by the wider membership for so doing; the group has at least 100,000 members or constitutes one percent of a country's population."

[5]To achieve this, I use the LEDA package's links to the Ethnologue database.

[6]Because an AMAR group (or subgroup) can be assigned to multiple languages, the resulting linguistic distance is an average of each language's distance to the groups forming the government.

[7]Each EPR group can be matched to multiple AMAR groups/subgroups, because AMAR's definition of ethnicity is more fine-grained compared to EPR's.

[8]To date, the AMAR website provides only a list of ethnic groups, subgroups, and country, without any additional covariates. A richer dataset (called AMAR Phase I data) is not yet available for download.

I re-run the baseline specification using this alternative dataset where potential rebels are defined based on AMAR's definition of ethnicity. The estimation sample includes 1,466 ethnic groups in 42 African countries observed over 57 years.[9] Table B-2 below replicates Table 1 in the paper and shows that the average effects are positive and significant when considering a wider set of ethnic groups as potential rebels. In my preferred specification in column (3), a one standard deviation increase in linguistic distance increases conflict over power by 0.14 standard deviations.[10] As should be expected, given that the sample now includes a higher number of ethnic groups that do not react to cultural distance, the effect size is smaller than in the main specification, but still large and significant.

Because it includes politically irrelevant groups, this dataset also provides me with the opportunity to explore whether the overall effect is driven by groups that are affected by national institutions the most, and that would perceive a potential benefit from engaging in conflict, as conjectured in the previous subsection. Since non-politically relevant groups are never found to be involved in conflict, the results in Table B-2 are (mechanically) entirely driven by politically relevant groups. However, I can test for heterogeneous effects along an alternative dimension: African historiography has emphasized that national institutions have limited reach in areas far from the capital city (see Michalopoulos and Papaioannou's (2014) review of the work by Herbst (2000), Mamdani (1996), and Rothchild and Olorunsola (2019)). Michalopoulos and Papaioannou (2014) find that colonial and contemporary national institutions are significantly less influential in areas far from the capital cities: national institutions do not shape regional development in

[9]The number of countries included in the estimation sample is lower compared to the baseline specification because AMAR does not distinguish between Sudan and South Sudan from 2011 onwards, as the ethnic group list AMAR provides is time invariant. Comoros is absent from the AMAR estimation sample because AMAR only lists one ethnic group for this country, which gets dropped as a singleton.

[10]The average prevalence of conflict in the sample is lower than in the main dataset, but perhaps not as low as one would expect given the considerably higher number of ethnic groups included as potential rebels. This is because AMAR provides a more fine-grained definition of ethnicity than EPR, resulting in multiple AMAR groups assigned to a single EPR group by the LEDA package and thus being categorized as involved in conflict when matched to ACD2EPR, as described above.

TABLE B-2: *Cultural Distance and Ethnic Civil Conflict over Power (AMAR)*

	Dependent variable: Ethnic conflict over power					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance^{W}	0.026**	0.029**	0.027**			
	(0.013)	(0.011)	(0.012)			
Linguistic distance^{UW}				0.023*	0.025**	0.025**
				(0.013)	(0.011)	(0.011)
Mean of dep. var.	0.039	0.039	0.039	0.039	0.039	0.039
Country-year FE	yes	yes	yes	yes	yes	yes
Ethnicity-country FE	yes	yes	yes	yes	yes	yes
Ethnicity-country trend		yes	yes		yes	yes
Geographic controls			yes			yes
Observations	77,601	77,601	77,601	77,601	77,601	77,601
Adjusted R-squared	0.525	0.603	0.604	0.525	0.603	0.604

Notes: The unit of observation is an AMAR ethnic group-country-year. For a description of geographic controls, see Table 1. Two-way clustered standard errors by year and country are reported in parenthesis.*** (**) (*) indicate significance at the 1% (5%) (10%) level.

the hinterland and the reach of the state is limited in more remote areas.

In light of this work, we should expect two empirical patterns to emerge: (i) proximity to the capital should be correlated with an ethnic group's political relevance; (ii) proximity to the capital should be a source of heterogeneity in the extent to which groups react to cultural distance. If a group is unaffected by national institutions and public policies, then cultural distance to the government should not affect a group's decision to rebel.

As illustrated by Table B-3 below, the further away a group is from the capital in the AMAR sample, the less likely it is to be included in the EPR database as being politically relevant.

TABLE B-3: *Distance to the Capital and Political Relevance*

	Politically relevant
ln(1 + distance to the capital)	-0.087***
	(0.028)
Mean of dep. var.	0.568
Country fixed effects	yes
Observations	1,466
Adjusted R-squared	0.186

Notes: The unit of observation is an AMAR ethnic group in a country. The dependent variable equals 1 if an AMAR group is matched to any EPR group by the LEDA package, and 0 otherwise. Standard errors clustered at the country level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

To assess heterogeneity in the violent response to cultural distance based on distance to the capital, I estimate the following:

$$\text{Conflict}_{rct} = \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \beta \text{LD}_{rct} + \pi \text{Below median dist capital}_{rc} \times \text{LD}_{rct} + \gamma \text{G}_{rct} + \epsilon_{rct} \quad (1)$$

where "Below median dist to capital" equals 1 for groups whose distance to the capital is smaller than the sample median, and 0 otherwise.[11] Columns 3 and 4 in Table B-4 report the results. As can be inferred from the positive and significant estimate of π and the small and insignificant estimate of β, the effect of cultural distance on conflict is entirely driven by groups located closer to the capital city.

Taken together, these findings show that considering all ethnic groups as potential rebels, irrespective of their political relevance, does not alter the main message of the paper: the average effect of cultural distance on conflict remains positive and significant, albeit smaller in magnitude when compared to my baseline specification. The heterogeneous effects highlighted in Table B-4, however, show that the

[11]The median distance to the capital for groups in the AMAR sample is 373.6 km. This is comparable to the median distance among geographic cells in Michalopolous and Papaioannou (2013) (367.4 km).

effect is driven by groups that are more likely to benefit from national institutions, i.e., by those located closer to the national capital. These results are consistent with a cost-benefit framework in which the response to cultural distance through conflict is a costly action that is undertaken only by groups with positive expected benefits.

TABLE B-4: *Cultural Distance and Conflict over Power (AMAR): Heterogeneity*

	Conflict over power	
	(1)	(2)
Linguistic distance	-0.002	0.000
	(0.011)	(0.008)
Linguistic distance × Below median dist. to capital	0.040**	0.036**
	(0.017)	(0.013)
Mean of dep. var.	0.039	0.039
Distance type	w	uw
Observations	77,601	77,601
Adjusted R-squared	0.605	0.605

Notes: The unit of observation is an ethnic group-country-year. Both columns include country-year fixed effects, ethnicity by country fixed effects, ethnicity-by-country time trends, and geographic controls. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

Finally, Table B-5 below reports that positive and significant average effects of cultural distance on conflict are present when using the Geo-referencing of Ethnic Groups (GREG) dataset, the Ethnographic Atlas, or the Ethnologue to define potential rebels.

It should be noted that these data sources come with important limitations for the purpose of my analysis. GREG and the Ethnographic Atlas are characterized by an outdated definition of ethnicity: the Atlas codes ethnic groups prior to European contact, while GREG is based on the Atlas Narodov Mira, which categorized ethnic groups and their settlements in the early 1960s. Moreover, GREG only includes a subsample of groups for which information on settlements is available. Contrary to AMAR and EPR, therefore, these two datasets may not reflect the current configuration of ethnic groups in contemporary countries (Weidmann et

al., 2010), which makes them less suitable for my analysis.

The Ethnologue, conversely, lists *all* languages in a country and, as such, it provides the most "complete sample" of linguistic groups that can be compiled. However, contrary to the other datasets on ethnicity, the Ethnologue does not provide information on whether or not individuals speaking a given language share an ethnic identity, which makes the Ethnologue ill-suited for an analysis of ethnic conflict at the ethnicity level. For instance, the Krio language in Sierra Leone is spoken by everyone in the country.[12] Krio constitutes a standalone linguistic group in the Ethnologue, but its speakers do not identify under a shared ethnic identity: the Mende, the Temne, the Kissi people, just to mention a few, constitute distinct ethnic groups, even though their members are fluent in Krio according to the Ethnologue website. In other cases, members identifying with the same ethnic group may speak multiple languages. An example is the ethnic group called "Ndowe" in Equatorial Guinea, whose members speak the Kwasio, Kombe, and Batanga languages. The Ethnologue considers each of these languages as a separate entity, even though their speakers identify with the same ethnicity.

Despite the limitations of these alternative data sources, the fact that the main results hold across various alternative definitions of ethnicity is reassuring.

[12]See the description of Krio on the Ethnologue website: https://www.ethnologue.com/language/kri/

TABLE B-5: *Cultural Distance and Ethnic Civil Conflict over Power: GREG, Ethnographic Atlas, and Ethnologue*

	Dependent variable: Ethnic conflict over power											
	GREG				Ethnographic Atlas				Ethnologue			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Linguistic distance^{W}	0.024***	0.019***			0.021***	0.018***			0.008***	0.008***		
	(0.007)	(0.007)			(0.005)	(0.006)			(0.002)	(0.002)		
Linguistic distance^{UW}			0.021***	0.016**			0.019***	0.016***			0.008***	0.008***
			(0.007)	(0.007)			(0.005)	(0.005)			(0.002)	(0.002)
Mean of dep. var.	0.039	0.039	0.039	0.039	0.036	0.036	0.036	0.036	0.007	0.007	0.007	0.007
Country-year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Ethnicity-country FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Ethnicity-country trend	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Geographic controls		yes		yes		yes		yes		yes		yes
Observations	23,781	23,781	23,781	23,781	52,153	52,153	52,153	52,153	130,913	130,913	130,913	130,913
Adjusted R-squared	0.577	0.578	0.577	0.578	0.633	0.633	0.633	0.633	0.456	0.457	0.456	0.457

Notes: The unit of observation is an ethnic group-country-year. The sample in columns 1-4 includes 438 ethnic groups by country. The sample in columns 5-8 includes 992 ethnic groups by country. The sample in columns 9-12 includes 2440 linguistic groups by country. Linguistic distance is standardized. Magnitudes: in column 2, a one standard deviation increase in linguistic distance increases conflict over power by 0.09 standard deviations; in column 6, a one standard deviation increase in linguistic distance increases conflict over power by 0.10 standard deviations; in column 10, a one standard deviation increase in linguistic distance increases conflict over power by 0.10 standard deviations. Standard errors clustered by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-2 Additional Analyses Supporting the Validity of the Estimation Strategy

The source of variation I leverage in my analysis comes from changes in the ethnic identity of the government. One concern with my strategy could be that there might be third factors generating both government transitions and, at the same time, conflict among those groups that become more culturally distant after the transition. To gauge the idea behind my identification strategy and the required identifying assumption, consider my event study specification, where the event of interest is a government transition. This is an event that occurs at the country-year level and that generates two sub-groups of ethnicities: those that become more culturally distant from the government and those that do not (i.e., become less distant or whose distance remains unchanged). Because my strategy compares the reaction to a government change among groups that become more culturally distant to the reaction among those that do not, it does not require government changes to occur at random: the required assumption is that, in the absence of a government transition, these two subgroups of ethnicities would have exhibited the same trend in conflict. While this assumption is not directly testable, my event-study specification provides suggestive evidence of its plausibility. Equivalently, in the continuous case, the insignificant coefficient on the lead measure of linguistic distance in Table A-4 indicates a lack of anticipatory effects, suggesting that the key identifying assumption is satisfied.

In this section, I provide additional evidence supporting the validity of my estimation strategy. To achieve this, I explore whether the ethnic-specific reaction to cultural distance is different depending on whether government transitions occur for more idiosyncratic reasons or not. Suppose that there is an omitted variable—discontent or dissatisfaction, which is not directly observable—that causes both a government change and the ensuing conflict among groups that become more culturally distant after the transition. Insofar as discontent towards the government is widespread across ethnic groups within a country, this omitted variable would be absorbed by country-by-year fixed effects, which also control for the overall effect

that a government transition has on all ethnic groups in a country. However, one concern could be the following: there is pent-up discontent among some potential rebels → the ethnic identity of the government changes as a result of this pent-up discontent → the government becomes *more* culturally distant from the dissatisfied potential rebels → these potential rebels start expressing their *pre-existing* discontent against the *new* government by fighting against it. In this scenario, conflict could be the result of some pre-existing dissatisfaction, and not of the increased cultural distance. For this to be a concern in my analysis, pent-up discontent would need to (i) systematically generate government transitions that disfavour dissatisfied ethnic groups (i.e., dissatisfied groups will always become more culturally distant from the newly-appointed government) and (ii) generate conflict against a newly-appointed government, and not against the one that caused the discontent in the first place.

Even though it seems implausible that these two conditions would emerge simultaneously and systematically, I propose some heterogeneity analyses to test for this possibility. The goal of this exercise is to identify government transitions that may be more "exogenous"—and thus likely unaffected by the just-described concerns—and assess whether the reaction to cultural distance is significantly different around these transitions.

First, I start by using Jones and Olken's (2005) data on the deaths of leaders. The timing of a transition occurring as a result of the death of a leader is arguably exogenous, as emphasized in Jones and Olken (2005). This data is available from 1945 to 2000 and contains information on the cause of death of leaders around the world. Among the 17 deaths of African leaders that occurred between 1961 and 2000, only 7 generated changes in the ethnic identity of the government in my data (see Table B-6 below). If the above-described concern was present, we should observe the effect of cultural distance on conflict to be muted or significantly smaller in cases where a transition occurs for more idiosyncratic reasons such as the death of a leader. I do not find this to be the case: the coefficient on the interaction term between cultural distance and an indicator capturing a transition

induced by the death of a leader is small and not significant, which suggests that the effect of cultural distance is not muted when government transitions occur for more idiosyncratic reasons.[13] However, I would caution against drawing strong conclusions from this test, given the few government transitions induced by the death of a leader.

Second, I conduct a similar analysis distinguishing between government transitions that are the result of "close elections" and those that are not. While the deaths of leaders makes the *timing* of the transition exogenous, close elections make the *identity* of the winner exogenous. For this exercise, I aim at identifying government transitions that occur at the same time as elections where winners obtained a relatively narrow margin of victory.[14] To achieve this, I consult the African Elections Database, which provides information on elections held in Africa until 2013.[15] In addition to listing dates of various types of elections (presidential elections, national assembly/house of representatives/parliamentary elections, constitutional referenda), the database includes information on the number of votes received by presidential candidates and the number of seats assigned to each party in the legislature. I start by identifying those changes in the ethnic identity of the government in my sample that were induced by elections.[16] Since most countries in my setting are characterized by presidential political systems where presidents exercise executive power, presidential elections are commonly those that generate a change in the ethnic identity of the government. I classify as "close elections" those transitions where the margin of victory of the winning presidential candi-

[13]This indicator takes value 1 for the year in which the transition takes place as well as for two periods before and two periods after.

[14]Close elections are commonly used in the literature to identify the effect of narrowly elected politicians or parties on downstream outcomes of interest, typically using regression discontinuity designs, based on the idea that the outcome of such elections is as close as random. Recent critiques question the randomness of close elections and emphasize that candidates and parties with specific characteristics (e.g., incumbents) are systematically more likely to win close elections (see De la Cuesta and Imai (2016) for a summary). Moreover, in my setting, close elections may be more likely to be contested and thus spur conflict. The reader should keep all these caveats in mind when interpreting the results proposed in the remainder of this section.

[15]https://africanelections.tripod.com/about.html

[16]53 percent of changes in the ethnic identity of the government occur at the same time as a a presidential and/or parliamentary election.

date is smaller than 20 percentage points, 10 percentage points, or 5 percentage points.[17] In non-presidential systems, or when only a parliamentary election occurs in conjunction with a change in the ethnic identity of the government, I classify transitions as coming from "close elections" whenever the difference between the percentage of seats obtained by the winning party and the percentage of seats obtained by the second most voted party is smaller than 20, 10 or 5 percentage points. As illustrated in Table B-7, 24 government transitions can be attributed to close elections with less than a 20 percentage points margin, 16 with less than a 10 percentage points margin, and 8 with less than a 5 percentage points margin in my sample. Columns 3-6 in Table B-8 show that the reaction to cultural distance around these more "exogenous" transitions is not statistically different from the reaction around other types of government changes.

Third, I test for heterogeneity by whether a government transition occurs at the end of a "natural" political cycle or not. The idea is that, if a transition occurs earlier than what a typical political term would prescribe, it is more likely that it was the result of the discontent and dissatisfaction described in the example above. If the effect of cultural distance I uncover was entirely driven by these transitions, one may worry that omitted factors could be driving the results. Based on this idea, I construct two measures. First, I categorize countries into whether their average duration of governments is above or below 5 years, i.e., the median length of governments in the sample. Second, I exploit information on the length of political terms by country[18] and distinguish between government transitions that occur earlier than the country-specific term would have prescribed and transitions that occur at the end of or after the term. Reassuringly, when investigating heterogeneity along these dimensions, I do not find that the effect of cultural distance is entirely explained by countries that experience more frequent

[17] When multiple election rounds occur, I consider the margin of victory in the decisive round.

[18] I use information collected by Nohlen et al. (1999): information on terms (in years) is in Table 2. The table refers to presidential terms; since most countries are characterized by presidential forms of government, it is safe to assume that the length of government terms is comparable. Interestingly, the most common term length is five years, precisely as the median length of governments in the sample.

government transitions or by government transitions that occur before the end of a "natural" political cycle. Table B-8 shows that there is no sizeable and significant heterogeneity, which is reassuring of the fact that government changes occurring at higher frequencies are not driving the results.

Fourth, I focus more closely on the specific example proposed above: there is pent-up discontent with the government among a subgroup of potential rebels → the ethnic identity of the government changes as a result of this pent-up discontent → the government becomes *more* culturally distant from the sub-group of potential rebels that were dissatisfied → the same potential rebels start expressing their *pre-existing* discontent against the new government by fighting against it. Suppose that, as in this scenario, the government leadership retaliates against unsatisfied groups by changing the composition of the government and forming coalitions with ethnic groups that are culturally distant from the dissatisfied ones. The latter, in turn, react by entering conflict against the new coalition. If these groups were rebelling due to their pre-existing dissatisfaction (which caused the government to become more culturally distant), and not much due to a reaction to cultural distance, my results would have to be interpreted differently. To explore this possibility, I distinguish between two types of government transitions: those where the government leadership (i.e., ethnic group(s) holding monopoly, dominant, or senior government power) does not change and those where it does. Since the just-described hypothetical scenario is realistically occurring only where the government leadership remains unchanged, I assess whether the effect of cultural distance on conflict is concentrated among those types of transitions. As shown in Table B-8 below, there is no significant heterogeneity along this dimension.

Fifth, to further alleviate any remaining concerns associated with the potential endogeneity of government changes, in Table B-9 I re-run my analysis using as identifying variation only the changes in linguistic distance that are induced by arguably more "exogenous" transitions, i.e., those occurring due to the death of a leader or close elections.[19] To achieve this, I start by constructing a variable (called

[19]To allow for enough variation to conduct this exercise, I consider elections where the margin

Linguistic distance change, or ΔLD_{rct}) that captures the year-to-year changes in linguistic distance to the government for each ethnic group:

$$\Delta \text{LD}_{rct} = \text{LD}_{rct} - \text{LD}_{rct-1}$$

Table B-9, columns 1 and 2 show that the main results hold when re-running the baseline specification using ΔLD_{rct} as explanatory variable. Next, I replace *Linguistic distance change* with *"Exogenous" linguistic distance change*, or $\Delta \text{LD}_{rct}^{exogenous}$, a variable that captures only the changes in linguistic distance induced by "exogenous" transitions:

$$\Delta \text{LD}_{rct}^{exogenous} = \begin{cases} \text{LD}_{rct} - \text{LD}_{rct-1} & \text{if leadership change at } t \text{ is "exogenous"} \\ 0 & \text{otherwise (no change/"endogenous" change)} \end{cases}$$

Despite the considerably lower identifying variation that can be exploited by focusing solely on these changes—only 31 out of 138 transitions are induced by the death of a leader or close elections—and keeping the caveats associated with close elections in mind (see footnote 14 in this section), the results in columns 3 and 4 show that the effect of cultural distance on conflict remains positive and significant. Columns 5 and 6 restrict the sample to the period for which information on close elections is available, and conclusions remain valid.

Taken together, the analyses proposed in this section suggest that the effects I uncover do not seem to be driven by more "endogenous" government transitions and that an effect of cultural distance on conflict can be found also among the most arguably "exogenous" transitions in the sample.

of victory of the winning party is smaller than 20 percentage points.

TABLE B-6: *Deaths of Leaders and Changes in the Ethnic Identity of the Government*

Country	**Leader's name**	**Year**	**Cause of death**	**Change in the ethnic identity of government**
Angola	Agostinho Neto	1979	Cancer of the pancreas	no
Botswana	Sir Seretse Khama	1980	Cancer of the stomach	no
Comoros	Prince Jaffar	1975	While on pilgrimage to Mecca	no
Comoros	Mohamad Taki	1998	Heart attack	yes
Cote d'Ivoire	Felix Houphouet-Boigny	1993	Following surgery for prostate cancer	yes
Egypt	Gamal Abdel Nasser	1970	Heart attack	no
Gabon	Leon Mba	1967	Cancer (in Paris)	yes
Guinea	Sekou Toure	1984	Heart attack during surgery in Cleveland	yes
Kenya	Jomo Kenyatta	1978	While sleeping	yes
Mauritania	Ahmed Ould Bouce	1979	Plane crash in sandstorm over Atlantic	yes
Morocco	Mohammed V	1961	Following operation to remove growth in throat	no
Morocco	Hassan II	1999	Heart attack	no
Mozambique	Samora Machel	1986	Plane crash near Maputo	no
Niger	Seyni Kountche	1987	Cancer (brain tumor)	no
Nigeria	Sani Abacha	1998	Heart attack (some say poisoned)	yes
Sierra Leone	Sir Milton Margai	1964	After "brief illness"	no
South Africa	Johannes G. Strijdom	1958	Heart disease	no

Notes: Source: (Jones and Olken, 2005, p.846-848)

TABLE B-7: *Government Transitions with Close Elections*

Country	Political System	Year of transition	Election	Margin of victory (pp)
Benin	Presidential system without a prime minister	1970	Presidential	4.25
Benin	Presidential system without a prime minister	1996	Presidential & National Assembly	4.98
Central African Republic	Presidential system with a prime minister	1994	Presidential & National Assembly	6.98
Congo, DRC	Semi-presidential system of government	2007	Presidential & National Assembly	16.10
Cote d'Ivoire	Presidential system with a prime minister	2012	National Assembly	18.04
Ghana	Presidential system without a prime minister	2001	Presidential & Parliamentary	13.80
Ghana	Presidential system without a prime minister	2009	Presidential & Parliamentary	0.46
Guinea	Presidential representative democratic republic	2010	Presidential	5.04
Guinea	Presidential representative democratic republic	2011	Presidential	5.04
Guinea-Bissau	President-parliamentary system	2006	Presidential	4.70
Kenya	Presidential system without a prime minister	2008	Presidential & National Assembly	2.35
Liberia	Presidential system without a prime minister	2006	Presidential & Legislative	18.80
Malawi	Presidential system without a prime minister	1995	Presidential & National Assembly	13.71
Mauritius	Parliamentary democracy	1977	Legislative Assembly	3.22
Niger	Semi-presidential system of government	1994	Presidential & National Assembly	8.84
Niger	Semi-presidential system of government	1996	National Assembly	6.02
Niger	Semi-presidential system of government	2000	Presidential & National Assembly	19.78
Niger	Semi-presidential system of government	2012	Presidential & National Assembly	16.08
Nigeria	Presidential system without a prime minister	1979	Presidential & National Assembly	4.59
Nigeria	Presidential system without a prime minister	1984	Presidential & National Assembly	16.42
Sierra Leone	Presidential system without a prime minister	1968	House of Representatives	8.49
Sierra Leone	Presidential system without a prime minister	2008	Presidential & House of Representatives	9.24
Uganda	Presidential system with a prime minister	1980	National Assembly	0.14
Zimbabwe	Presidential system without a prime minister	2000	House of Assembly	1.60

Notes: Source: African Elections Database (https://africanelections.tripod.com/about.html).

TABLE B-8: *Heterogeneity Analysis by Type of Government Transition*

Sample restriction:	Dep. var: conflict over power									
	until 2000		*until 2013*				*none*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Linguistic distance	0.082*	0.082*	0.075***	0.081***	0.077**	0.079***	0.086***	0.095***	0.080***	0.075***
	(0.046)	(0.046)	(0.026)	(0.029)	(0.029)	(0.028)	(0.019)	(0.035)	(0.021)	(0.021)
Linguistic distance ×										
death of leader		-0.009								
		(0.005)								
Linguistic distance ×										
close elections (20pp margin)				-0.022						
				(0.015)						
close elections (10pp margin)					-0.008					
					(0.016)					
close elections (5pp margin)						-0.020				
						(0.015)				
Linguistic distance ×										
below median government length								-0.017		
								(0.037)		
transition before term									0.026	
									(0.017)	
no change in leadership										0.037
										(0.035)
Observations	6,384	6,384	9,134	9,134	9,134	9,134	9,990	9,990	9,990	9,990
Adjusted R-squared	0.638	0.637	0.518	0.518	0.518	0.518	0.507	0.507	0.507	0.508

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Linguistic distance* captures linguistic distance between each potential rebel and the ethnic groups at the government, weighted by the position of power of each ethnic group in case the government is a coalition. The variable *death of leader* is an indicator that equals 1 for the two years before, the year of, and two years after a government transition induced by the death of a leader; *close election (20pp margin)*, *close election (10pp margin)*, *close election (5pp margin)* is an indicator that equals 1 for the two years before, the year of, and two years after a government transition induced by close elections with a 20pp, 10pp, and 5pp margin of victory; *below median government length* is an indicator that equals 1 for countries whose governments last on average less than the median duration in the sample (5 years); *transition before term* is an indicator that equals 1 for the two years before, the year of, and two years after a government transition that occurred before the country-specific government term; *no change in leadership* is an indicator that equals 1 for the two years before, the year of, and two years after a government transition where the ethnic identity of the leadership remains unchanged. All specifications include ethnicity-country fixed effects, country-year fixed effects, ethnicity-by-country-specific year trends, and geographic controls. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE B-9: *Exploiting only Arguably Exogenous Changes in Linguistic Distance as Source of Variation*

Sample restriction:	Dep. var: conflict over power					
	none				*until 2013*	
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance change	0.226**	0.201**			0.215**	
	(0.096)	(0.084)			(0.089)	
"Exogenous" linguistic distance change			0.151***	0.138**		0.140**
			(0.044)	(0.057)		(0.064)
Mean of dep. var.	0.048	0.048	0.048	0.048	0.051	0.051
Country-year FE	yes	yes	yes	yes	yes	yes
Ethnicity-country FE	yes	yes	yes	yes	yes	yes
Ethnicity-country trend	yes	yes	yes	yes	yes	yes
Geographic controls		yes		yes	yes	yes
Observations	9,676	9,676	9,676	9,676	8,591	8,591
Adjusted R-squared	0.514	0.515	0.513	0.514	0.530	0.529

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Linguistic distance change* captures year-by-year changes in linguistic distance between each potential rebel and the ethnic groups in government (weighted by the position of power of each ethnic group in case the government is a coalition) and is constructed as follows: Linguistic distance$_t$ - Linguistic distance$_{t-1}$. *"Exogenous" linguistic distance change* captures only the year-by-year changes in linguistic distance between each potential rebel and the ethnic groups in government that are induced by either the death of leader or a close election (20pp margin). All other (more "endogenous") changes are set to zero. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-3 Cultural Distance and the Intensive Margin of Conflict

In this section, I explore whether cultural distance influences the intensive margin of conflict and not only whether conflict takes place or not. The theoretical framework linking cultural distance to conflict by Spolaore and Wacziarg (2017), on which this paper is based, focuses on the "probability of conflict between populations." In their game-theoretical framework, a group can decide to challenge/not challenge the other group which, in turn, can respond by fighting/not fighting. By modelling the probability that two groups enter conflict with each other, this framework does not provide a clear prediction on whether distance matters for the intensive margin of a hypothetical conflict that emerges. In fact, the same authors show empirically that genetically closer countries are more likely to enter conflict with each other over rival goods, as predicted by their model (Spolaore and Wacziarg, 2016); in supplementary analyses, however, they show that most of the effect stems from the extensive margin and that genetic distance does not explain the scale of the war. I find similar results in the context of this paper.

Data on conflict fatalities is, unfortunately, only available for a limited time period (from 1989 onwards). Conveniently, however, UCDP dyadic provides a categorical variable that captures the intensity of the conflict for my entire study period, which is based on the underlying number of battle-related deaths. This variable can take two values: "Minor: between 25 and 999 battle-related deaths" or "War: at least 1,000 battle-related deaths in a given year." Based on this indicator, I construct two alternative binary variables: one that equals 1 for small-scale conflicts (minor), and one that equals 1 for large-scale conflicts (war).

I re-run my baseline specification using four alternative outcome variables: the variable provided by UCDP on the intensity of the conflict (ranging between 0 and 2, where 0 denotes no conflict, 1 small-scale conflict, and 2 large-scale conflict); a variable capturing the log number of battle-related deaths (from 1989 onwards only), and two indicators for small-scale and large-scale conflict, respectively. These latter two specifications will allow me to tease out whether cultural

distance triggers only small-scale conflicts, only large-scale conflicts, or both. Finally, I examine whether linguistic distance explains conflict intensity conditional on conflict taking place.

Table B-10 presents the results. Column 1 shows that replacing the binary outcome variable with a variable capturing conflict intensity does not alter the message of the paper: linguistic distance remains positively and significantly associated with conflict over government power. Column 2, which employs battle-related deaths as an outcome, also shows a consistent estimate, albeit imprecisely estimated due to the limited time period under consideration (the p-value is 0.19). Columns 3 and 4 show that linguistic distance does not seem to be driving only one type of conflict: conflicts emerging from a group's linguistic distance to the government can be of both small and large scale. This intuition is confirmed in column 5: once conditioning on conflict, linguistic distance is not a powerful predictor of the intensity of such conflict. These results are in line with Spolaore and Wacziarg's (2016) findings on genetic distance and international conflict between countries mentioned above.

Despite these findings, more research is needed to assess how cultural distance affects the type of violence employed in conflicts once they break out. The intensive margin of violence (i.e., how conflict manifests once it takes place) has received less attention in the literature compared to the extensive margin (i.e., whether or not conflict happens at all). A notable exception is the work by Moscona et al. (2020), who empirically test the hypothesis that African societies organized along segmentary lineages are more likely to experience conflict. Here, the main hypothesis also yields a prediction about the scale of the conflicts that break out: as ethnographic accounts suggest, individuals are obliged to help fellow lineage members when they become involved in conflicts, and this generates conflicts that are larger in scale and longer in duration.

TABLE B-10: *Cultural Distance and Conflict Intensity*

	Dependent variable:				
	Intensity	ln(1+battle deaths)	Small-scale	Large-scale	Intensity
	(1)	(2)	(3)	(4)	(5)
Linguistic distance	0.137***	0.309	0.034**	0.051***	0.008
	(0.035)	(0.232)	(0.015)	(0.019)	(0.033)
Mean of dep. var.	0.068	0.314	0.028	0.020	1.46
Country-year FE	yes	yes	yes	yes	yes
Ethnicity-country FE & trends	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes
Observations	9,990	5,900	9,990	9,990	326
Adjusted R-squared	0.534	0.647	0.377	0.487	0.982

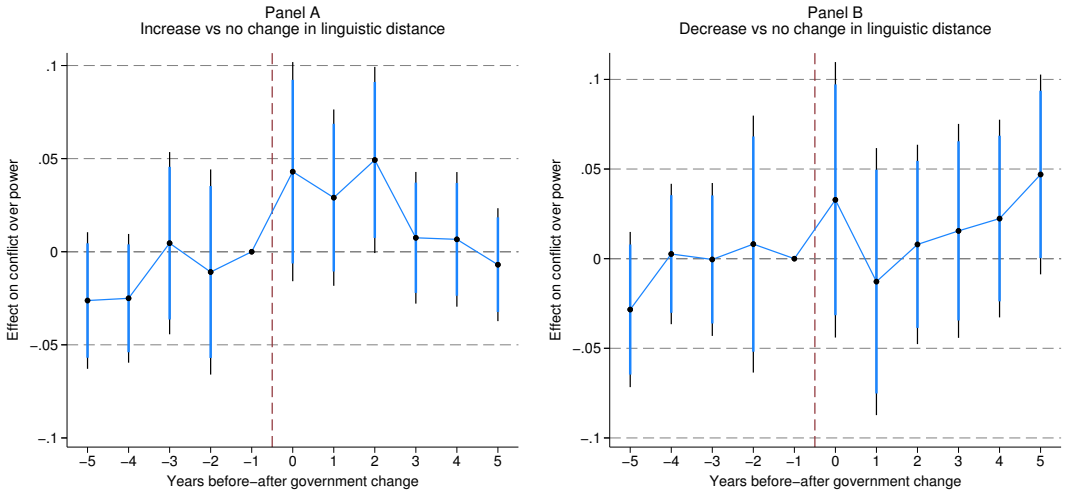
Notes: The unit of observation is an ethnic group-country-year. In column 1, the dependent variable takes value 0 if no conflict over government power occurs, 1 if a small-scale conflict (between 25 and 999 battle-related deaths) occurs between a potential rebel and the government in a specific year and 2 if a large-scale conflict occurs (at least 1,000 battle-related deaths in a given year); in column 2, the dependent variable is ln(1+battle-related deaths) and is only available from 1989 onwards; in column 3, the dependent variable is an indicator that takes value 1 if a small-scale conflict takes place and 0 otherwise; in column 4, the the dependent variable is an indicator that takes value 1 if a small-scale conflict takes place and 0 otherwise; in column 5, the sample is restricted to potential rebels that engage in conflict against the government and the dependent variable takes either value 1 (small-scale conflict) or 2 (large scale conflict). Two-way clustered standard errors by year and country (columns 1-4) are reported in parenthesis. In column 5, robust standard errors are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-4 Event Study Design: Additional Results and Robustness

In this section, I report some additional results and propose robustness tests supporting the validity of the event-study design.

I start by exploring whether the effect I uncover is explained by groups that become more culturally distant from the government rebelling more after a transition, by groups that become culturally closer rebelling less, or both. To tease this out, I start by estimating equation 4 removing from the sample ethnic groups whose linguistic distance decreases after a transition. I thus compare the reaction of groups becoming culturally more distant only to that of groups who face the transition but whose linguistic distance remains unchanged. I then estimate an equivalent of equation 4, but replacing More Dist_{rce} with Less Dist_{rce}, which is an indicator that equals 1 if a group becomes culturally closer following transition e and 0 if an ethnic group's linguistic distance remains unchanged. Ethnic groups whose linguistic distance increases after transition e are removed from this sample. Panel A and B in Figure B-1 show the results of these exercises: a positive and sizeable effect of cultural distance can only be found among groups becoming culturally more distant from the government.

FIGURE B-1: *Effect of Becoming More vs Less Distant from the Government (I)*



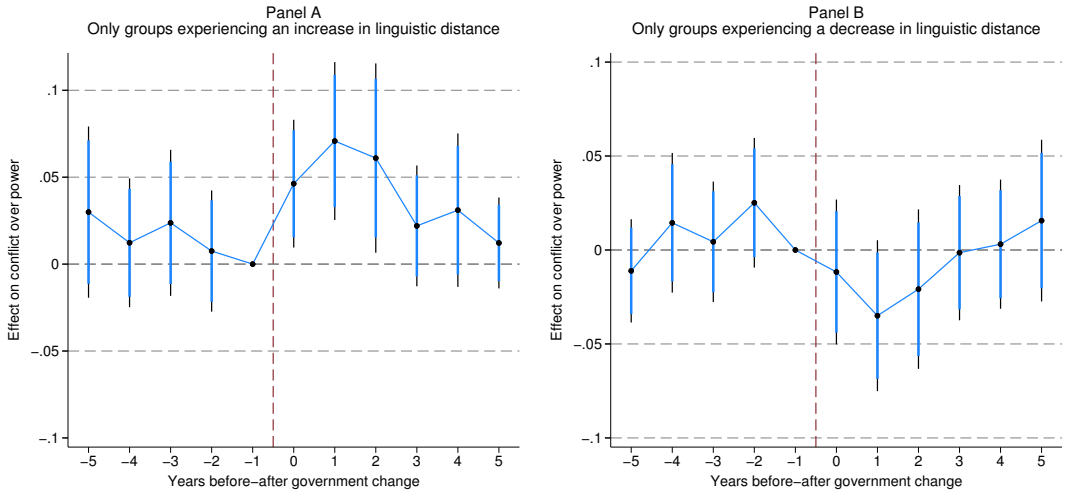
Notes: Sample size: 3,916 in panel A (groups whose linguistic distance increases account for 74 percent of the sample); and 3,999 in panel B (groups whose linguistic distance decreases account for 73.5 percent of the sample). Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

Alternatively, I split the sample and estimate the following event study specification separately for (i) groups that become more culturally distant from the government after a transition and (ii) for groups becoming culturally closer to the government after a transition.

$$\text{Confl}_{rcte} = \sum_{\substack{k=-5,\\ k\neq -1}}^{5} \alpha_k \text{T}_{et}^{k} + \lambda_{ct} + \zeta_{rc} + \theta_{rc}t + \gamma \text{G}_{rct} + \epsilon_{rcte} \tag{2}$$

where T_{et}^{k} is an indicator equal to 1 if calendar year t is k years before (or after) government transition e and zero otherwise. Panels A and B in Figure B-2 display estimates of α_k for the two subsample of groups, respectively. While groups becoming closer to the government seem to become less likely to rebel after a transition, the largest and most statistically significant effect is present among groups becoming more culturally distant. Reassuringly, this specification confirms the lack of any significant pre-transition trend in conflict among both sub-groups of ethnicities.

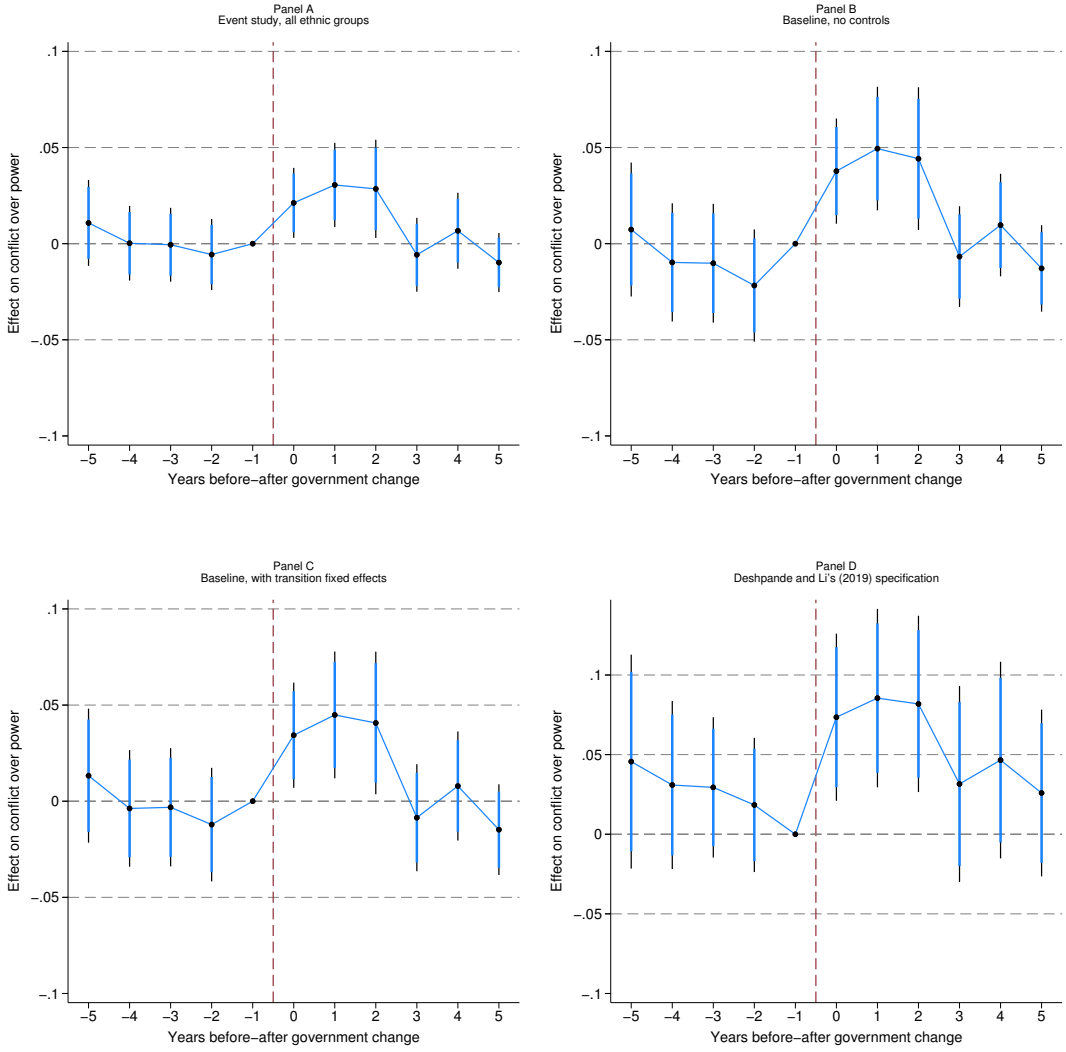
FIGURE B-2: *Effect of Becoming More vs Less Distant from the Government (II)*



Notes: Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

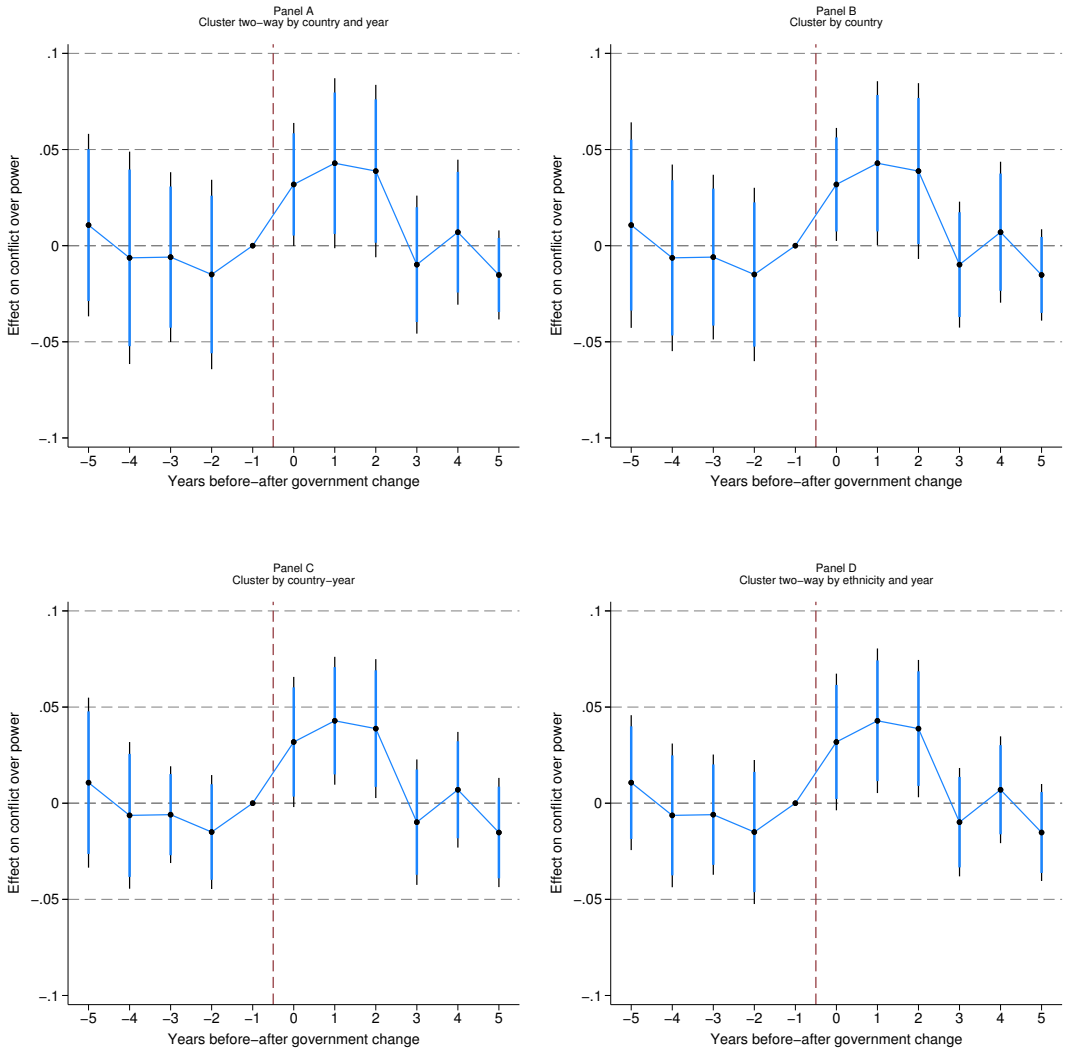
Next, Figure B-3, Figure B-4, and Figure B-5 below show that results are robust to alternative specifications, clustering of the standard errors, and estimation sample. In Figure B-3, I first re-run the specification in equation 2 above keeping all ethnic groups in the sample, but setting T_{et}^{k} to zero for control ethnic groups (Panel A). Second, I remove geographic controls from the baseline specification (Panel B). Third, I add fixed effects for each government transition (Panel C). Fourth, I remove ethnicity-specific year trends and I cluster standard errors at the government transition level to estimate a specification identical to Deshpande and Li's (2019) (Panel D). Figure B-4 shows that inference remains similar with alternative levels of clustering of the standard errors, analogously to what shown in Table A-3 (columns 3-6) for the baseline specification. Figure B-5 re-runs the event study keeping only those events and ethnic groups that allow for perfect compositional balance across periods within the event window, i.e. keeping only ethnic groups that can be followed throughout the whole time window around a government transition. The results remain remarkably similar, which suggests that the effects are not an artifact of issues related to the sample composition.

FIGURE B-3: *Robustness Tests*



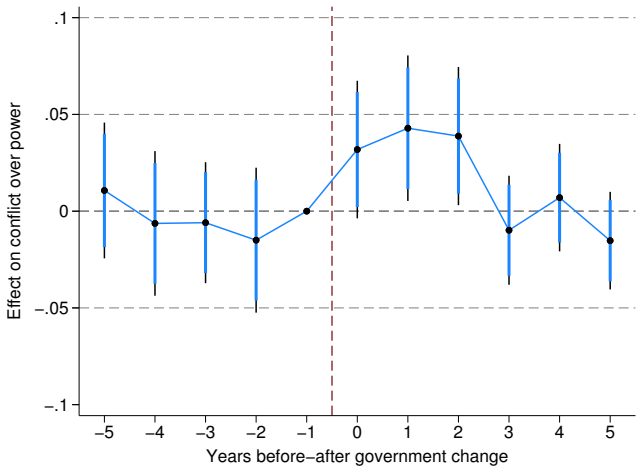
Notes: Panel A shows estimates of α_k in equation 2 keeping all ethnic groups in the sample, but setting T_{et}^k to zero for control ethnic groups; Panel B removes geographic controls from equation 4; Panel C adds government transition fixed effects; Panel D removes ethnicity trends and clusters standard errors at the government transition level as in Deshpande and Li (2019). Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level in Panels A, B, and C and at the government transition level in Panel D.

FIGURE B-4: *Alternative Clustering*



Notes: The graphs report the baseline event-study specification employing different level of clustering of the standard errors. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively.

FIGURE B-5: *Event Study with Compositional Balance*



Notes: The graph reports the baseline event-study specification on a sample with compositional balance, i.e., a sample of 143 ethnic groups that can be followed throughout the whole time window around a government transition.

Finally, I propose a stacked event study with compositional balance and clean controls to address potential concerns with my event-study specification related to the negative-weighting issues discussed by recent literature. As emphasized in Section 5.1, there are some challenges I need to overcome to be able to estimate an event study in my setting, which makes it not easily comparable to more standard event studies. In a standard staggered design, units receive treatment once and treatment is absorbing. However, this is not the case in my setting. The event of interest here is a government transition (indexed by e), and each government transition can be thought of as a "sub-experiment." The goal of the event study is to compare the conflict behavior of groups that experience an increase in linguistic distance after a transition (these are the "treated" ethnic groups) to that of groups within the same country that do not (these are the "control" ethnic groups), focusing on the 5 periods before and 5 periods after each transition or sub-experiment e. Because government transitions occur repeatedly over the sample period, and the same ethnic group may be treated in some transitions but not others, my setting is not comparable to a standard staggered design where units receive treatment once and treatment is absorbing. Moreover, as described in Section 5.1, to be able to estimate dynamic effects for every government transition in the sample I have to take into account another special feature of my setting: some observations (i.e., ethnicity-country-years) can be classified as both belonging to the "after" period of a government transition and to the "before" period of another, subsequent gov-

ernment transition. As such, the dataset I construct to estimate the event study is not a panel dataset at the unit-time level; it is an "augmented" dataset where the variables uniquely identifying the observations are ethnicity, country, year, and *transition e*.

Most of the existing estimators proposed by the literature to circumvent the negative weighting issue have been developed for designs with staggered treatment adoptions and absorbing treatments (i.e., where groups' treatment can only increase over time and can change at most once) and/or for panel datasets where each observation is at the unit-time level (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2021; De Chaisemartin and d'Haultfoeuille, 2020). Unfortunately, these are not directly applicable to my case given the just-described peculiarities of my setting and dataset structure.

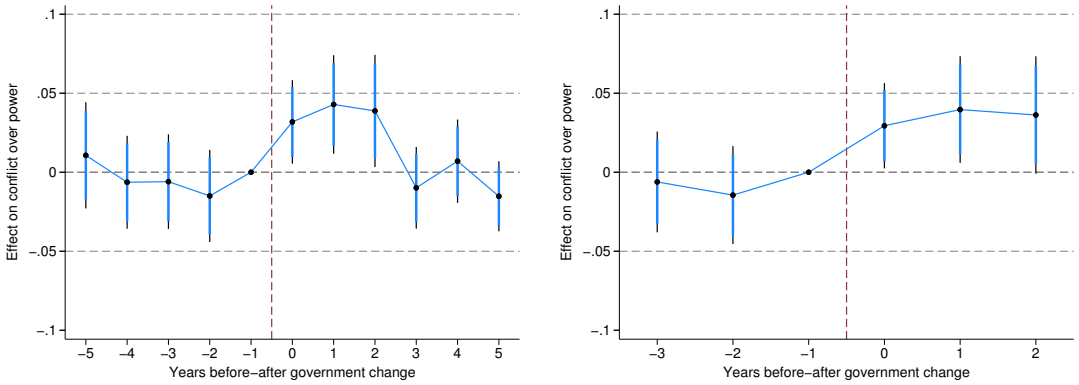
I now turn to discussing how the negative-weighting issue may materialize in my case and adopt a solution. The negative-weighting problem highlighted by the recent literature comes from "forbidden comparisons," whereby treated units are sometimes compared to already treated units in staggered designs. If treatment effects are heterogeneous, forbidden comparisons can generate biased estimates. In my setting, the issue of forbidden comparisons could arise if an ethnic group whose linguistic distance increases due to transition e (i.e., a *treated* ethnic group) is compared to an ethnic group whose cultural distance does not increase during transition e (this makes it a *control* ethnic group), but whose distance has increased due to a different transition e' that happened in close temporal proximity in the same country. This could materialize in situations where government changes happen close to one another, and the sub-experiment e is therefore "contaminated" with unclean controls.

To ensure that this potential issue does not affect my results, I refine my event-study sample and specification and estimate a *stacked event study design* with compositional balance and clean controls, as described by Baker et al. (2022), Freedman et al. (2023) and Wing et al. (2024), and employed in recent applied literature (see, e.g., Cengiz et al. (2019); Deshpande and Li (2019); Butters et al. (2022)). Stacked event studies are one of the solutions to the negative-weighting issue that have been proposed by the literature, and is straightforwardly applicable to my setting. In a stacked design, the researcher circumvents the issue of forbidden comparisons by focusing only on *valid* sub-experiments characterized by "clean controls" and "compositional balance" around each event. Clean controls are units that do not experience treatment during the sub-experiment time window; compositional balance ensures that all units are observed across the whole time

window.[20] The researcher then stacks all these valid sub-experiments into a single dataset and estimates an event study.

To apply the stacked approach, I restrict the focus to a 6-year time window around each government transition (rather than a 10-year window as in my baseline specification). This window allows me to maximize the number of valid sub-experiments and, at the same time, (i) test for pre-trends and (ii) estimate dynamic effects for the post-treatment period for which I had identified positive and significant effects in Section 5.1 (time 0, 1, and 2, respectively). A too wide time window would force many sub-experiments out of the sample, by increasing the probability of contaminating the sub-experiment with unclean controls. For comparison, Figure B-6 shows that, in my baseline event-study specification, restricting the time window does not alter the results.

FIGURE B-6: *Baseline Event Study Design, Narrower Window*



Notes: The figure plots coefficient estimates of τ_k described in equation 4 focusing on 3 periods before and after a government transition. The sample size is 6,223. Groups whose linguistic distance increases after a transition e account for 43 percent of the sample. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnicity-by-country level.

Next, I construct a separate dataset for each valid sub-experiment. A valid sub-experiment, in my setting, satisfies two conditions: (i) control units are ethnic groups within the same country that do not experience an increase in linguistic distance to the government in the specified time window; (ii) all the ethnic groups included in the sub-experiment are observed for the entire time window. Finally, I append all these valid sub-experiments into a single dataset. This dataset is

[20]Compositional balance is not always imposed in the cited applied settings, but I adopt it following Baker et al.'s (2022), Freedman et al.'s (2023) and Wing et al.'s (2024) recommendations. As noticed by Baker et al. (2022), compositional balance guards against the bias arising due to heterogeneous treatment effects that appears in the ΔATT term in Goodman-Bacon (2021).

entirely comparable to my baseline event-study dataset, but differs in that there is compositional balance and only clean controls.

Using this refined dataset, I estimate an event-study specification similar to my baseline specification, but with some important differences, as required in stacked designs: I replace ethnicity fixed effects with ethnicity×sub-experiment fixed effects (ζ_{rce}), and I replace event-time fixed effects with a set of event-time×sub-experiment fixed effects (μ_{te}).[21] I apply Wing et al.'s (2024) corrective sample weights to avoid that different implicit weights are applied to treatment and control trends.[22] Wing et al. (2024) highlight that this weighting procedure is equivalent to the Callaway and Sant'Anna (2021) method applied to a stacked event-study design with compositional balance.[23]

As in Butters et al. (2022), I start by estimating the following difference-in-differences specification using the stacked dataset:

$$\text{Confl}_{rcte} = \tau \text{Post}_{te} \times \text{MoreDist}_{rce} + \mu_{te} + \zeta_{rce} + \theta_{rc} t + \gamma \text{G}_{rcte} + \epsilon_{rcte} \tag{3}$$

where Confl_{rcte} is a conflict indicator for potential rebel r in country c observed in year t and sub-experiment e. Post_{te} indicates the post-government transition period in each sub-experiment e and MoreDist_{rce} indicates treated ethnic groups in each sub-experiment. To trace out effect dynamics, I then estimate the event-study version of equation 3:

$$\text{Confl}_{rcte} = \sum_{\substack{k=-3, \\ k\neq -1}}^{3} \tau_k \text{T}_{et}^{k} \times \text{MoreDist}_{rce} + \mu_{te} + \zeta_{rce} + \theta_{rc} t + \gamma \text{G}_{rcte} + \epsilon_{rcte} \tag{4}$$

[21] This is equivalent to Cengiz et al. (2019) and Butters et al. (2022) implement in their applications. I also include geographic controls and ethnicity-specific year trends, so that this specification is fully comparable to the one presented in the paper.

[22] As described in Wing et al. (2024) in their Table 1, the weights are constructed as follows. Treated units receive a sample weight equal to 1. Control units receive the following weight: $(N_e^D/N^D)/(N_e^C/N^C)$, where N_e^D is the number of groups that first adopt treatment in sub-experiment e; N^D is the total number of groups that ever adopt treatment; N_e^C and N^C are the analogous counts for the control groups.

[23] Together, compositional balance and corrective weighting address the issue highlighted by Baker et al. (2022) with regards to using OLS to weight treatment effects in a stacked DID setting.

where ζ_{rce} denote ethnicity×sub-experiment fixed effects and μ_{te} denote time×sub-experiment fixed effects. Standard errors are clustered at the ethnic-group-by-sub-experiment level.

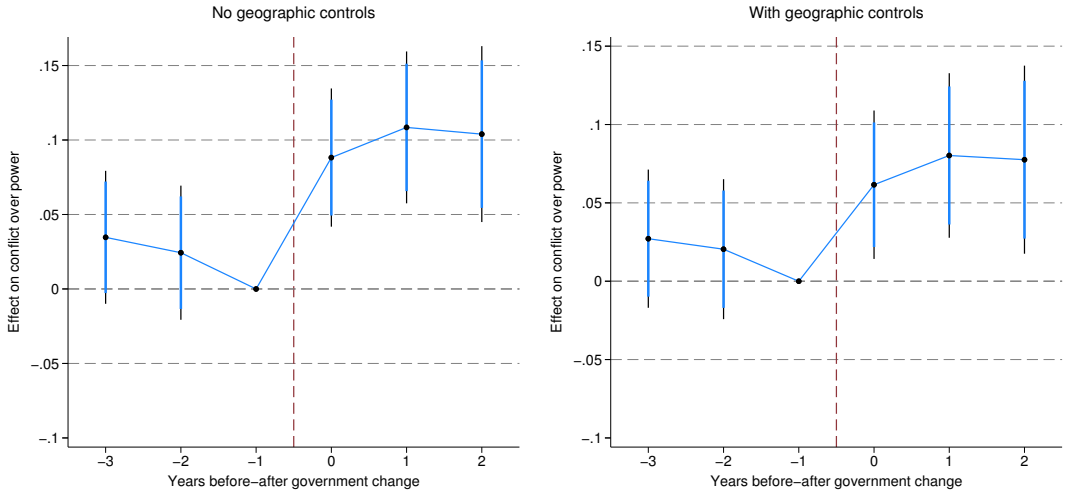
Results are reported in Table B-12 and Figure B-8. For comparison, Table B-11 and Figure B-7 report the results from running specifications described in equation 3 and 4 on the original dataset, i.e., without imposing clean controls or compositional balance. Despite the restrictions to the sample imposed by the clean controls and compositional balance rules, the estimates remain positive and significant. Given these reassuring findings, I conclude that it is unlikely that negative weighting issues are driving my results.

TABLE B-11: *Stacked DID Design: Full Sample*

	Conflict over power	
	(1)	(2)
More Distant × Post	0.081***	0.057***
	(0.021)	(0.021)
Mean dependent variable	0.060	0.060
Geographic controls		yes
Observations	4,002	4,002
Adjusted R-squared	0.465	0.477

Notes: The unit of observation is an ethnic group-country-year-sub-experiment. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. *More Distant* is an indicator that equals 1 if an ethnic group becomes more culturally distant from the government in a given sub-experiment, and 0 otherwise. *Post* is an indicator that equals 1 for time periods after a government change in each sub-experiment. The dataset includes 138 distinct sub-experiments. The sample includes 34 African countries and 210 distinct ethnic groups. Standard errors clustered at the ethnic-group-by-sub-experiment level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

FIGURE B-7: *Stacked Event Study Design: Full Sample*



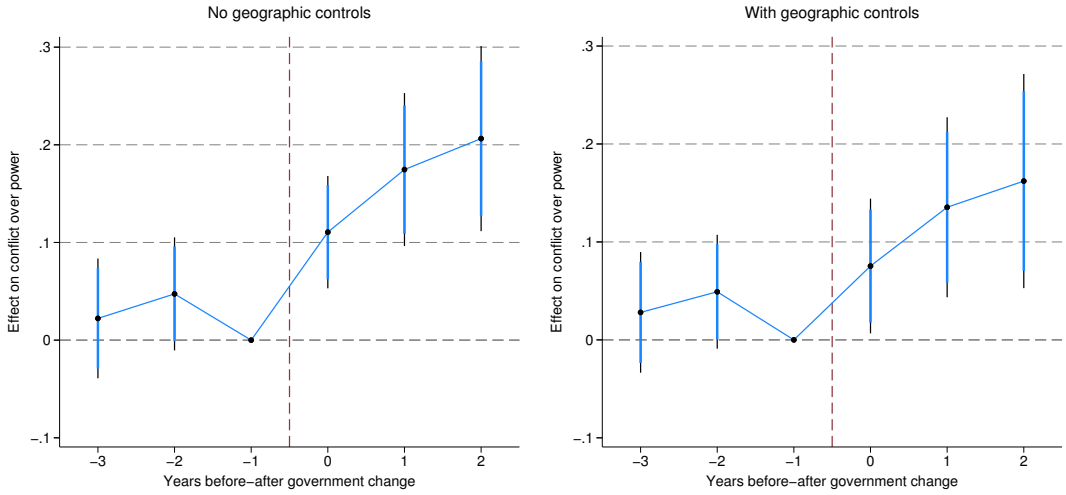
Notes: The figures plot coefficient estimates of τ_{ke} described in equation 4. The left graph does not include geographic controls, while the graph on the right does. Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnic-group-by-sub-experiment level.

TABLE B-12: *Stacked DID Design: Clean Controls and Compositional Balance*

	Conflict over power	
	(1)	(2)
More Distant × Post	0.129***	0.087**
	(0.031)	(0.038)
Mean dependent variable	0.065	0.065
Geographic controls		yes
Observations	1,674	1,674
Adjusted R-squared	0.594	0.600

Notes: The unit of observation is an ethnic group-country-year-sub-experiment. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. *More Distant* is an indicator that equals 1 if an ethnic group becomes more culturally distant from the government in a given sub-experiment, and 0 otherwise. *Post* is an indicator that equals 1 for time periods after a government change in each sub-experiment. The dataset includes 54 distinct sub-experiments. The sample includes 23 African countries and 150 distinct ethnic groups. The estimation is a weighted-least-squares stacked regression that follows the weighting procedure suggested by Wing et al. (2024). Standard errors clustered at the ethnic-group-by-sub-experiment level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

FIGURE B-8: *Stacked Event Study: Clean Controls and Compositional Balance*



Notes: The figures plot coefficient estimates of τ_{ke} described in equation 4. The left graph does not include geographic controls, while the graph on the right does. The estimation is a weighted-least-squares stacked regression that follows the weighting procedure suggested by Wing et al. (2024). Bold blue vertical bars and thin black vertical bars denote 90 and 95 confidence intervals respectively. Standard errors are clustered at the ethnic-group-by-sub-experiment level.

B-5 Estimation Using Partitioned Ethnicities

This section describes in detail the procedure I adopted to obtain the results in columns 1 and 2 of Table A-3. My main estimation strategy does not allow to account for unobserved time-varying ethnic-specific shocks. If these occurred simultaneously to a change in government and were correlated to an ethnic group's decision to rebel, they could confound my results. To address this, I re-run the analysis on a subset of ethnic groups that were split by country borders during the Scramble for Africa. This strategy, similar to the one employed by Michalopoulos and Papaioannou (2014) and analogous to the one adopted by Dickens (2018), exploits the fact that the same ethnicity is simultaneously exposed to different governments in different countries. As the ethnic identity of governments changes in some countries but not others, the quasi-random allocation of borders provides an exogenous source of within-group variation in cultural distance to the government.[24] As cultural distance of partitioned groups varies over time and between countries, this variation allows me to control for a full set of ethnicity-year fixed effects.

I merge contemporary ethnicities to the Murdock Atlas, which contains information on ethnic settlements prior to European contact and therefore provides the most reliable source to identify groups that were partitioned during the Scramble for Africa.[25] The map in Figure B-9 displays the Murdock counterparts of the EPR ethnic groups that were split across countries. This restricted sample includes 76 distinct Murdock ethnic groups partitioned across 33 African countries. Because the definition of ethnicity in the Murdock Map is finer than in the EPR database, a given EPR group can be matched to multiple Murdock groups. This keeps the resulting estimation sample still large (N=7,434), even though almost 60 percent of the original EPR groups included in my baseline analysis are dropped due to not having been split across countries.

With this data at hand, I estimate the following empirical model:

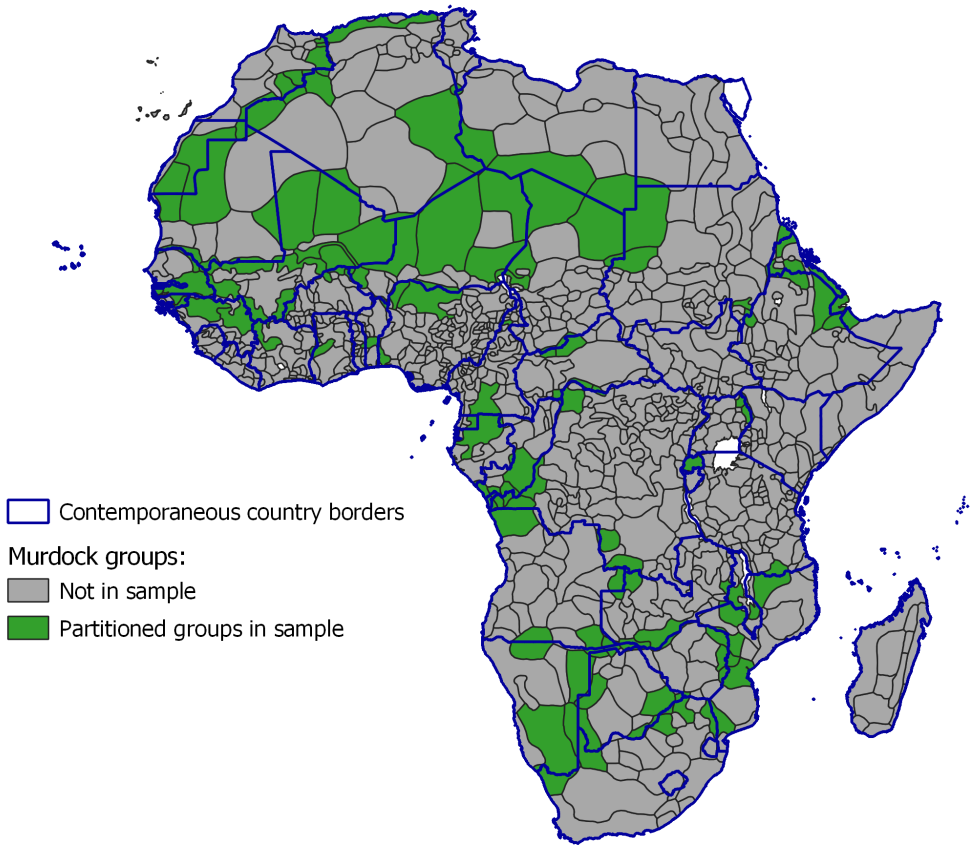
$$\text{Conflict}_{rct} = \eta_{cr} + \lambda_{ct} + \zeta_{rt} + \beta \text{LD}_{rct} + \Gamma \text{G}_{rct} + \epsilon_{rct} \tag{5}$$

The dependent variable is a measure of conflict for ethnic group r in country c and year t. The main independent variable, LD_{rct} is a measure of linguistic

[24] For additional evidence and details on the randomness in the allocation of colonial boundaries, see the discussion in Michalopoulos and Papaioannou (2016) and in Dickens (2018).

[25] This is also the data source employed by Michalopoulos and Papaioannou (2016), who were the first exploiting partitioned groups for identification.

FIGURE B-9: *Murdock Ethnic Groups Split by Country Borders*



Notes: The figure displays the ethnic groups in the Murdock Ethnographic Atlas that were split by country borders and that I include in the analysis.

distance to the government in period t. As in the main specification, I add a full set of country-year fixed effects (λ_{ct}), as well as a set of ethnicity-country fixed effects (η_{cr}). Moreover, since the same potential rebel is present in multiple countries, this specification controls for a full set of ethnicity-year fixed effects (ζ_{rt}), and thus accounts for any time-variant ethnic-specific shocks. G_{rct} indicates the same set of geographic controls outlined in Section 4. Standard errors are clustered at the ethnic-group-by-country level. Results are reported in Table A-3, columns 1 and 2 and in Table B-13 below, columns 1 and 4. Qualitatively, the results are similar to those in my baseline specification, but the magnitudes differ: a one standard deviation increase in linguistic distance increases conflict by 0.22 standard deviations, a smaller effect compared to the 0.4 standard deviations found in the baseline specification.

This difference may be due to the different nature of the estimation sample. Including ethnicity-year fixed effects has the disadvantage of forcing the sample to be restricted to a peculiar set of ethnicities, i.e., those that were split into multiple countries. This set of ethnic groups is selected, as partitioned African groups exhibit important differences in a large range of political economy outcomes compared to their non-split counterparts (Michalopoulos and Papaioannou, 2014, 2016). Given these selection-bias and external-validity concerns, I opt for including all ethnic groups in the main analysis, and interpret the results presented in this section as a robustness exercise in support of the baseline findings.

However, in light of the difference in magnitude, one may worry that the true effect is in fact smaller when holding ethnicity-year characteristics fixed. Because the only difference between my baseline specification and the specification that only includes partitioned ethnic groups comes from replacing ethnicity-by-country year trends with ethnicity-year fixed effects, I can test for the sensitivity of the estimates to this replacement in the sample of partitioned ethnicities, i.e., where both options are feasible. Reassuringly, as can be seen in Table B-13 by comparing column 1 to column 2 (and column 4 to column 5 for the unweighted linguistic distance measure), replacing ethnicity-year fixed effects with ethnicity-by-country time trends barely changes the magnitude and the significance of the estimate. Columns 3 and 6 further show that results are similar when including ethnicity year trends (without them being country specific). This leads me to conclude that the difference in magnitude when including ethnicity-year fixed effects is mostly due to the different nature of the estimation sample.

TABLE B-13: *Robustness to Substituting Ethnicity-year Fixed Effects with Ethnicity-year Trends*

	Dependent variable: Ethnic conflict over power					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance^{W}	0.043**	0.039**	0.030**			
	(0.021)	(0.015)	(0.012)			
Linguistic distance^{UW}				0.037**	0.036**	0.025**
				(0.017)	(0.014)	(0.011)
Country-year FE	yes	yes	yes	yes	yes	yes
Ethn.-country FE	yes	yes	yes	yes	yes	yes
Ethn.-year FE	yes			yes		
Ethn.-by-country year trend		yes			yes	
Ethn. year trend			yes			yes
Geographic controls	yes	yes	yes	yes	yes	yes
Observations	7,434	7,434	7,434	7,434	7,434	7,434
Adjusted R-squared	0.836	0.705	0.694	0.836	0.703	0.693

Notes: The unit of observation is an Murdock ethnic group in a country and year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. The sample includes 33 African countries and 166 distinct Murdock ethnic groups by country (or 76 distinct Murdock groups). Standard errors clustered at the ethnicity-country level are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-6 Francois et al.'s (2015a) Data on the Ethnic Composition of the Government

As a robustness test, I use Francois et al.'s (2015a) data on the ethnicity of African cabinet members to construct an alternative measure of linguistic distance. For 15 out of the 44 countries included in my estimation sample, Francois et al. (2015a) provide information on the ethnic identity of more than 90% of national ministers from the 1960s (i.e., independence) up to 2004. I manually match each cabinet minister's ethnicity to one or more Ethnologue languages (up to 8). Conveniently, Francois et al. (2015a) provide information on the number of top and lower government positions occupied by each ethnic group in a given country and year. This information allows me to generate a linguistic distance measure similar to the one I employ in my main specification and assign a higher weight to ethnicities occupying top government positions.[26]

I then re-run my baseline specification using this alternative measure. The sample (N=2,763) includes 15 countries (Benin, Cameroon, Cote d'Ivoire, Democratic Republic of Congo, Gabon, Ghana, Guinea, Liberia, Nigeria, Republic of Congo, Sierra Leone, Tanzania, Togo, Kenya, and Uganda) and 91 potential rebels (out of the 265 in my main sample). Reassuringly, my baseline linguistic distance measure and this alternative one are highly correlated (correlation coefficient: 0.75).

Table B-14 shows that results hold for this subsample of countries and years when employing my baseline linguistic distance measure (columns 1-3). Reassuringly, the linguistic distance measure based on Francois et al.'s (2015a) data (columns 4-6) leads to similar results. Coefficients are larger in columns 4 and 5; albeit not statistically significant, the coefficient on column 6 is still positive and large in magnitude. Despite the restricted coverage, these findings suggest that my main results are robust to employing a linguistic distance measure relying on an alternative data source on the ethnic composition of the executive.

[26] In a similar vein as for my EPR-based measure, when constructing the weights I assume that holding a top-government position is equivalent to holding two lower-government positions.

TABLE B-14: *Robustness using Francois et al.'s (2015a) Data on Ethnic Composition of the Government*

	Dependent variable: Ethnic conflict over power					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance$^{\text{Baseline}}$	0.047**	0.031***	0.050**			
	(0.018)	(0.004)	(0.018)			
Linguistic distance$^{\text{Francois et al.}}$				0.080**	0.084**	0.040
				(0.035)	(0.034)	(0.037)
Mean of dep. var.	0.037	0.037	0.037	0.037	0.037	0.037
Country-year fixed effects	yes	yes	yes	yes	yes	yes
Ethnicity fixed effects	yes	yes	yes	yes	yes	yes
Ethnicity year trend			yes			yes
Geographic controls		yes	yes		yes	yes
Observations	2,763	2,763	2,763	2,763	2,763	2,763
Adjusted R-squared	0.349	0.352	0.438	0.346	0.347	0.430

Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Linguistic distance*^{Baseline} captures linguistic distance between each potential rebel and the ethnic groups at the government based on the EPR dataset. *Linguistic distance*$^{Francois\ et\ al.}$ captures linguistic distance between each potential rebel and the ethnic groups at the government based on the Francois et al.'s (2015a) data. The linguistic distance measures are standardized. Geographic controls include the logged geodesic distance between each ethnic group and the government, absolute distance in elevation, absolute distance in ruggedness, and absolute distance in the caloric suitability index (CSI). The sample includes 91 distinct ethnic groups in 15 African countries over 44 years (1961-2004). Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-7 Instrument Construction, Validity, and Robustness

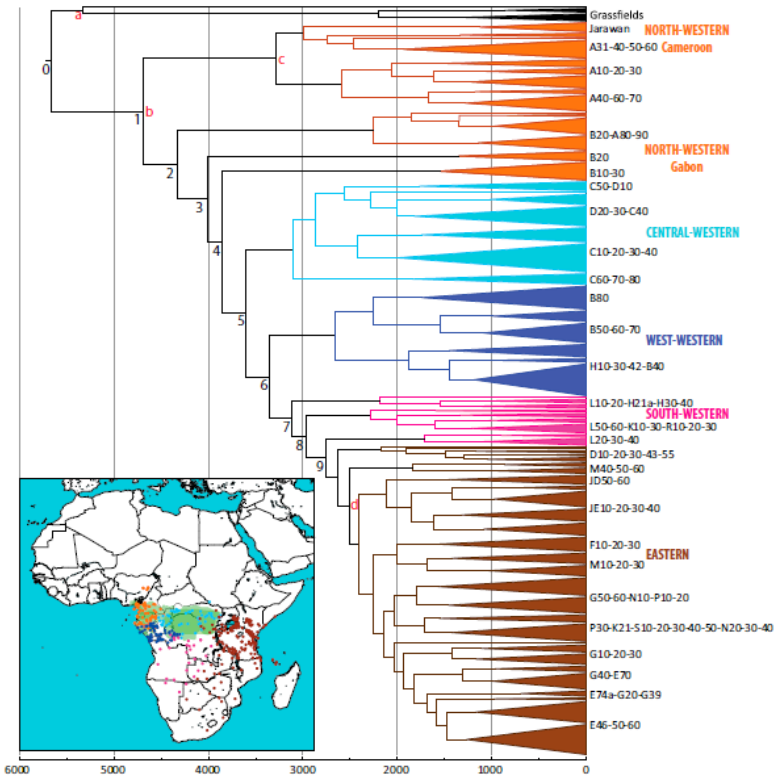
B-7.1 The Bantu-expansion Route

To build my instrument, I exploit the Bantu-expansion route reconstructed by Grollemund et al. (2015b). The authors start by constructing a phylogenetic tree that relates a sample of 424 Bantu languages, including now-extinct languages, one another. Each subsequent branch of the phylogenetic tree is then located on a timeline. The methodology adopted to construct this tree makes use of lexical data (i.e., a sample of 100 lexical items), and is based on a method that was originally developed to investigate the origin of virus outbreaks from molecular sequence data. The authors calibrate their model using archaeological evidence (e.g., node 1 is dated back to 4000-5000 before present in the Grassfields region of Cameroon, on the basis of an archaeological site called Shum Laka, the principal site associated with Bantu homelands). The resulting tree is reported in Figure B-10 below and is available in more detail in the Grollemund et al.'s (2015b) Supporting Information document (pg. 4, Figure S1).

To reconstruct the geographic location of the expansion, the authors collect data on the present-day geographic location of Bantu languages and consider them to be the destination (or arrival) points of the Bantu expansion route. The assumption is that these points are unaffected by recent events and indeed reflect the end-destination of the Bantu expansion. This assumption seems to be accepted by other scholars in the literature. For instance, Koile et al. (2022), who make use of Grollemund et al.'s (2015b) data, explicitly state this assumption in their study: "Latitude and longitude data on the current location of the languages studied were taken from ref. 3 [Grollemund et al. (2015)] [...]. We assume that these locations have not been substantially modified by the impact of recent events such as the Atlantic slave trade and colonialism" (Koile et al., 2022, p.6).

Based on the phylogenetic tree in combination with the assumed end-destination points (i.e., the contemporary locations of Bantu languages), the authors reconstruct the probable ancestral geographical locations of each of the internal nodes of the phylogenetic tree using a Brownian motion model. Similar methodologies have been adopted by other influential studies to infer ancestral longitudes and latitudes in a phylogenetic context: a prominent example is Bouckaert et al.'s (2012) study on the expansion of the Indo-European language family. Grollemund et al. (2015b) infer the ancestral latitude and longitude for each internal node of

FIGURE B-10: *Phylogenetic Tree of Bantu Languages Reconstructed by Grollemund et al. (2015)*



Notes: Source: Grollemund et al. (2015b), Figure 1.

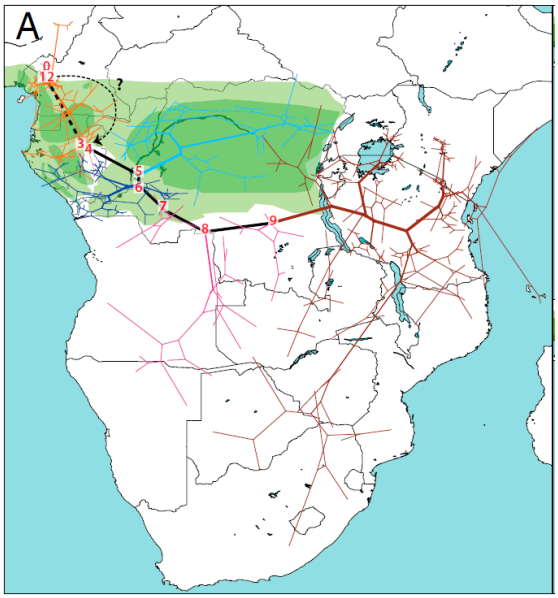
the tree, which they connect using straight lines.[27] The result of this procedure is depicted in Figure B-11, which illustrates the Bantu-expansion path I digitize and use to construct my instrument.

B-7.2 Instrument Construction

To build the Bantu instrument, I start by constructing a measure of exposure to the Bantu expansion for each ethnic group's homeland. I merge contemporary ethnicities to their ancestral homelands in the Murdock map, the most ancient available record of ethnic settlements in Africa. I then overlay ethnic groups' ancestral homelands on the historical migration route of the Bantu reconstructed

[27]Additional methodological details can be found in Grollemund et al. (2015b) and in the article's Supplementary Appendix.

FIGURE B-11: *Ancestral Bantu Migration Route Reconstruction*



Notes: Source: Grollemund et al. (2015b), Figure 2.

by Grollemund et al. (2015b).[28] The left panel of Figure A-5 illustrates the route of the Bantu migration and contemporaneous country borders. For each ethnic group in the Murdock map, I construct a Bantu Index capturing the intensity of exposure to the Bantu migration route. I calculate the length of the path crossing each ethnic homeland, divide it by the size of the settlement, and then normalize this index to range between 0 and 1:

$$\text{Bantu Index}_{EA} = \frac{\text{Path Length}}{\text{Homeland Area}} \tag{6}$$

Figure A-5 (right) shows the distribution of the Bantu index across Africa. There is considerable within-country, cross-ethnicity variation in ethnic groups' exposure to the Bantu expansion. Since each EPR group is associated to multiple groups in the Ethnographic Atlas, I proceed as follows to obtain an index at the EPR-group level:

$$\text{Bantu Index}_{EPR} = \sum_{i=1}^{N} \alpha_i \text{Bantu Index}_{EA_i} \tag{7}$$

[28] See Appendix B for details on how Grollemund et al. (2015b) reconstructed the route.

where α_i is a weight reflecting the accuracy of the match between each Murdock group and the EPR group.[29]

Next, I construct the instrument, i.e., the absolute distance in the Bantu Index between a potential rebel and the government. To this end, I follow the same procedure I adopted for the linguistic distance measure. First, I compute the Bantu distance between a potential rebel (r) and each ethnic group forming the government (g_i):

$$\text{BD}_{rg_i} = |\text{Bantu Index}_r - \text{Bantu Index}_{g_i}| \tag{8}$$

Second, I compute a weighted Bantu distance between each potential rebel and the government:

$$BD^W = \sum_{i=1}^{N} p_{g_i} \times BD_{rg_i} \tag{9}$$

where N denotes the total number of ethnicities forming the government and p_{g_i} is a weight reflecting the position of power of group g_i in the coalition. Alternatively, I construct the unweighted version:

$$BD^{UW} = \sum_{i=1}^{N} \frac{BD_{rg_i}}{N} \tag{10}$$

B-7.3 Instrument Validity

The validity of the IV estimates rests on the assumption that the differential exposure to the Bantu expansion affects conflict today only through its impact on cultural distance, conditional on the controls included in the regression. The first concern is that geographical differences between ancestral homelands might have determined the path of the Bantu migration and, at the same time, might also affect the likelihood of conflict between groups today. Grollemund et al. (2015b) emphasize how the route of the expansion was mostly determined by emerging savannah corridors, which appeared in the rainforest due to a climatic shock and then disappeared once the climatic crisis was over. Despite this suggestive evidence of exogeneity in the way the path evolved, my preferred specification addresses this

[29]The accuracy is given by the percentage of nodes in the EPR-group linguistic tree covered by the Murdock-group linguistic tree. $\sum_{i=1}^{N} \alpha_i = 1$. If the match is equally accurate between each Murdock group, then Bantu $\text{Index}_{EPR} = \sum_{i=1}^{N} \frac{\text{Bantu Index}_{EA_i}}{N}$. See Appendix Section C-2 for details on the LEDA package, the merging procedure, and the accuracy of the match.

concern by controlling for a set of geographic controls.

Another concern could be that the Bantu, in gradually expanding, intentionally and systematically avoided pre-existing tribes with certain characteristics that make them more (or less) prone to conflict today. Suppose the Bantu avoided particularly bellicose tribes, and favored territories inhabited by more peaceful populations. Then my instrument would be systematically picking up differences in bellicosity. If these differences have a direct effect on conflict today, the exclusion restriction is violated.

I examine whether this example is at play in my setting by delving deeper in the reduced-form analysis. If the Bantu systematically avoided certain territories along unobservable characteristics correlated with contemporary conflict, then only being more (or less) affected to the Bantu expansion should have a reduced-form effect on conflict, but not both. I test this by splitting the absolute distance in the Bantu index into two: being *more* exposed to the Bantu expansion than the ethnicities at the government, and being *less* exposed.

I run an alternative reduced-form estimation using these two measures, instead of the absolute difference as in my main specification. As shown in Table B-15, I do not find evidence of a different reduced-form effect of being more versus less exposed to the Bantu expansion. Both coefficients hold the same sign and are of similar size, and the F statistic indicates that these coefficients are not different from each other. This suggests that what matters for conflict between ethnic groups is their *absolute* distance in the Bantu index—as conjectured in Section 6.1—and not the fact of being more (or less) affected by the Bantu migration compared to the other group.

TABLE B-15: *Reduced Form: Splitting the Bantu Distance into Two*

	Ethnic conflict over power	
	(1)	(2)
Bantu more	0.053	0.047
	(0.039)	(0.033)
Bantu less	0.054*	0.053*
	(0.031)	(0.030)
F statistic (Bantu more - Bantu less=0)	0.00	0.01
Prob > F	0.980	0.905
Country-year fixed effects	yes	yes
Ethnicity fixed effects	yes	yes
Ethnicity year trends	yes	yes
Geographic controls		yes
Observations	4,027	4,027
Adjusted R-squared	0.502	0.502

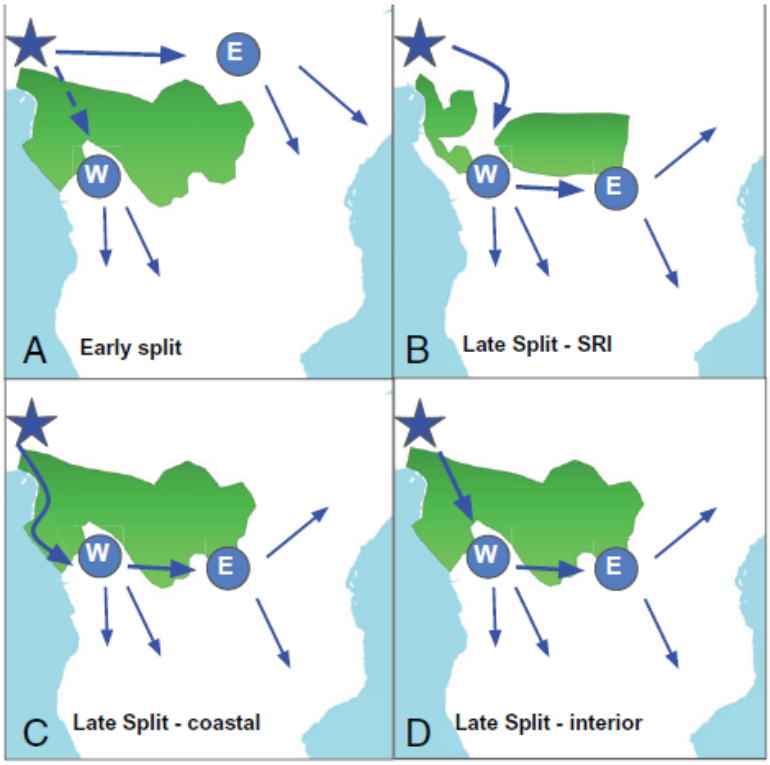
Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government and 0 otherwise. *Bantu more* and *Bantu less* is a continuous variable that denotes whether the rebel was more or less exposed to the Bantu expansion than the government, respectively. Standard errors clustered by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-7.4 Robustness Tests in Light of Debates on the Bantu Expansion Route

The exact location of the Bantu expansion route remains debated in the literature. An important topic in the debate pertains the so-called early- versus late-split hypotheses (Koile et al., 2022). Figure B-12, taken from Koile et al. (2022), provides a stylized illustration of these competing hypotheses. According to the early-split hypothesis (panel A), East Bantu people and West Bantu people, as observed today, are the result of an early split that occurred when Bantu people left their homeland in West Africa. While West Bantu speakers directly entered the rainforest, the East Bantu branch moved eastward, north of the Congo rainforest toward the African Great Lakes region and then headed south afterwards. Conversely, the

late-split hypothesis pertains that the split between East and West Bantu peoples occurred only after the original Bantu tribe successfully traversed the rainforest. Even among scholars in support of the late-split hypothesis, there is further debate as of whether the migratory route occurred through a coastal route (panel C), a savannah corridor called the "Sangha River Interval" (panel B), or through the rainforest but before the Sangha River Interval corridor opened up.

FIGURE B-12: *Illustration of the Early vs Late Split Hypotheses*



Notes: Stylized illustration of the competing hypotheses around the Bantu expansion. Source: Koile et al. (2022).

Reassuringly for the purpose of this paper, the early-split hypothesis advanced by Vansina (1995) and Holden (2002) (depicted in panel A of Figure B-12) has been systematically rejected by recent studies based using both genetic and lexical data (De Filippo et al., 2012; Currie et al., 2013; Grollemund et al., 2015b; Whiteley et al., 2019; Koile et al., 2022). However, while Grollemund et al.'s (2015b) study finds support for the "Shangha River Interval" hypothesis, Koile et al.'s (2022) findings are consistent with a rainforest route taken by the Bantu tribe way before the opening of the Shanga river corridor occurred. Since these two explanations

only differ in what they hypothesize with regards to the *timing* of the expansion, the predicted geographical location of the path itself is quite similar.[30] This is reassuring, and suggests that a hypothetical analysis based on a path consistent with Koile et al.'s (2022) hypothesis would yield similar results. Unfortunately, this direct test is not feasible, because Koile et al. (2022) only provide a stylized geographical reconstruction of the path[31] given that the aim of their analysis is to shed light on the timing of the expansion.

To nevertheless gauge whether my IV findings are sensitive to the three most debated hypotheses—late split coastal, and late split interior, and late split Sangha River Interval—I re-run the IV specification dropping the countries in whose territories the exact location of the path is most uncertain: Cameroon, Gabon, Equatorial Guinea, and Congo. Table B-16 shows that dropping these countries does not alter the IV results.

[30]In fact, Koile et al. (2022) write: "Previous geographic reconstructions of the Bantu expansion (Grollemund et al. 2015; Currie 2013; Whiteley et al. 2019) consistently find a migration route in the interior of the rainforest, far from the proposed coastal route. Although these studies tend to interpret this result as supporting the late split through the Sangha River Interval hypothesis (this is not the case for Whiteley et al. (2019), which suggests rivers and valleys as natural corridors of migration), this interpretation yields an inconsistent dating of the migration events (Grollemund et al. 2015; Ehret 2015)."

[31]See Figure 3 in Koile et al. (2022).

TABLE B-16: *Robustness to Dropping Countries where the Path is Debated*

	Dependent variable: Ethnic conflict over power				
	(1)	(2)	(3)	(4)	(5)
Linguistic distance	0.177***	0.220***	0.203***	0.204***	0.196***
	(0.032)	(0.040)	(0.036)	(0.036)	(0.028)
Mean of dep. var.	0.050	0.048	0.048	0.050	0.057
	First stage				
	Linguistic distance				
Bantu distance	0.343***	0.332***	0.330***	0.337***	0.346***
	(0.078)	(0.088)	(0.081)	(0.083)	(0.085)
Kleibergen-Paap F-statistic	14.48	10.98	12.74	12.72	12.86
Country dropped:	Cameroon	Congo	Equatorial Guinea	Gabon	All
Observations	3,687	3,846	3,927	3,806	3,185

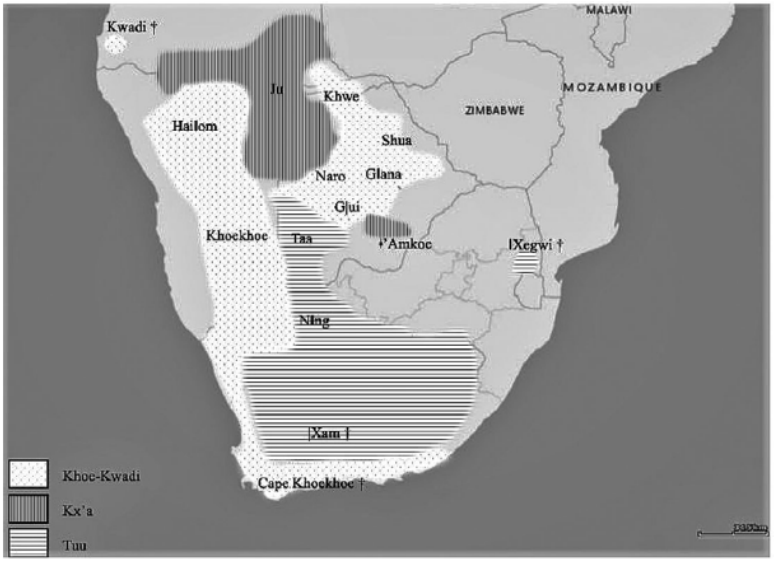
Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. Each column drops one country from the sample where the precise path of the Bantu expansion is debated. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-7.5 Robustness Tests in Light of Debates on the Similarity of Pre-Bantu Populations

One of the underlying assumptions of my instrument is that, *within contemporary countries*, pre-Bantu groups were on average culturally similar to each other, and culturally dissimilar to the Bantu. This is partly supported by archaeological evidence showing that pre-Bantu societies shared the same mode of subsistence, i.e., hunting and gathering, while Bantu people were predominantly farmers (Diamond, 1997). However, the cultural landscape of Central, Southern, and Southeast Africa before the Bantu expansion remains largely unknown. Linguistic and genetic evidence suggest that Central Africa was inhabited by Pygmy foragers, Southern Africa by Khoisan-speaking hunter-gatherers, and Southeast Africa by Cushitic-speaking pastoralists and Nilo-Saharan-speaking herders (Bostoen and Gunnink, 2022). Therefore, the fact that pre-Bantu populations shared the same linguistic family within these macro-regions (with the exception of Southeast Africa, where both Cushitic and Nilo-Saharan languages coexist among non-Bantu speakers), also suggests that the assumption is plausible.

To directly test the assumption behind my instrument, one would want to consult an ethnolinguistic map of pre-Bantu Africa and assess whether groups were culturally comparable before the expansion within nowadays' country borders. Unfortunately, to the best of my knowledge, such information is not available. In the absence of systematic data on pre-Bantu populations, the contemporary geographical distribution of non-Bantu linguistic families can help visualize the non-Bantu cultural landscape within today's countries.[32] Figures B-13, B-14, B-15, and B-16 below show the location of the Khoisan linguistic family (and the respective three subfamilies Khoe-Kwadi, Tuu, and Kx'a), the Ubangi and Sudanic languages spoken by Pygmies, the Cushitic languages, and the Nilo-Saharan languages, respectively. The maps indicate that, by and large, controlling for country fixed effects in the first stage ensures comparability (in terms of linguistic-family affiliation) between groups that were not affected by the Bantu expansion. Among the countries included in my IV sample, only Kenya and Tanzania display multiple non-Bantu linguistic families (i.e., both Cushitic and Nilotic languages, see Figure B-15 and B-16).

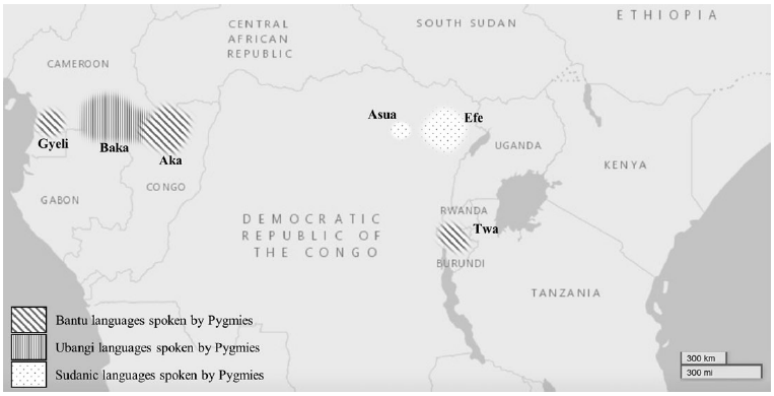
FIGURE B-13: *Distribution of the Khoisian Languages in Southern Africa*



Notes: Source: Bostoen and Gunnink (2022)

[32]This is obviously only an approximation of how the linguistic landscape of pre-Bantu expansion Africa might have looked like.

FIGURE B-14: *Distribution of the Languages Spoken by Pygmies in Central Africa*



Notes: Source: Bostoen and Gunnink (2022)

FIGURE B-15: *Distribution of Cushitic Languages*

Notes: Source: https://en.wikipedia.org/wiki/Cushitic_languages

FIGURE B-16: *Distribution of Nilotic Languages*

Notes: Source: https://en.wikipedia.org/wiki/Nilotic_languages#/media/File:Nilotic_languages.png

Based on this, a first robustness test consists of re-running the IV analysis without Kenya and Tanzania, where pre-Bantu populations might have been dissimilar, as suggested by the maps above. Table B-17 below indicates that the IV results hold for the sample of countries in which my assumption about the cultural similarity among pre-Bantu groups is more likely to be satisfied.

TABLE B-17: *Robustness to Dropping Countries where pre-Bantu Populations may have been Dissimilar (I)*

	Conflict over power
Linguistic distance	0.202***
	(0.036)
	First stage
	Linguistic distance
Bantu Distance	0.290***
	(0.030)
Observations	3,555
Kleibergen-Paap F-statistic	73.29

Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. The sample excludes Kenya and Tanzania. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

Among linguists, there is also some debate with regards to whether Khoisan languages should be considered as a single linguistic family. A common feature of the languages belonging to the three subfamilies Khoe-Kwadi, Tuu, and Kx'a are the presence of phonemic clicks. However, linguists tend to consider Khoe-Kwadi-speaking populations as culturally distinct (Fehn et al., 2022). Moreover, recent genetic studies reveal diversity among pre-Bantu populations in the Namibian desert, part of which can be attributed to the Kwadi branch of Khoe-Kwadi linguistic family (Oliveira et al., 2023). As shown in Figure B-13, Khoe-Kwadi languages coexists with Tuu and Kx'a languages in Namibia, South Africa, Botswana, and Angola. If the Khoe-Kwadi speakers were more different from Tuu and Kx'a speakers than to the Bantu, the assumption behind my instrument would not be satisfied in these countries. I re-run the analysis dropping these countries from the sample, and find that the main results hold (Table B-18).

TABLE B-18: *Robustness to Dropping Countries where pre-Bantu Populations may have been Dissimilar (II)*

	Conflict over power
Linguistic distance	0.203***
	(0.037)
	First stage
	Linguistic distance
Bantu Distance	0.366***
	(0.090)
Observations	3,061
Kleibergen-Paap F-statistic	12.33

Notes: the unit of observation is an ethnic group-country-year. The dependent variable is a binary variable that takes value 1 if an ethnic group is fighting the government for gaining power. The sample excludes Namibia, South Africa, Botswana, and Angola. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

Finally, I explore the plausibility of the assumption behind my instrument by running the first stage on the full population of Murdock ethnic groups, and subsequently dropping ethnic-group pairs unaffected by the Bantu migration. I construct a dataset considering all the possible within-country ethnic-homelands pairs (ij) in the Ethnographic Atlas as unit of observation (see Figure 2 in the paper), and run a gravity specification including side i and j fixed effects. If pre-Bantu populations were systematically more dissimilar one another than each to the Bantu, removing ethnic-group pairs not exposed to the Bantu expansion (with a "Bantu Distance" equal to zero) should increase the magnitude and the significance of the coefficient on "Bantu Distance." Instead, I find the opposite: Table B-19 shows that both the magnitude and the significance of the coefficient on "Bantu Distance" decrease, which suggests that ethnic-group pairs not exposed to the Bantu expansion contribute significantly to the strength and relevance of the instrument.

TABLE B-19: *First Stage Dropping Pairs not Exposed to the Bantu Expansion: Murdock Atlas Sample*

	Linguistic distance	
	(1)	(2)
Bantu distance	0.179***	0.062**
	(0.043)	(0.031)
Sample:	full	restricted
Group i-by-country fixed effects	yes	yes
Group j-by-country fixed effects	yes	yes
Observations	22,124	19,716
Adjusted R-squared	0.745	0.818

Notes: the unit of observation is a pair of Murdock ethnic groups (group i and group j) in a country. The dependent variable is linguistic distance between two ethnic groups in a pair. *Bantu distance* denotes the absolute distance in the exposure to the Bantu expansion between two ethnic groups in a pair. Column 2 removes pairs with a *Bantu distance* equal to zero. Standard errors clustered two-way by ethnicity i-by-country and ethnicity j-by-country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

B-8 Ancestral Arrangements, Distance in Preferences over Public Goods, and Conflict over Power

This section proposes a preliminary investigation aimed at decomposing culture into various sub-domains and assessing which component of cultural distance is most powerful in explaining conflict over government power. The theoretical literature on diversity and conflict described in Section 2 does not take a clear stance on which set of cultural traits we should expect to matter more in generating (or alleviating) conflict.[33] However, these models emphasize that diversity triggers conflict because it generates heterogeneity of preferences over public goods. Based on this work, therefore, one would want to isolate the component of cultural distance that is responsible for shaping a group's preferences over public goods. To the best of my knowledge, there is no clear evidence on which specific cultural component shapes such preferences.[34]

To nevertheless try to gauge the role of different cultural components in explaining ethnic conflict, I leverage a common approach used in the literature to measure culture, which relies on deep-rooted, pre-industrial arrangements that have persistently shaped norms, values, and preferences. This approach builds on literature highlighting the persistence of cultural traits (Voigtländer and Voth, 2012) and the importance of historical factors for the evolution of culture (Guiso et al., 2008; Becker et al., 2016; Nunn and Wantchekon, 2011b; Grosjean, 2011; Alesina et al., 2013; Enke, 2019; Le Rossignol and Lowes, 2022; Cao et al., 2021; Guarnieri and Tur-Prats, 2023).

I use information from the Ethnographic Atlas and group ethnic characteristics into mutually-exclusive cultural domains. While this approach has the advantage of opening the "black box" and allowing the researcher to decompose culture, it also comes with important challenges. To my knowledge, there is no established literature discussing which ancestral characteristics might have shaped preferences over public goods, i.e., the cultural component that theory suggests I should aim

[33] This literature takes a *genealogical* approach, at the core of which is the idea that all populations have a shared common ancestor, but as different groups have split from each other, they have diverged in their set of "traits that are transmitted from one generation to another, including language, values, and norms" (Spolaore and Wacziarg, 2017, p. 9). My choice of using linguistic distance as a measure of cultural distance in the main analysis is motivated by this literature.

[34] These challenges and data limitations are well explained in Esteban et al. (2012), who also opt for using linguistic distance in their cross-country empirical analysis linking polarization to conflict, and cite a list of other empirical studies employing the same measure (Greenberg, 1956; Fearon and Laitin, 2000; Fearon, 2003; Desmet et al., 2009).

at isolating. More generally, there is no consensus on how to decompose culture. I therefore consider all the variables from the Ethnographic Atlas[35] and opt for the off-the-shelf classification provided by the D-PLACE database (Kirby et al., 2016) to group these variables into the following mutually-exclusive domains: the economy, politics, kinship, marriage, settlement patterns, and rituals.[36] Tables B-20, B-21 and B-22 display which traits fall within each domain. I group all the variables belonging to each domain into different indices using PCA. I take the predicted score of the first principal component and normalize it to range between 0 and 1.

The idea behind this exercise is that each of these domains might have shaped different cultural traits in the long-run: for instance, Enke (2019) shows that different forms of kinship have generated different moral systems; Grosjean (2014) and Cao et al. (2021) show that the "culture of honor" is tied to traditional herding practices; Le Rossignol and Lowes (2022) show that societies relying on transhumant pastoralism display more limited moral universalism. Here, I do not take a stance on which specific cultural trait has been shaped by each of the domains I construct. Instead, I let the data tell us which of these domains (i) most correlates with distance in preferences over public goods and (ii) is most predictive of conflict. If the asset of culture that triggers conflict is indeed distance in preferences over public goods, we should expect the findings in point (i) to be consistent with those in point (ii).

I start by assessing whether the proposed cultural domains correlate with linguistic distance in my dataset. Table B-23 shows the results after regressing linguistic distance on each cultural domain, controlling for country-year fixed effects and ethnicity fixed effects. As can be seen, linguistic distance is highly correlated with distance in each cultural domain, which confirms that linguistic distance is a good "summary" measure of cultural distance. Yet, when repeating the same exercise using my survey-based measure of distance in preferences over public goods instead of linguistic distance as dependent variable, it seems that the strongest and most significant correlation is the one with the *political* domain of culture (see Table B-24). Finally, to assess which of these distances matter the most in explaining conflict, I re-run my baseline specification replacing linguistic distance

[35]I only include the 51 traits that do not have missing values for more than 20 percent of the ethnic groups.

[36]To achieve mutually-exclusive domains, I do not consider other D-PLACE categories that are nested in the just-listed ones. See https://d-place.org/contributions/EA for the original classification. I acknowledge that there may be alternative valid ways of grouping ethnic characteristics into mutually-exclusive cultural components.

with each component of culture as explanatory variable. Results are displayed in Table B-25. I find that, while nearly all the distances are positively correlated with conflict, the largest and most significant coefficient is the one on the distance in the political domain.

While this analysis further supports the mechanism put forward in the paper, more research is needed on the deep-rooted determinants of preferences over public goods and how ancestral arrangements pertaining the political domain and community organization might have shaped these preferences in the African continent. These results can be linked to literature exploring the determinants of preferences over redistribution (summarized in Alesina and Giuliano (2011)), which can be considered to be an important aspect of preferences over the type of public good to be provided. Across contexts, this literature has linked individual preferences to historical events such as the Great Depression (Giuliano and Spilimbergo, 2009); other cultural traits such as individualism (Alesina and Glaeser, 2004); indoctrination in certain political regimes (Alesina and Fuchs-Schündeln, 2007); the intergenerational transmission of views about inequality and social mobility (Benabou and Tirole, 2006); the organization of the family and an individual's independence (Todd and Garrioch, 1985; Esping-Andersen, 1999; Alesina and Giuliano, 2010); perception of fairness (Alesina and Glaeser, 2004; Alesina and Angeletos, 2005); or the desire to obtain social standing (Corneo and Grüner, 2000, 2002). Arguably, the variables in the Atlas belonging to the political domain listed in Table B-22 may be linked to some of these determinants: the number of jurisdictional levels in the local community (v32), ranging from independent nuclear families to extended clan-barrios, may go hand in hand with the organization of the family, individuals' independence and thus their views about the role of government; ancestral caste and class differentiation (v66, v67, v68, v69), or the ancestral presence of slave status (v70, v71) may be deep-rooted factors influencing perceptions of fairness, the desire to obtain social standing, or views about inequality and social mobility; finally, political complexity in ancestral times (v33) may be directly shaping views about the role of government today.

Because the classification of ancestral characteristics I propose in this analysis is somewhat arbitrary, given that the D-PLACE database does not provide a theoretical justification for this grouping, I invite the reader to not over-interpret these results. Yet, this explorative analysis may serve as a starting point for future investigations on which component of culture is more likely to explain conflict, as well as how pre-industrial arrangements may have shaped political preferences in the long run.

TABLE B-20: *Grouping of Murdock Atlas Variables: Economy*

Economy/Subsistence/Property/Wealth Transactions	
v1	Dependence on the gathering of wild plants and small land fauna, relative to other subsistence activities.
v2	Dependence on hunting, including trapping and fowling, relative to other subsistence activities.
v3	Dependence on fishing, including shellfishing and the pursuit of large aquatic animals, relative to other subsistence activities.
v4	Dependence on animal husbandry, relative to other subsistence activities.
v5	Dependence on agriculture, relative to other subsistence activities.
v28	Intensity of cultivation.
v29	Principal type of crop cultivated.
v39	Indicates whether or not animals are employed in plow cultivation, and whether plow cultivation is aboriginal or dates to the post-contact period.
v40	The predominant type of animals kept.
v41	Indicates whether or not domestic animals milked.
v42	Dominant mode of subsistence.
v74	The rule or practice governing the disposition or transmission of a man's property in land (exclusive of any dower right of his widow).
v75	Following the inheritance rule for real property (land), this variable indicates how real property is distributed among several individuals of the same category.
v76	The rule or practice governing the disposition or transmission of a man's movable property (exclusive of any dower right of his widow).
v77	Following the inheritance rule for movable property, this variable indicates how movable property is distributed among several individuals of the same category.

TABLE B-21: *Grouping of Murdock Atlas Variables: Marriage and Kinship*

Marriage	
v6	Prevailing type of transfer or exchange at marriage.
v7	Alternative types of transfer or exchange at marriage, or supplementary practices.
v8	The prevailing form of domestic or familial organization.
v9	Marital composition of family units, regardless of whether families live as independent nuclear units or as part of larger extended families.
v10	Pattern of marital residence in the first years or so of marriage.
v11	The prevailing pattern of transfer of residence at marriage.
v12	The prevailing profile of marital residence in the society.
v13	Culturally patterned alternatives to, or numerically significant deviations from, the prevailing pattern of transfer of residence at marriage.
v14	Culturally patterned alternatives to, or numerically significant deviations from, the prevailing profile of marital residence.
v23	The rules or practices governing the marriageability or non-marriageability of a man's first cousins and, in some cases, second cousins.
v24	Number of types of cousins (first and/or second cousins) a man is permitted to marry.
v25	Preferred rather than just permitted marriages to particular first (and/or second) cousins.
v26	Specific type(s) of cousin preferred as spouses.
Kinship	
v15	The prevalence of local endogamy, agamy, and exogamy, together with the presence or absence of localized kin groups.
v16	Indicates whether or not societies organized as clan-communities (i.e., where communities represent a single, localized, exogamous kin group or clan) are also segmented into clan barrios.
v17	The largest type of patrilineal kin group reported for the society.
v18	The largest patrilineal kin group characterized by exogamy.
v19	The largest type of matrilineal kin group reported for the society.
v20	The largest matrilineal kin group characterized by exogamy.
v21	The presence or probable absence, and the typology, of cognatic kin groups.
v22	The presence or probable absence, and the typology, of secondary cognatic kin groups.
v43	Major mode of descent.

TABLE B-22: *Grouping of Murdock Atlas Variables: Politics, Rituals, Settlement Patterns*

Politics/Community Organization	
v32	The number of jurisdictional levels in the local community, with 2 representing the theoretical minimum (e.g., independent nuclear or polygynous families) and 4 representing the theoretical maximum (e.g., nuclear families nested within extended families and clan-barrios).
v33	The number of jurisdictional levels beyond the local community, with 1 representing the theoretical minimum (e.g., none/autonomous bands or villages) and 4 representing the theoretical maximum (e.g., villages nested within parishes, districts, provinces, and a complex state). This variable also provides a measure of political complexity, ranging from 1 for stateless societies, through 2 or 3 for petty and larger paramount chiefdoms or their equivalent, to 4 or 5 for large states.
v66	The degree and type of class differentiation, excluding purely political and religious statuses.
v67	Secondary type of class differentiation present in the society, excluding purely political and religious statuses.
v68	The degree and type of caste differentiation.
v69	Secondary type of caste differentiation.
v70	The forms and prevalence of slave status, treated quite independently of both class and caste status.
v71	Indicates whether slavery present at the time of observation of the society, or only in the past.
Rituals	
v37	The practice of circumcision or any of its variants, when culturally normative and not sporadic or merely optional, as well as the age at which it typically occurs.
Settlement Patterns	
v30	The prevailing type of settlement pattern.
v79	Ground plan of the prevailing type of dwelling.
v80	Floor level of the prevailing type of dwelling.
v82	Shape of roof of the prevailing type of dwelling.
v83	Roofing materials of the prevailing type of dwelling.

TABLE B-23: *Cultural Components in the Ethnographic Atlas and Linguistic Distance*

	Dep var: Linguistic distance					
Distance in the following domain:	(1)	(2)	(3)	(4)	(5)	(6)
Economy	0.143***					
	(0.028)					
Marriage		0.068***				
		(0.016)				
Politics			0.113***			
			(0.035)			
Kinship				0.056***		
				(0.012)		
Settlement patterns					0.123***	
					(0.029)	
Rituals						0.076***
						(0.016)
Mean of dep. var.	0.412	0.411	0.418	0.403	0.412	0.414
Country-year fixed effects	yes	yes	yes	yes	yes	yes
Ethnicity-country fixed effects	yes	yes	yes	yes	yes	yes
Observations	7,858	7,924	7,171	7,883	7,973	8,006
Adjusted R-squared	0.928	0.923	0.942	0.918	0.925	0.929

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is linguistic distance between each group and the government and ranges between 0 and 1. *Economy*, *Marriage*, *Politics*, *Kinship*, *Settlement patterns*, and *Rituals* capture distances in various cultural domains based on the Murdock Ethnographic Atlas (see Tables B-20, B-21, and B-21) and are standardized. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE B-24: *Cultural Components in the Ethnographic Atlas and Distance in Preferences over Public Goods*

	Dep var: Distance in preferences over public goods					
Distance in the following domain:	(1)	(2)	(3)	(4)	(5)	(6)
Economy	0.016					
	(0.012)					
Marriage		0.000				
		(0.008)				
Politics			0.037***			
			(0.010)			
Kinship				-0.000		
				(0.007)		
Settlement patterns					0.031*	
					(0.017)	
Rituals						0.006
						(0.006)
Mean of dep. var.	0.079	0.078	0.075	0.080	0.079	0.080
Country-year fixed effects	yes	yes	yes	yes	yes	yes
Ethnicity-country fixed effects	yes	yes	yes	yes	yes	yes
Observations	4,678	4,808	4,316	4,722	4,943	4,886
Adjusted R-squared	0.883	0.882	0.925	0.880	0.888	0.882

Notes: The unit of observation is an ethnic group-country-year. The dependent variable (ranging between 0 and 0.46) is the distance in preferences over public goods, an index constructed at the ethnic-group level based on the Afrobarometer survey. *Economy*, *Marriage*, *Politics*, *Kinship*, *Settlement patterns*, and *Rituals* capture distances in various cultural domains based on the Murdock Ethnographic Atlas (see Tables B-20, B-21, and B-21) and are standardized. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

TABLE B-25: *Cultural Components in the Ethnographic Atlas and Conflict Over Government Power*

	Dep var: Conflict over power					
Distance in the following domain:	(1)	(2)	(3)	(4)	(5)	(6)
Economy	0.024					
	(0.030)					
Marriage		0.018				
		(0.027)				
Politics			0.144***			
			(0.048)			
Kinship				0.032		
				(0.034)		
Settlement patterns					-0.040	
					(0.049)	
Rituals						0.028
						(0.047)
Mean of dep. var.	0.056	0.055	0.057	0.057	0.053	0.055
Country-year FE	yes	yes	yes	yes	yes	yes
Ethnicity-country FE & trends	yes	yes	yes	yes	yes	yes
Geographic controls	yes	yes	yes	yes	yes	yes
Observations	7,858	7,924	7,171	7,883	7,973	8,006
Adjusted R-squared	0.534	0.534	0.537	0.530	0.525	0.544

Notes: The unit of observation is an ethnic group-country-year. The dependent variable is an indicator that equals 1 if a group rebels over government power, and zero otherwise. *Economy*, *Marriage*, *Politics*, *Kinship*, *Settlement patterns*, and *Rituals* capture distances in various cultural domains based on the Murdock Ethnographic Atlas (see Tables B-20, B-21, and B-21) and are standardized. Two-way clustered standard errors by year and country are reported in parenthesis. *** (**) (*) indicate significance at the 1% (5%) (10%) level.

Appendix C: Data Description, Sources, and Summary Statistics

C-1 Data Sources

Ethnic groups: Data on ethnic groups comes from the Ethnic Power Relations (EPR) dataset family (Wucherpfennig et al., 2012; Vogt et al., 2015). EPR lists each politically relevant ethnic group for each country in each year, and the respective access (or lack thereof) to executive government power. EPR defines an ethnic group as an identity group that defines itself or is defined by others along linguistic, religious or racial characteristics. I download the EPR group-level data from the GrowUp platform (Girardin et al., 2015).
Source: https://growup.ethz.ch/rfe

Linguistic distance: I retrieve data on languages (up to 3) spoken by each ethnic group from the Ethnic Dimensions Dataset (Bormann et al., 2017; Ethnic Power Relations Dataset Family, n.d.). I compute a measure of linguistic distance (called cladistic distance) using information on linguistic trees coming from the Ethnologue database (Lewis et al., 2020). I describe the methodology for the construction of the linguistic distance measure in Section 3.
Sources: https://icr.ethz.ch/data/epr/ed/ and https://www.ethnologue.com/browse/names

Conflict: Data on ethnic civil conflicts comes from the UCDP/PRIO Armed Conflict Dataset (version 19.1) (Harbom et al., 2008; Allansson et al., 2017; Uppsala Conflict Data Program, n.d.b), with information on the identity of the combatants involved in a conflict (rebel groups and governments). Information on the ethnic identity of rebel groups comes from the ACD2EPR dataset (Vogt et al., 2015; Ethnic Power Relations Dataset Family, 2018a).
Sources: https://ucdp.uu.se/downloads/ and https://icr.ethz.ch/data/epr/acd2epr/

Geodesic distance: I compute geodesic distances between each group's centroid, which I calculate based on ethnic settlements provided by the GeoEPR dataset (Wucherpfennig et al., 2011; Ethnic Power Relations Dataset Family, 2018b). Throughout the analysis, I use log geodesic distance.
Source: https://icr.ethz.ch/data/epr/geoepr/

Absolute difference in elevation: I combine GeoEPR with data on elevation at the grid level (2.5 arc-minute resolution) provided by the Worldclim Global Climate Database (Fick and Hijmans, 2017; WorldClim, 2020a). I measure average elevation of each ethnic group, and then calculate the absolute distance in elevation between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

Absolute difference in ruggedness: I produce a measure of ruggedness for each ethnic group using the standard deviation of the Worldclim elevation data. I calculate the absolute distance in ruggedness between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

Absolute difference in the Caloric Suitability Index: Data on the potential agricultural output come from the Caloric Suitability Index (CSI) provided by Galor and Özak (2016). This measure reflects the potential caloric yield of a grid cell based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). The Index results from a combination of climatic and geographic variables unaffected by human activity. I use the pre-1500 average CSI measure that includes cells with zero productivity. I calculate the average CSI for each ethnic group, and calculate the absolute distance in CSI between each ethnic group and the ethnicities forming the government.
Source: https://ozak.github.io/Caloric-Suitability-Index/

Absolute difference in temperature: Temperature data comes from the GrowUp platform (Girardin et al., 2015), which in turn exploits data from the PRIO-GRID dataset. This is a time-varying measure of mean temperature for each EPR group, which allows me to calculate the absolute distance in mean temperature between each ethnic group and the ethnicities forming the government.
Source: https://growup.ethz.ch/rfe

Absolute difference in rainfall: Rainfall data comes from the WorldClim Global Climate Database (historical monthly level data) (WorldClim, 2020b). I compute a time-varying yearly measure of average rainfall for each group, and calculate the absolute distance in mean rainfall between each ethnic group and the ethnicities forming the government.
Source: https://www.worldclim.org/data/worldclim21.html

Ancestral homelands of contemporary ethnicities: I retrieve the geolocation of ethnic groups' ancestral homelands from the Murdock Map, digitized by Nunn and Wantchekon (2011b,a). I merge EPR ethnicities to the corresponding Murdock groups' counterparts using the LEDA R-package (see details in Section C-2).
Source: https://scholar.harvard.edu/nunn/pages/data-0

Rainfall in the growing season: I construct a measure at the ethnic group and year level that captures the amount of rain received during the main crop's growing season, similar to the one proposed by Guariso and Rogall (2017). I retrieve information about the starting and ending month of the growing season for the group's settlement's main crop from the GrowUp platform (Girardin et al., 2015), which in turn exploits data from the PRIO-GRID dataset. Monthly rainfall data comes from the WorldClim Global Climate Database (historical monthly level data) (WorldClim, 2020b). The final variables I exploit in the analysis are the following: the absolute distance in rainfall between each potential rebel and the groups forming the government whenever the potential rebel receives more rain on average during the growing season than the government; and the absolute distance in rainfall between each potential rebel and the groups forming the government whenever the potential rebel receives less rain during the growing season than the government.
Sources: https://growup.ethz.ch/rfe and https://www.worldclim.org/data/worldclim21.html

Presence of oil fields: Data on the total number of oil field in each ethnic group's settlement area comes from the GrowUp platform (Girardin et al., 2015), which in turn exploits data from the PRIO-GRID dataset. This variable is similar to the one employed by Morelli and Rohner (2015). The final variables I exploit in the analysis are the following: the absolute distance in the number of oil fields between each potential rebel and the groups forming the government whenever the potential rebel holds more oil fields than the government; and the absolute distance in the number of oil fields between each potential rebel and the groups forming the government whenever the potential rebel holds more oil fields than the government.
Source: https://growup.ethz.ch/rfe

The Bantu expansion route. Data on the Bantu expansion route comes from work by Grollemund et al. (2015b,a). They reconstructed the route by collecting the indigenous geographic location of 424 Bantu languages, including now-extinct languages. Using Bayesian techniques on a sample of 100 lexical items, they construct a pyhlogenetic tree connecting all these languages. Based on this, they reconstruct the probable ancestral geographical locations of each of the internal nodes of the phylogenetic tree through a model calibrated using archaeological evidence (e.g., node 1 is dated back to 4000-5000 before present in the Grassfields region of Cameroon, on the basis of an archaeological site called Shum Laka, the principal site associated with Bantu homelands). Using other archaeological sites, they calibrate other branching points of the tree. Using a Brownian motion model, they infer the ancestral latitude and longitude for each internal node of the tree, which they then connect using straight lines. Additional methodological details can be found in Grollemund et al. (2015b) and in Appendix Section B-7.

Opinions on government performance: Data on individuals' opinion on the performance of the current government comes from seven rounds of the Afrobarometer survey (Afrobarometer Data, 1999-2017; BenYishay et al., 2017), conducted between 1999 and 2017 in 27 African countries. See Table A-7 for additional details on the questions asked in each round. Refer to Section 7 for details on how I construct a measure of an individual's degree of disagreement with policies implemented by the current government. I manually merge the self-reported ethnolinguistic affiliation of Afrobarometer respondents to the corresponding Ethnologue language.
Source: http://afrobarometer.org/data

Took or intended to take action against the government: Data on whether individuals took action to express their dissatisfaction against the government comes from seven rounds of the Afrobarometer survey (Afrobarometer Data, 1999-2017; BenYishay et al., 2017). In two rounds of the survey (6 and 7), respondents are asked the following question: "Here is a list of actions that people sometimes take as citizens when they are dissatisfied with government performance. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: (i) Joined others in your community to request action from government; (ii) Contacted the media, like calling a radio program or writing a letter to a newspaper, (iii) Contacted a government official to ask for help or make a complaint; (iv) Refused to pay a

tax or fee to government; (v) Participated in a demonstration or protest march." In the previous survey rounds, questions about similar types of activities are also present, but without the interviewer making clear that the actions are taken as a result of dissatisfaction with the performance of the government and the listed activities tend to be a subset of those elicited in rounds 6 and 7. Questions in previous rounds are posed as follows. Round 1: "I will read out a list of things that people sometimes do as citizens. Please tell me how often you, personally, have done any of these things during the last five years: Write to newspaper, Attend demonstration, Contact government or party official"; Round 2: "Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: Attended a demonstration or protest march, Got together with others to raise an issue, Contact official of a government ministry"; Round 3: " Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: Attended a demonstration or protest march, Contact official of a government ministry, Join others to raise an issue"; Round 4: "Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: Got together with others to raise an issue, Attend a demonstration or protest march, Contact official of a government agency." Round 5: "Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have done any of these things during the past year. If not, would you do this if you had the chance: Attended a demonstration or protest march, Refused to pay tax or fee to government, Join others to raise an issue, Contact official of a government agency." I manually merge the self-reported ethnolinguistic affiliation of Afrobarometer respondents to the corresponding Ethnologue language.
Source: http://afrobarometer.org/data

Age: Age of respondent at the time of survey.

Gender: An indicator variable equal to one if a respondent is female.

Rural: An indicator variable for rural locations.

Opinions on the role of government: Data on individuals' opinion on the general role of government comes from seven rounds of the Afrobarometer survey (Afrobarometer Data, 1999-2017; BenYishay et al., 2017), conducted between 1999 and 2017 in 27 African countries. Refer to Section 7 and Table A-9 for details on how I construct a measure of an ethnic group's view of the role of government. I manually merge the self-reported ethnolinguistic affiliation of Afrobarometer respondents to the corresponding Ethnologue language.
Source: http://afrobarometer.org/data

National budget spent on public (vs particularistic) goods: I construct a binary variable at the country and year level (*Public goods*) based on information included in the V-Dem dataset, which collects a variety of indicators coded by country experts (for details, see Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others (2023)). The variable I use to construct the *Public goods* indicator comes from the following question asked to country experts: "Considering the profile of social and infrastructural spending in the national budget, how "particularistic" or "public goods" are most expenditures?". The V-Dem dataset considers particularistic spending as narrowly targeted on a specific corporation, sector, social group, region, party, or set of constituents. Such spending may be referred to as "pork", "clientelistic", or "private goods." Instead, V-Dem considers public-goods spending as "intended to benefit all communities within a society, even though it may be means-tested so as to target poor, needy, or otherwise underprivileged constituents." V-Dem suggests experts to consider the entire budget of social and infrastructural spending when answering this question. The variable *Public goods* takes value 1 if "most social and infrastructure expenditures are public-goods but a significant portion (e.g., 1/4 or 1/3) is particularistic" or "all social and infrastructure expenditures are public-goods in character. Only a small portion is particularistic.". The variable *Public goods* takes value 0 if "most all of the social and infrastructure expenditures are particularistic" or "most social and infrastructure expenditures are particularistic, but a significant portion (e.g. 1/4 or 1/3) is public-goods" or "social and infrastructure expenditures are evenly divided between particularistic and public-goods programs."
Source: https://v-dem.net/data/the-v-dem-dataset/country-year-v-dem-fullothers-v13/

Distribution of public services across social groups: I construct a binary variable at the country and year level (*Equal distribution*) based on information

included in the V-Dem dataset, which collects a variety of indicators coded by country experts (see Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others (2023) for details). The variable I use to construct the *Equal distribution* indicator comes from the following question asked to country experts: "Are basic public services, such as order and security, primary education, clean water, and healthcare, distributed equally across social groups?". V-Dem considers a social group as "differentiated within a country by caste, ethnicity, language, race, region, religion, migration status, or some combination thereof." V-Dem emphasizes that "social group identity is contextually defined and is likely to vary across countries and through time. Nonetheless, at any given point in time there are social groups within a society that are understood—by those residing within that society—to be different, in ways that may be politically relevant." Given the relevance of ethnicity in African politics (see Francois et al. (2015a) and literature therein cited), this measure is arguably appropriate for my setting. The variable *Public goods* takes value 1 if public services are distributed somewhat equally, relatively equally, or equally and zero otherwise.
Source: https://v-dem.net/data/the-v-dem-dataset/country-year-v-dem-fullothers-v13/

Alternative definitions of ethnicity: To assess the robustness of my findings to an alternative definition of ethnicity, I use the All Minority at Risk (AMAR) dataset (Birnir et al., 2018; The All Minorities at Risk Project, 2015). AMAR includes all contemporary *socially relevant* ethnic groups, defined as groups where members "notice and condition their actions on ethnic distinction in everyday life." AMAR includes a global sample of ethnic groups, in turn divided into subgroups. I consider both groups and subgroups as potential rebels and I match AMAR African potential rebels to their spoken languages using the LEDA package. As an additional robustness test, I define potential rebels using the Geo-referencing of Ethnic Groups (GREG) dataset (Geo-referencing of ethnic groups, n.d.), based on the Soviet Atlas Narodov Mira, the Murdock Ethnographic Atlas (Nunn, 2008b,a; Giuliano and Nunn, 2018b,a), and the Ethnologue. For additional details, refer to Section B-1.
Source (AMAR): http://www.mar.umd.edu/amar_project.asp
Source (GREG): https://icr.ethz.ch/data/greg/
Source (Ethnographic Atlas): https://scholar.harvard.edu/nunn/pages/data-0
Source (Ethnologue): https://www.ethnologue.com/

Deaths of leaders and close elections: Data on the deaths of leaders comes

from Jones and Olken (2005), Table II. Data on close elections comes from the African Elections Database. For additional details on how I categorize elections as "close," see section B-2.
Source (close elections): https://africanelections.tripod.com/about.html

Intensive margin of violence: Data on conflict fatalities comes from the UCDP Battle-Related Deaths Dataset (version 23.1) and is only available from 1989 onwards (Davies et al., 2024; Uppsala Conflict Data Program, n.d.a). Data on conflict intensity comes from the UCDP Dyadic Dataset (Uppsala Conflict Data Program, n.d.b), which provides a categorical variable capturing the intensity of the conflict: "Minor: between 25 and 999 battle-related deaths" or "War: at least 1,000 battle-related deaths in a given year." Based on this indicator, I construct two alternative binary variables: one that equals 1 for small-scale conflicts (minor), and one that equals 1 for large-scale conflicts (war).
Source: https://ucdp.uu.se/downloads/

Alternative data on the ethnic identity of the government: An alternative coding for the ethnic identity of the government comes from Francois et al. (2015a,b). They provide data on the ethnicity of African cabinet members for 15 out of the 44 countries included in my estimation sample, Francois et al. (2015a) from the 1960s to 2004. I manually match each cabinet minister's ethnicity to one or more Ethnologue languages.
Source: https://www.econometricsociety.org/publications/econometrica/browse/2015/03/01/how-power-shared-africa

Ancestral ethnic characteristics: Information on ancestral ethnic characteristics I leverage to decompose culture comes from the Murdock Ethnographic Atlas (Nunn, 2008b,a; Giuliano and Nunn, 2018b,a). For details on the ethnic traits and their classification into mutually-exclusive cultural domains, see Section B-8. I match EPR ethnicities to the corresponding Murdock groups' counterpart using the LEDA R-package (see details in Section C-2).
Source: https://scholar.harvard.edu/nunn/pages/data-0

C-2 Merging Ethnic Groups Across Datasets

To merge EPR ethnic groups to their ancestral homelands in the Murdock Map, I use a recent R-package created by Müller-Crepon et al. (2020) called Linking Ethnic Data for Africa (LEDA). Since ethnic categories vary considerably across datasets, Müller-Crepon et al. (2020) created an algorithm that systematically links ethnic groups across 11 dataset. Using a dictionary-based linking procedure, they match more than 8,100 ethnicities via the list of known language families, languages, and dialects from the 16th edition of the Ethnologue database.

Since links are formed through the linguistic tree, the LEDA package allows the researcher to choose the level of precision for each match. One option is to use the so-called *set overlap* rule, which generates a link between any two groups that share at least one language node at a specified level of the language tree. The higher the level specified, the higher is the completeness, but the lower the precision of the match. Throughout the analysis, I opt for precision of matches, and choose the *dialect* level (i.e., the lowest level) of the linguistic tree of ethnic group in dataset A for generating matches to the ethnic group(s) in dataset B.

Even after choosing the matching procedure that favors precision, some matches remain more precise than others. Conveniently, for each match, the package provides an indication of the degree of precision by indicating the extent to which the linguistic tree of an ethnic group in dataset A overlaps with the linguistic tree of an ethnic group in dataset B. Whenever LEDA matches an ethnic group in dataset A (e.g., the EPR dataset) to multiple groups in dataset B (e.g., the Murdock Map), I employ this accuracy measure as a weight when calculating statistics for group in dataset A based on multiple groups in dataset B (e.g., when computing a measure of Bantu exposure for each EPR group based on the Bantu Index of multiple Murdock groups). Groups in dataset B that constitute more accurate matches will get more weight than groups constituting weaker links.[37]

For additional information on the LEDA package, see:
http://www.carlmueller-crepon.org/project/ethnic_matching/

[37]This occurs often in my setting. Because the definition of ethnic group in the EPR dataset tends to be broader than the one in the Murdock Map, an EPR group is usually linked to multiple groups in the Murdock Map.

C-3 Summary Statistics

TABLE C-1: *Number of Government Changes by Country*

Country	Number of changes	Country	Number of changes
Angola	0	Madagascar	2
Benin	8	Malawi	1
Botswana	0	Mali	1
Burundi	0	Mauritania	0
Cameroon	2	Mauritius	8
Central African Republic	7	Morocco	0
Chad	7	Mozambique	1
Comoros	2	Namibia	1
Congo	9	Niger	10
Congo, DRC	8	Nigeria	9
Cote d'Ivoire	5	Senegal	2
Djibouti	1	Sierra Leone	6
Egypt	0	South Africa	2
Equatorial Guinea	0	South Sudan	0
Eritrea	1	Sudan	2
Ethiopia	2	Tanzania	1
Gabon	2	The Gambia	0
Ghana	7	Togo	3
Guinea	5	Uganda	6
Guinea-Bissau	6	Zambia	1
Kenya	6	Zimbabwe	4
Liberia	2		
Libya	0	**Total**	**140**

Notes: The table reports the number of changes in the ethnic identity of governments experienced by each country in the sample over the period 1961-2017.

TABLE C-2: *Summary Statistics*

	Mean (1)	Std. Dev. (2)	Min (3)	Max (4)	Obs. (5)
Conflict Variables					
Conflict over government power	0.048	0.213	0	1	9,990
Conflict over territory	0.024	0.153	0	1	9,990
Linguistic Distance					
Linguistic Distance^{W}	0.441	0.310	0	1	9,990
Linguistic Distance^{UW}	0.442	0.306	0	1	9,990
Geographic Controls					
Log geodesic distance^{W}	5.524	1.271	0	7.284	9,990
Log geodesic distance^{UW}	5.519	1.284	0	7.274	9,990
Absolute distance ruggedness^{W}	132.9	153.3	0	1,223	9,990
Absolute distance ruggedness^{UW}	135.2	154.9	0	1,238	9,990
Absolute distance elevation^{W}	216.2	239.2	0	1,385	9,990
Absolute distance elevation^{UW}	216.8	237.9	0	1,329	9,990
Absolute distance CSI^{W}	581.1	1149	0	9,209	9,990
Absolute distance CSI^{UW}	582.4	1128	0	9,206	9,990

Notes: The sample includes 265 distinct ethnic groups in 44 African countries over a period of 57 years (1961-2017). The superscript W indicates average distances between potential rebels and each ethnic group in the government coalition weighted by the role of ethnic groups in power (where senior partners receive double the weight of each junior partner); the superscript UW indicates unweighted average distances between potential rebels and each ethnic group in the government coalition.

TABLE C-3: *Summary Statistics: Bantu Region*

	Mean (1)	Std. Dev. (2)	Min (3)	Max (4)	Obs. (5)
Conflict Variables					
Conflict over power	0.047	0.211	0	1	4,027
Conflict over territory	0.007	0.085	0	1	4,027
Linguistic Distance					
Linguistic Distance^{W}	0.327	0.270	0	1	4,027
Linguistic Distance^{UW}	0.333	0.272	0	1	4,027
Bantu Distance					
Bantu Distance^{W}	0.053	0.054	0	0.491	4,027
Bantu Distance^{UW}	0.060	0.061	0	0.673	4,027
Geographic Controls					
Log geodesic distance^{W}	5.628	1.320	0	7.284	4,027
Log geodesic distance^{UW}	5.631	1.320	0	7.274	4,027
Absolute distance ruggedness^{W}	160.1	119.2	0	676.1	4,027
Absolute distance ruggedness^{UW}	165.4	127.0	0	688.7	4,027
Absolute distance elevation^{W}	249.9	234.4	0	1,385	4,027
Absolute distance elevation^{UW}	255.2	234.0	0	1,327	4,027
Absolute distance CSI^{W}	359.1	586.3	0	3,646	4,027
Absolute distance CSI^{UW}	392.8	646.0	0	3,440	4,027

Notes: The sample includes 86 distinct ethnic groups in 18 African countries corresponding to regions covered by Bantu languages according to the Guthrie classification (Angola, Botswana, Burundi, Cameroon, Congo, Congo DRC, Equatorial Guinea, Gabon, Kenya, Malawi, Mozambique, Namibia, Rwanda, South Africa, Tanzania, Uganda, Zambia, Zimbabwe) over a period of 57 years (1961-2017). The superscript W indicates average distances between potential rebels and each ethnic group in the government coalition weighted by the role of ethnic groups in power (where senior partners receive double the weight of each junior partner); the superscript UW indicates unweighted average distances between potential rebels and each ethnic group in the government coalition.

References

Afrobarometer Data, 1999-2017.

Alesina, Alberto and Edward Ludwig Glaeser, *Fighting poverty in the US and Europe: A world of difference*, Oxford University Press, USA, 2004.

_ **and George-Marios Angeletos**, "Fairness and redistribution," *American economic review*, 2005, *95* (4), 960–980.

_ **and Nicola Fuchs-Schündeln**, "Good-bye Lenin (or not?): The effect of communism on people's preferences," *American Economic Review*, 2007, *97* (4), 1507–1528.

_ **and Paola Giuliano**, "The power of the family," *Journal of Economic growth*, 2010, *15*, 93–125.

_ **and** _ , "Preferences for redistribution," in "Handbook of social economics," Vol. 1, Elsevier, 2011, pp. 93–131.

_ **,** _ **, and Nathan Nunn**, "On the origins of gender roles: Women and the plough," *The quarterly journal of economics*, 2013, *128* (2), 469–530.

Allansson, Marie, Erik Melander, and Lotta Themnér, "Organized violence, 1989–2016," *Journal of Peace Research*, 2017, *54* (4), 574–587.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Becker, Sascha O, Katrin Boeckh, Christa Hainz, and Ludger Woessmann, "The empire is dead, long live the empire! Long-run persistence of trust and corruption in the bureaucracy," *The Economic Journal*, 2016, *126* (590), 40–74.

Benabou, Roland and Jean Tirole, "Belief in a just world and redistributive politics," *The Quarterly journal of economics*, 2006, *121* (2), 699–746.

BenYishay, A, R Rotberg, J Wells, Z Lv, S Goodman, L Kovacevic, and D Runfola, "Geocoding Afrobarometer rounds 1-6: Methodology & data quality. Williamsburg, VA: AidData at William & Mary. Retrieved November 20, 2018," 2017.

Birnir, Jóhanna K, David D Laitin, Jonathan Wilkenfeld, David M Waguespack, Agatha S Hultquist, and Ted R Gurr, "Introducing the AMAR (All Minorities at Risk) Data," *Journal of Conflict Resolution*, 2018, *62* (1), 203–226.

Bormann, Nils-Christian, Lars-Erik Cederman, and Manuel Vogt, "Language, religion, and ethnic civil war," *Journal of Conflict Resolution*, 2017, *61* (4), 744–771.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2021.

Bostoen, Koen and Hilde Gunnink, "The impact of autochthonous languages on Bantu language variation: A comparative view on Southern and Central Africa," in "The Cambridge handbook of language contact: volume 1: population movement and language change," Cambridge University Press, 2022, pp. 152–186.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson, "Mapping the origins and expansion of the Indo-European language family," *Science*, 2012, *337* (6097), 957–960.

Butters, R Andrew, Daniel W Sacks, and Boyoung Seo, "How do national firms

respond to local cost shocks?," *American Economic Review*, 2022, *112* (5), 1737–1772.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Cao, Yiming, Benjamin Enke, Armin Falk, Paola Giuliano, and Nathan Nunn, "Herding, warfare, and a culture of honor: Global evidence," Technical Report, National Bureau of Economic Research 2021.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The effect of minimum wages on low-wage jobs," *The Quarterly Journal of Economics*, 2019, *134* (3), 1405–1454.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American economic review*, 2020, *110* (9), 2964–2996.

Coppedge, Michael and Gerring, John and Knutsen, Carl Henrik and Lindberg, Staffan I. and others , "V-Dem Dataset v13," 2023.

Corneo, Giacomo and Hans Peter Grüner, "Social limits to redistribution," *American Economic Review*, 2000, *90* (5), 1491–1507.

_ **and** _ , "Individual preferences for political redistribution," *Journal of public Economics*, 2002, *83* (1), 83–107.

Currie, Thomas E, Andrew Meade, Myrtille Guillon, and Ruth Mace, "Cultural phylogeography of the Bantu Languages of sub-Saharan Africa," *Proceedings of the Royal Society B: Biological Sciences*, 2013, *280* (1762), 20130695.

Davies, Shawn, Garoun Engström, Therese Pettersson, and Magnus Öberg, "Organized violence 1989–2023, and the prevalence of organized crime groups," *Journal of Peace Research*, 2024, *61* (4), 673–693.

Deshpande, Manasi and Yue Li, "Who is screened out? Application costs and the targeting of disability programs," *American Economic Journal: Economic Policy*, 2019, *11* (4), 213–248.

Desmet, Klaus, Shlomo Weber, and Ignacio Ortuño-Ortín, "Linguistic diversity and redistribution," *Journal of the European Economic Association*, 2009, *7* (6), 1291–1318.

Diamond, Jared, "Guns, Germs, and Steel: The Fates of Human Societies," 1997.

Dickens, Andrew, "Ethnolinguistic favoritism in african politics," *American Economic Journal: Applied Economics*, 2018, *10* (3), 370–402.

Enke, Benjamin, "Kinship, cooperation, and the evolution of moral systems," *The Quarterly Journal of Economics*, 2019, *134* (2), 953–1019.

Esping-Andersen, Gøsta, *Social foundations of postindustrial economies*, Oxford university press, 1999.

Esteban, Joan, Laura Mayoral, and Debraj Ray, "Ethnicity and conflict: An empirical study," *American Economic Review*, 2012, *102* (4), 1310–42.

Ethnic Power Relations Dataset Family, "ACD2EPR 2018," 2018.

_ , "GeoEPR 2018 dataset," 2018.

_ , "The Ethnic Power Relations - Ethnic Dimensions (EPR-ED) dataset Version 2018," n.d.

Fearon, James D, "Ethnic and cultural diversity by country," *Journal of economic*

growth, 2003, *8*, 195–222.

_ , "Ethnic mobilization and ethnic violence," 2008.

_ **and David D Laitin**, "Violence and the social construction of ethnic identity," *International organization*, 2000, *54* (4), 845–877.

Fehn, Anne-Maria, Beatriz Amorim, and Jorge Rocha, "The linguistic and genetic landscape of southern Africa," *Journal of Anthropological Sciences*, 2022, *100*, 243–265.

Fick, Stephen E and Robert J Hijmans, "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas," *International journal of climatology*, 2017, *37* (12), 4302–4315.

Filippo, Cesare De, Koen Bostoen, Mark Stoneking, and Brigitte Pakendorf, "Bringing together linguistic and genetic evidence to test the Bantu expansion," *Proceedings of the Royal Society B: Biological Sciences*, 2012, *279* (1741), 3256–3263.

Francois, Patrick, Ilia Rainer, and Francesco Trebbi, "How is power shared in Africa?," *Econometrica*, 2015, *83* (2), 465–503.

_ , _ , and _ , "Replication data for: "How Is Power Shared in Africa?"," 2015.

Freedman, Seth M, Alex Hollingsworth, Kosali I Simon, Coady Wing, and Madeline Yozwiak, "Designing Difference in Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines," Technical Report, National Bureau of Economic Research 2023.

Galor, Oded and Ömer Özak, "Replication data for: The Agricultural Origins of Time Preference," 2016.

Geo-referencing of ethnic groups, "GREG shapefile," n.d.

Girardin, Luc, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann, and Manuel Vogt, "Growup-geographical research on war, unified platform," *ETH Zurich. http://growup. ethz. ch*, 2015, p. 33.

Giuliano, P. and N. Nunn, "Ancestral characteristics of modern populations. Underlying ethnicity-level data," 2018.

Giuliano, Paola and Antonio Spilimbergo, "Growing up in a recession: Beliefs and the macroeconomy," Technical Report, National Bureau of Economic Research 2009.

_ **and Nathan Nunn**, "Ancestral characteristics of modern populations," *Economic History of Developing Regions*, 2018, *33* (1), 1–17.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of econometrics*, 2021, *225* (2), 254–277.

Greenberg, Joseph H, "The measurement of linguistic diversity," *Language*, 1956, *32* (1), 109–115.

Grollemund, R., S. Branford, K. Bostoen, A. Meade, C. Venditti, and M. Pagel, "Fig. 2. Ancestral migration route reconstruction," 2015.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel, "Bantu expansion shows that habitat alters the route and pace of human dispersals," *Proceedings of the National Academy of Sciences*, 2015, *112* (43), 13296–13301.

Grosjean, Pauline, "The weight of history on European cultural integration: a gravity approach," *American Economic Review*, 2011, *101* (3), 504–508.

_ , "A history of violence: The culture of honor and homicide in the US South," *Journal of the European Economic Association*, 2014, *12* (5), 1285–1316.

Guariso, Andrea and Thorsten Rogall, "Rainfall inequality, political power, and ethnic conflict in Africa," 2017.

Guarnieri, Eleonora and Ana Tur-Prats, "Cultural distance and conflict-related sexual violence," *The Quarterly Journal of Economics*, 2023, *138* (3), 1817–1861.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "Social capital as good culture," *Journal of the European economic Association*, 2008, *6* (2-3), 295–320.

Harbom, Lotta, Erik Melander, and Peter Wallensteen, "Dyadic dimensions of armed conflict, 1946—2007," *Journal of peace research*, 2008, *45* (5), 697–710.

Herbst, Jeffrey, "Economic incentives, natural resources and conflict in Africa," *Journal of African Economies*, 2000, *9* (3), 270–294.

Holden, Clare Janaki, "Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2002, *269* (1493), 793–799.

Jones, Benjamin F and Benjamin A Olken, "Do leaders matter? National leadership and growth since World War II," *The Quarterly Journal of Economics*, 2005, *120* (3), 835–864.

Kirby, Kathryn R, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, Carlos A Botero, Claire Bowern, Carol R Ember et al., "D-PLACE: A global database of cultural, linguistic and environmental diversity," *PloS one*, 2016, *11* (7), e0158391.

Koile, Ezequiel, Simon J Greenhill, Damián E Blasi, Remco Bouckaert, and Russell D Gray, "Phylogeographic analysis of the Bantu language expansion supports a rainforest route," *Proceedings of the National Academy of Sciences*, 2022, *119* (32), e2112853119.

la Cuesta, Brandon De and Kosuke Imai, "Misunderstandings about the regression discontinuity design in the study of close elections," *Annual Review of Political Science*, 2016, *19* (1), 375–396.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter, "Valid t-Ratio Inference for IV," *American Economic Review*, 2022, *112* (10), 3260–3290.

Lewis, M Paul, Gary F Simons, and Charles D Fennig, "Ethnologue: Languages of the World," *Dallas, Texas: SIL International. Online version: http://www.ethnologue.com*, 2020.

Mamdani, Mahmood, "Indirect rule, civil society, and ethnicity: The African dilemma," *Social Justice*, 1996, *23* (1/2 (63-64), 145–150.

Michalopoulos, Stelios and Elias Papaioannou, "National institutions and subnational development in Africa," *The Quarterly journal of economics*, 2014, *129* (1), 151–213.

_ **and** _ , "The long-run effects of the scramble for Africa," *American Economic Review*, 2016, *106* (7), 1802–48.

Morelli, Massimo and Dominic Rohner, "Resource concentration and civil wars," *Journal of Development Economics*, 2015, *117*, 32–47.

Moscona, Jacob, Nathan Nunn, and James A Robinson, "Segmentary Lineage

Organization and Conflict in Sub-Saharan Africa," *Econometrica*, 2020, *88* (5), 1999–2036.

Müller-Crepon, Carl, Yannick Pengl, and Nils-Christian Bormann, "Linking Ethnic Data from Africa," 2020.

Nickell, Stephen, "Biases in dynamic models with fixed effects," *Econometrica: Journal of the econometric society*, 1981, pp. 1417–1426.

Nohlen, Dieter, Bernard Thibaut, and Michael Krennerich, *Elections in Africa: A data handbook*, OUP Oxford, 1999.

Nunn, N., "The long-term effects of Africa's slave trades: Murdock map," 2008.

Nunn, Nathan, "The long-term effects of Africa's slave trades," *The Quarterly journal of economics*, 2008, *123* (1), 139–176.

_ **and Leonard Wantchekon**, "Replication data for: The Slave Trade and the Origins of Mistrust in Africa," 2011.

_ **and** _ , "The slave trade and the origins of mistrust in Africa," *American Economic Review*, 2011, *101* (7), 3221–52.

Oliveira, Sandra, Anne-Maria Fehn, Beatriz Amorim, Mark Stoneking, and Jorge Rocha, "Genome-wide variation in the Angolan Namib Desert reveals unique pre-Bantu ancestry," *Science Advances*, 2023, *9* (38), eadh3822.

Rossignol, Etienne Le and Sara Lowes, "Ancestral livelihoods and moral universalism: Evidence from transhumant pastoralist societies," Technical Report, National Bureau of Economic Research 2022.

Rothchild, Donald and Victor A Olorunsola, *State versus ethnic claims: African policy dilemmas*, Routledge, 2019.

Ruggles, Steven, Lara Cleveland, Rodrigo Lovaton, Sula Sarkar, Matthew Sobek, Derek Burk, Dan Ehrlich, Quinn Heimann, and Jane Lee, "Integrated Public Use Microdata Series, International: Version 7.5 [dataset]," *Minneapolis, MN: IPUMS*, 2024.

Spolaore, Enrico and Romain Wacziarg, "War and relatedness," *Review of Economics and Statistics*, 2016, *98* (5), 925–939.

_ **and** _ , "The political economy of heterogeneity and conflict," Technical Report, National Bureau of Economic Research 2017.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of econometrics*, 2021, *225* (2), 175–199.

The All Minorities at Risk Project, "AMAR Ethnic Groups List," 2015.

Todd, Emmanuel and David Garrioch, "The explanation of ideology: Family structures and social systems," *(No Title)*, 1985.

Uppsala Conflict Data Program, "UCDP Battle-Related Deaths Dataset version 23.1," n.d.

_ , "UCDP Dyadic Dataset version 19.1 (1946-2018)," n.d.

Vansina, Jan, "New linguistic evidence and 'the Bantu expansion'," *The Journal of African History*, 1995, *36* (2), 173–195.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin, "Integrating data on ethnicity,

geography, and conflict: The ethnic power relations data set family," *Journal of Conflict Resolution*, 2015, *59* (7), 1327–1342.

Voigtländer, Nico and Hans-Joachim Voth, "Persecution perpetuated: the medieval origins of anti-Semitic violence in Nazi Germany," *The Quarterly Journal of Economics*, 2012, *127* (3), 1339–1392.

Weidmann, Nils B, Jan Ketil Rød, and Lars-Erik Cederman, "Representing ethnic groups in space: A new dataset," *Journal of Peace Research*, 2010, *47* (4), 491–499.

Whiteley, Peter M, Ming Xue, and Ward C Wheeler, "Revising the Bantu tree," *Cladistics*, 2019, *35* (3), 329–348.

Wing, Coady, Seth M Freedman, and Alex Hollingsworth, "Stacked Difference-in-Differences," Technical Report, National Bureau of Economic Research 2024.

WorldClim, "WorldClim version 2.1, Elevation 2.5 minutes," 2020.

_ , "WorldClim version 2.1, Precipitation 2.5 minutes (1961-2017)," 2020.

Wucherpfennig, Julian, Nils B. Weidmann, Luc Girardin, Lars-Erik Cederman, and Andreas Wimmer, "Politically Relevant Ethnic Groups across Space and Time: Introducing the GeoEPR Dataset," *Conflict Management and Peace Science*, 2011, *28* (5), 423–437.

_ **, Nils W Metternich, Lars-Erik Cederman, and Kristian Skrede Gleditsch**, "Ethnicity, the State, and the Duration of Civil War," *World Politics*, 2012, *64* (1), 79–115.