Ethnolinguistic Favoritism in African Politics

Andrew Dickens[†]

11 October 2016

CLICK HERE FOR THE MOST RECENT VERSION

I document evidence of ethnic favoritism in a panel of 163 ethnolinguistic groups partitioned across 35 African countries. In contrast to previous studies, I construct a computerized lexicostatistical measure of linguistic similarity between each ethnic group and the national leader as a novel measure of ethnic proximity. I exploit the arbitrary placement of African political borders as a source of exogenous within-group variation, where the similarity of the same partitioned group varies over time according to the ethnolinguistic identity of the national leader on each side of the border. To quantify patronage at the group level, I isolate time variation in night light luminosity resulting from changes in the ethnolinguistic identity of a leader. Using a triple-difference estimator I find that a one standard deviation increase in linguistic similarity yields a 7.0 percent increase in luminosity, which corresponds to a 2.1 percent increase in group-level GDP per capita. I then use the continuity of linguistic similarity to show that favoritism exists among groups that are not coethnic to the leader, where the mean effect of non-coethnic similarity is one quarter the size of the coethnic effect. I corroborate this evidence using individual-level data and establish that it's where an individual lives and the attached ethnolinguistic identity that predicts favoritism, not the identity of the individual respondent. I relate these results to the literature on coalition building, and provide evidence that ethnicity is one of the guiding principles behind high-level government appointments.

[†]York University, Department of Economics, Toronto, ON. E-mail: adickens@yorku.ca. I am indebted to Nippe Lagerlöf for his encouragement and detailed feedback, and Tasso Adamopoulos, Berta Esteve-Volart and Ben Sand for their guidance throughout this project. I thank Matthew Gentzkow and two anonymous referees for helpful suggestions. Detailed feedback from James Fenske was especially helpful. I also thank Greg Casey, Mario Carillo, Raphaël Franck, Oded Galor, Andrew Hencic, Fernando Leibovici, Stelios Michalopoulos, Stein Monteiro, Laura Salisbury, Assaf Sarid, Andrey Stoyanov, Francesco Trebbi and David Weil for helpful comments, in addition to seminar participants at Brown University, NEUDC, RES Symposium for Junior Researches, Economic History Association's Annual Meeting, 3rd annual PODER Summer School, Canadian Economics Association's Annual Meeting and York University. This research is funded by the Social Science and Humanities Research Council of Canada. All errors are my own.

1 Introduction

Understanding why global poverty is so concentrated in Africa remains one of the most crucial areas of inquiry in the social sciences. One long-standing explanation is that Africa's high level of ethnic diversity is a major source of its underdevelopment and political instability (Easterly and Levine, 1997; Collier and Gunning, 1999; Posner, 2004; Alesina and La Ferrara, 2005, among others). Yet recent evidence documents that the source of underdevelopment is not ethnic diversity per se, but rather Africa's high degree of inequality between ethnic groups (Alesina et al., 2016). This suggests that ethnic diversity is only an impediment to economic development when some ethnicities prosper at the expense of others.

Ethnic inequality not only contributes to the under-provision of the overall level of public resources (Baldwin and Huber, 2010), but it provokes discriminatory policies that advantage some groups over others (Alesina et al., 2016). Discriminatory policies of this type are a form of ethnic favoritism, which has been the subject of a few influential papers that document evidence of public resource distribution across ethnic lines in Africa (Franck and Rainer, 2012; Burgess et al., 2015; Kramon and Posner, 2016). The provision of resources on the basis of ethnicity – rather than on a need or marginal value basis – suggests that some ethnic groups are being systematically favored over others. Hence, a better understanding of how ethnic patronage is distributed and to whom is important because it sheds light on the extent to which favoritism occurs and how some ethnic groups benefit at the expense of others.

In this paper I revisit the study of ethnic favoritism with three contributions. My first contribution is a novel measure of linguistic similarity that quantifies the relative similarity of all ethnic groups to the national leader in each country, not just groups that share an ethnicity with the leader (coethnics). Because linguistic similarity is measured on the unit interval it encompasses the commonly used coethnic dummy variable, while extending measurement to all non-coethnic groups.[1] This extension is beneficial because the majority of Africans are never coethnic to their leader.[2] The continuity of this new measure implies that *any* change in the ethnic identity of a leader is associated with *some* change in the similarity of *all* groups in a country, an important source of variation that is not observable using a coethnic dummy variable. This measure also provides testable grounds for the central hypothesis of this paper: a group's well-being is increasing in their ethnic similarity to the national leader.

My second contribution relates to the evidence that ethnic favoritism is widespread

[1] People identify as coethnics because they share a common ancestry and language, hold similar cultural beliefs and pursue related economic activities (Batibo, 2005). In this way, linguistic similarity is a good measure of ethnic proximity because it is the most visible marker of ethnic identity.

[2] Using population data from the Ethnologue (16^{th} edition), I calculate that only 34 percent of the median sub-Saharan country's population was *ever* coethnic to their leader between 1992 and 2013.

throughout sub-Saharan Africa. I use the systematic partitioning of African ethnic groups across political borders to expand the scope of evidence relative to previous studies. In particular, I exploit the fact that the same ethnic group is split between neighboring countries and exposed to a different ethnic leader on each side of the border. As different ethnic leaders come and go from power, the relative similarity of a partitioned group varies over time. This source of variation allows for ethnicity-year fixed effects, a novel empirical specification that accounts for the long-run persistence of a group's pre-colonial history on group-level outcomes today (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013; Fenske, 2013). I use this variation in a triple difference set-up and document evidence of ethnic favoritism in two empirical settings: a panel of 163 ethnic groups split across 35 countries, and a repeated cross-section of individuals living in 20 groups split across 13 countries. I also use the continuity of similarity in both settings to show that favoritism exists among groups that are not coethnic to the leader, a new finding that is a contribution in itself. This speaks to why a continuous measure of ethnic similarity is important: ethnic favoritism is under-reported when using a coethnic dummy variable because non-coethnic favoritism goes undetected. In order to understand the impact of ethnic favoritism on development, it is important to understand the extent to which it occurs.

For my third contribution I disentangle the relative importance of location-based favoritism from individual-level favoritism. It is commonly assumed that the ethnic majority of a region defines the ethnic identity of that region, despite the fact that not all residents belong to the dominant group. While this is a reasonable assumption, it limits our understanding of how patronage is distributed because no distinction can be made between regional transfers and targeted transfers towards individuals. To understand who benefits from favoritism it is necessary to understand whether the benefits of similarity are exclusive to individuals living in their ethnic homeland, or if patronage is distributed more broadly by targeting individuals irrespective of location. To this end, I use variation among survey respondents who identify with an ethnicity that is different from the ethnic region in which they live.[3] I find that patronage is distributed according to the ethnic identity of a region rather than as a targeted transfer towards individuals from a particular ethnic group.

Throughout this empirical analysis I rely on the fact that the location of African borders are quasi-random (Englebert et al., 2002; Michalopoulos and Papaioannou, 2014, 2016, among others). The historical formation of Africa's borders began with the Berlin Conference of 1884-1885, where European powers divided up Africa with little regard for the spatial distribution of ethnic homelands (Herbst, 2000). This disregard led to the arbitrary

[3]This is analogous to Nunn and Wantchekon (2011), who use a similar source of variation to separate internal norms of an individual from the external norms of an individual's environment.

formation of national borders, which "did not reflect reality but helped create it" (Wesseling, 1996, p.364). One such reality was the partitioning of approximately 200 ethnic groups throughout Africa.

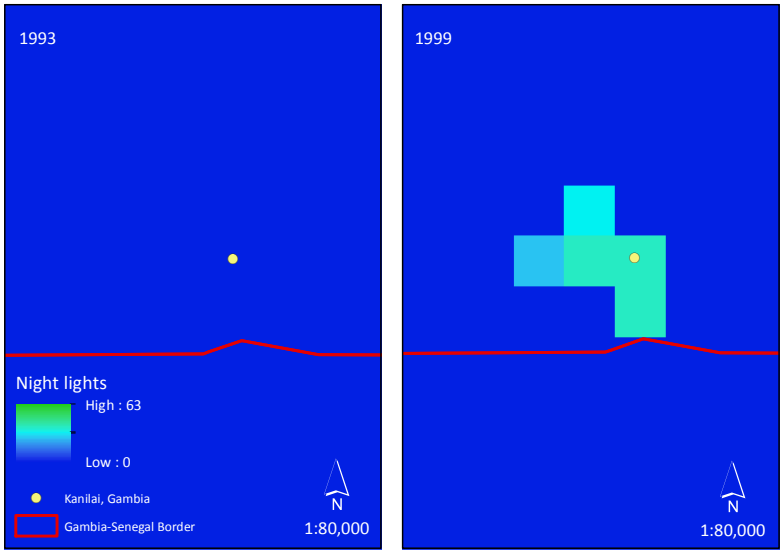
In the context of this study, the quasi-random nature of African border design generates exogenous variation because the ethnic identity of a national leader varies across borders within the same partitioned group. Because an ethnic group shares a common ancestry, and is relatively homogeneous in terms of cultural and biological factors, the fraction of a partitioned group on one side of the border is a suitable counterfactual observation for the other fraction of that same group across the border. Using ethnic groups partitioned across African borders as a source of exogenous variation is methodologically similar to Michalopoulos and Papaioannou (2014).

To exploit this within-group variation I use the 16th edition of the Ethnologue language map (Lewis, 2009). This map depicts the spatial distribution of ethnolinguistic homelands across the world. I use these subnational groups as a spatial unit of observation in Africa. Because no income data exists at this level of observation, I proxy an ethnic group's economic activity using annual satellite images of night light luminosity for the time period 1992-2013. These luminosity data are available at a very fine spatial resolution, which I can use to construct a panel of economic activity at the country-group level. Luminosity is frequently used as a measure of economic activity because of its strong empirical association with GDP per capita and other measures of living standards (Henderson et al., 2012; Michalopoulos and Papaioannou, 2013, 2014; Alesina et al., 2016, among others). Hodler and Raschky (2014) first used luminosity in this way to study patterns of regional favoritism.

Consider, as an example, the Jola-Fonyi language group partitioned across Gambia and Senegal. In 1993, both the Gambian and Senegalese Jola-Fonyi bear little resemblance to their respective leaders. For several years little changed in Senegal as President Diouf's reign continued. On the contrary, much changed for the Gambian Jola-Fonyi when Yahya Jammeh, a young officer in the National Gambian Army, overthrew President Jawara in a 1994 military coup. Jammeh was born in Kanilai, a small village near the southern border of Gambia and home to the Jola-Fonyi language group. Jammeh took much pride in his birth region – a "place that gained prominence overnight in Gambia" (Mwakikagile, 2010, p. 56). Jammeh repeatedly "feathered his nest" to such an extent that the Jola-Fonyi region surrounding Kanilai is one of few rural areas in Gambia with "electricity, street lighting, paved roads and running water – not to mention its own zoo and game preserve, wrestling arena, bakery and luxury hotel with a swimming pool" (Wright, 2015, p. 219).

Figure 1 provides visual evidence of this phenomenon. The two panels represent the same subsection of the Jola-Fonyi language group at two points in time, with the border

Figure 1: Change in Night Lights Intensity from 1993-1999



This figure documents the change in night light activity in the partitioned Jola-Fonyi language group in Gambia (north of the border) and Senegal (south of the border) between 1993 and 1999. In 1994, Yahya Jammeh assumed power of Gambia and soon after started reallocating funds to the Jola-Fonyi. Within 5 years of presidency the Gambian Jola-Fonyi exhibit much greater economic activity in terms of night lights than the Senegalese Jola-Fonyi on the south side of the border, whom had no change in leadership during this period.

dividing Gambia to the north and Senegal to the south. While there is no visible night light activity on either side of the border in 1993, there is a significant increase in lights on the Gambian side only 5 years after Jammeh assumed power. On the contrary, Diouf's presidency continued throughout this entire period and there is no observable change in night light activity in Senegal just south of the border. This demonstrated change in Figure 1 is exactly the within-group variation that I use to estimate the effect of similarity. In this case, the Senegalese Jola-Fonyi are the counterfactual observation for the Gambian Jola-Fonyi, who are equally dissimilar in language to their incumbent leader in 1993, and the effect of similarity on night light activity is estimated off of the change in linguistic similarity following Jammeh's rise to power.

My benchmark results imply that a standard deviation increase in linguistic similarity (23 percent) yields a 7 percent increase in luminosity and a 2 percent increase in group-level GDP per capita. I also use the continuity of linguistic similarity to document evidence of non-coethnic favoritism, where the mean non-coethnic effect is one quarter the coethnic premium. To the contrary I find no evidence of anticipatory effects in the data or evidence migration in response to leadership changes. To be sure this result is not a consequence of my new measure of similarity I construct two alternative measures: a standard binary

measure of coethnicity and a discrete similarity measure of the ratio of shared nodes on the Ethnologue language tree. While these alternative measures of similarity yield significant evidence of favoritism, my preferred lexicostatistical measure of similarity is more precisely estimated and the only measure to maintain significance in a series of horse race regressions.

I also test for a variety of mechanisms, but find no systematic evidence of the usual channels (e.g., democracy). However, I do find that my benchmark result is largely driven by leaders who have held office longer than the sample median of nine years. This implies that one determinant of favoritism is leadership tenure.

Next I turn to individual-level data from the Demographic and Health Survey (DHS). I use survey cluster coordinates to pinpoint the location of individual respondents on the Ethnologue map. Doing so allows me to construct a repeated cross-section of individuals living in partitioned ethnic groups across DHS survey waves. Narrowing the focus to these individuals allows me to exploit the same variation I use in my benchmark estimates. As an outcome I use an individual-level measure of access to public resources and ownership of assets. I corroborate my benchmark findings with this individual-level data, including evidence of non-coethnic favoritism. I also establish that patronage is distributed regionally and not as a targeted transfer towards individuals.

These findings speak to a sparse but growing body of evidence that ethnic favoritism is widespread throughout Africa. Franck and Rainer (2012) use a panel of ethnic groups in 18 countries to document evidence of favoritism throughout sub-Saharan Africa. What sets my paper apart from Franck and Rainer's (2012) is that I construct a panel of *partitioned* ethnic groups, so I have a minimum of two country-group observations for any partitioned group in a year. This feature of my data affords me ethnicity-year fixed effects. Because I can account for all observable and unobservable time-varying features of an ethnic group, I am able to rule out endogeneity concerns associated with the impact of pre-colonial group characteristics on contemporary development (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013; Fenske, 2013).

More commonly researchers focus on a single patronage good in a single country. Kramon and Posner (2016) find that Kenyans whom are coethnic to their leader attain higher levels of education, while Burgess et al. (2015) find that Kenyan districts associated with the leader's ethnicity receive two times the investment in roads during periods of autocracy. At an even finer level, Marx et al. (2015) document evidence of ethnic favoritism in housing markets within a large slum outside of Nairobi.

The rich micro-data these studies use provide clear evidence of ethnic favoritism and the channels through which patronage is distributed. Yet generalizing these results is difficult because of the highly localized analyses these studies employ. To this end, I exploit the

systematic partitioning of ethnic groups in Africa to expand the scope of evidence to 35 sub-Saharan countries. In a related manuscript, De Luca et al. (2015) document that ethnic favoritism is an axiom of politics on a global scale and not simply an African phenomenon. As this literature continues to grow, these localized studies coupled with the broader evidence of ethnic favoritism help to build consensus around ethnic favoritism in Africa.[4]

The notion that ethnic favoritism drives discriminatory policies that disadvantage some groups at the expense of others also relates this research to the literature on ethnic inequality and conflict. Alesina et al. (2016) document that the negative correlation between ethnic inequality and economic development is a global phenomenon, though most pronounced in Africa. Income differences between a country's ethnic groups can also impact the political process: ethnic inequality mitigates public good provision (Baldwin and Huber, 2010), diminishes the quality of governance (Kyriacou, 2013), and provokes the "ethnification" of political parties (Huber and Suryanarayan, 2014). At the heart of this literature is the long-standing instrumentalist view that conflict over scarce resources drives ethnic competition in Africa (Bates, 1974). Even the perception of ethnic favoritism exacerbates already existing ethnic tensions (Bowles and Gintis, 2004), which itself can further incite ethnic conflict (Esteban and Ray, 2011; Esteban et al., 2012; Caselli and Coleman, 2013).

I contribute to this line of research with evidence that regions, rather than individuals, tend to be targeted, and that non-coethnic groups that are similar to the leader stand to gain from their ethnic proximity. The fact that similar but not identical ethnic regions benefit from patronage suggests that ethnicity is more than just a marker of identity: similarity may create affinity or reduce coordination costs across related non-coethnic groups. This is consistent with the idea that these broader ethnic connections may solve collective action problems (Miguel and Gugerty, 2005) and bring about greater between-group trust (Habyarimana et al., 2009). The continuity of linguistic similarity captures these affinities that are otherwise unobservable with a coethnic dummy variable, thus highlighting one further benefit of this new measure.

This is also in line with the idea that leaders bring elites from outside of their ethnic group into the governing coalition in an effort to sustain power in the face of political instability (Joseph, 1987; Francois et al., 2015). In the discussion section of this paper I provide evidence that leaders appoint similar but not identical ethnic elites to high-level government positions for this purpose. In doing so, non-coethnic groups gain coethnic representation in government, where representatives speak on their behalf and channel resources to them (Ar-

[4]Yet consensus on ethnic favoritism if Africa has not been reached. Francois et al. (2015) document that leaders only provide a small premium to their coethnics, and otherwise political power is proportional to group size in Africa. Kasara (2007) finds that leaders are more likely to extract taxes from their own ethnic group because they have a better understanding of internal markets in their homeland.

riola, 2009). Although my focus is Africa, this deeper understanding of where favoritism is expected to take place has implications for distributive politics more broadly: it contributes to our knowledge of how targeted transfers can potentially magnify inequality between groups and thus is informative of a determining factor of comparative economic development.

The rest of this paper is structured as follows. Section 2 describes how I identify language group partitions and measure linguistic similarity. This section also documents patterns in the data. Section 3 outlines the empirical model and identification strategy, and Section 4 reports the benchmark estimates and robustness checks. Section 5 disentangles the relative importance of location-based favoritism from individual-level favoritism, and in Section 6 I link the findings to the literature on ethnic favoritism and coalition building, and provide suggestive evidence that non-coethnic favoritism works through the appointment of elites from outside of the leader's ethnic group. Section 7 concludes.

2 Data

In this section I describe the main variables of interest. For a complete description of all data and sources see Appendix A.

2.1 Language Group Partitions

I construct language group partitions using the 2009 Ethnologue (16^{th} edition) mapping of language groups from the World Language Mapping System (WLMS). These WLMS data depict the spatial distribution of linguistic homelands at the country-language group level (Figure 2). I focus on continental sub-Saharan Africa.[5] In total there are 2,384 country-language group observations reflecting 1,961 unique language groups in 42 continental African countries.[6]

I define a partition as a set of contiguous country-language group polygons, where each polygon in a set is part of the same language group but separated by a national border. I use ArcGIS to identify these partitioned groups, excluding country-language groups with a reported Ethnologue population of zero. The result is 486 remaining country-language group observations, made up of 227 language groups partitioned across 37 African countries.

[5] I use the United Nations classification of sub-Saharan countries. However, I include Sudan in the analysis because it is geographically part of sub-Saharan Africa and contains a number groups partitioned between Sudan and sub-Saharan countries.

[6] Because Western Sahara is a disputed territory I exclude it from this border analysis.

Figure 2: Language Groups

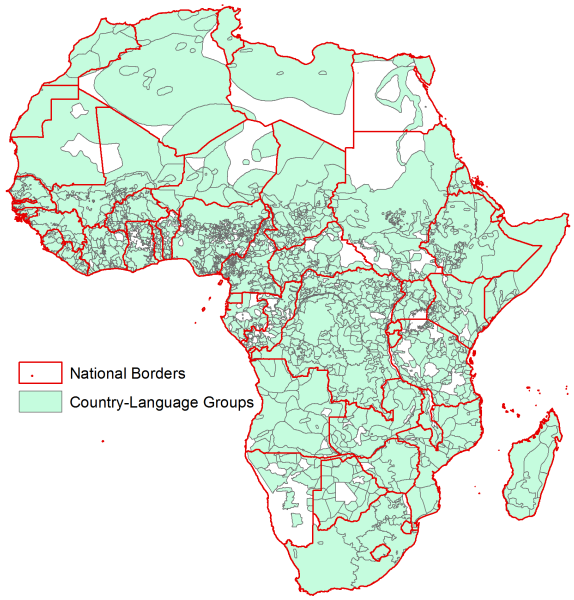
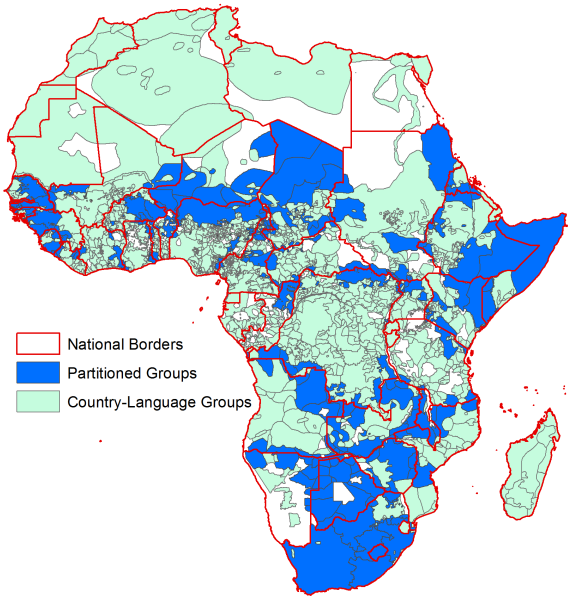


Figure 3: Language Partitions



2.2 Satellite Imagery of Night Light Luminosity

Satellite imagery of night light luminosity come from the National Oceanic and Atmospheric Administration's (NOAA) National Geophysical Data Center. Many others have used these data because of two features: night lights data exhibit a strong empirical relationship with GDP per capita and other measures of living standards (Henderson et al., 2012), and because these data are available at a spatial resolution of 30-arc seconds (approximately 1 square kilometre).[7] The fine resolution of these lights data facilitates a proxy measure of GDP per capita at any desired level of spatial aggregation. Because I require a measure of economic activity at the country-language group level – a level of aggregation where no official data on economic output exists – the availability of these data is indispensable to this study.

The yearly composite of night light luminosity is constructed by NOAA using daily images taken from U.S. Department of Defense weather satellites that circle the earth 14 times a day. These satellites observe every location on earth every night sometime between 20:30 and 22:00. Before distributing these data publicly, NOAA scientists remove observations contaminated by strong sources of natural light, e.g., the summer months when the sun sets late, light activity related to the northern and southern lights, forest fires, etc. All daily images that pass this screening process are then averaged for the entire year producing a

[7] Hodler and Raschky (2014) also show there is a strong empirical relationship between these night lights data and GDP at the subnational administrative region. Michalopoulos and Papaioannou (2014) further validate the use of night lights in Africa as a proxy measure of development with evidence that light intensity correlates strongly with individual-level data on electrification, presence of sewage systems, access to piped water and education.

satellite-year dataset for the time period 1992 to 2013. Light intensity receives a value of 0 to 63 at a resolution of 30-arc seconds. The result is a measure of night light intensity that only reflects human (economic) activity.[8]

Using these data I construct a panel of average luminosity for each country-language group partition. I use the Africa Albers Equal Area Conic projection to minimize distortion across the area dimension before calculating the average light luminosity of each country-language group polygon in each year.[9] I follow Michalopoulos and Papaioannou (2013, 2014) and Hodler and Raschky (2014) in adding 0.01 to the log transformation of the lights data because roughly 40% of these data have a value of zero in the benchmark sample. Doing so helps correct for the non-normal nature of the data and preserves sample size, and allows for a (near) semi-elasticity interpretation of the benchmark empirical model.

2.3 Assignment of a Leader's Ethnolinguistic Identity

There are 106 leaders to assign an ethnolinguistic identity for my sample of 35 countries between 1992-2013. The challenge of mapping ethnicity to language is that, in some instances, a single ethnic group speaks many languages. Because African language groups are often resident of well-defined territories (Lewis, 2009), an ethnolinguistic identity is typically attached to a person's birthplace (Batibo, 2005). As a first step towards assignment I locate the birthplace of a leader and collect latitude and longitude coordinates for each birthplace from www.latlong.net. I project these coordinates onto the Ethnologue map of Africa to back out the language group associated with each leader's birthplace.[10] I exclude leaders born abroad (4 leaders) since their ethnolinguistic group is not home to the country they govern.[11] Second, I identify a leader's ethnic identity using data from Dreher et al. (2015) and Francois et al. (2015), and in the few instances where neither source reports the ethnicity of a leader I fill in the gap using a country's Historical Dictionary. Finally, I take the following steps to assign a leader's ethnolinguistic identity using these data:

Step 1: I compare the birthplace linguistic identity with the ethnic identity for each of the 102 leaders. In 56.9 percent of the sample the name of the birth language and ethnic identity are equivalent (58 leaders). For these leaders the assignment is unambiguous.

[8] Henderson et al. (2012, p. 998) provide a thorough introduction to the NOAA night lights data.

[9] In some years data is available for two separate satellite, and in all such cases the correlation between the two is greater than 99 percent in my sample. To remove choice on the matter I use an average of both.

[10] Because most leaders enter/exit office mid-year, I assign the incumbent leader as whomever is in power on December 31st of the transition year. Hence, by assumption I drop any leader who exited office the same year she entered office because she was neither in power the previous year or December 31st of the transition year.

[11] These leaders include Ian Khama (Botswana), Francois Bozize Yangouvonda (Central African Republic), Nicephore Soglo (Benin), and Rupiah Banda (Zambia).

Step 2: For the remaining sample of unmatched leaders, I check if the birthplace language is a language spoken by the leader's ethnic group. In 12.7 percent of the sample this is true (13 leaders); I assign the birthplace language as the leader's ethnolinguistic identity.

Step 3: For the remaining 30.4 percent of unmatched leaders (31 leaders), the birthplace identity does not correspond to their ethnic identity. This is especially true for leaders born in a major city. For these leaders I drop the birthplace identity and map the ethnicity of a leader to a single language using the three-step assignment rule outlined in Appendix C.

2.4 Linguistic Similarity

Estimating linguistic similarity is difficult because languages can differ in a variety of ways, including vocabulary, pronunciation, grammar, syntax, phonetics and more. One common approach is to use a measure of the shared branches on a language tree as an approximation of linguistic similarity. Known as cladistic similarity, this measure was introduced to economists by Fearon and Laitin (1999), popularized by Fearon (2003) and has since become the convention.[12] The idea behind the cladistic approach is that two languages with a large number of shared nodes – and thus a recent splitting from a common ancestor – will be similar in terms of language because of their common ancestry. The data most commonly used is Fearon's (2003) cladistic measure of linguistic similarity, constructed using the Ethnologue's phylogenetic language tree. A cladistic measure is attractive because linguistic similarity is easily computed for any language pair, since language trees exist for virtually all known world language families (Lewis, 2009). See Appendix B for a formal definition of this measure.

My preferred measure is a computerized lexicostatistical measure of linguistic similarity developed by the Automatic Similarity Judgement Program (ASJP).[13] As a percentage estimate of a language pair's cognate words (i.e., words that share a common linguistic origin), the lexicostatistical method is a measure of the phonological similarity between two languages. Hence, a lexicostatistical measure can be thought of as a proxy for the ancestral relationship between two groups, or an implicit measure of the set of shared ancestral and cultural traits that are important to group identity.

The ASJP Database (Version 16) consists of 4401 language lists, where each list contains the same 40 implied meanings (i.e., words) for comparison across languages. The ASJP

[12] For example, Guiso et al. (2009); Spolaore and Wacziarg (2009); Desmet et al. (2012); Esteban et al. (2012) and Gomes (2014) all use a cladistic approach, among others.

[13] The lexicostatistical measure I use in this paper has been used to study factor flows in international trade (Isphording and Otten, 2013), job satisfaction of linguistically distinct migrants (Bloemen, 2013), language acquisition of migrants (Isphording and Otten, 2014), and the role of language in the flow of ideas (Dickens, 2016). See (Ginsburgh and Weber, 2016) for a discussion of this and other measures of linguistic distance.

research team has transcribed these lists into a standardized orthography called ASJPcode, a phonetic ASCII alphabet consisting of 34 consonants and 7 vowels. A standardized alphabet restricts variation across languages to phonological differences. Meanings are then transcribed according to pronunciation before language differences are estimated.[14]

Then for each language pair of interest I run the Levenshtein distance algorithm on the respective language lists, which calculates the minimum number of edits necessary to translate the spelling of each word from one language to another. To correct for the fact that longer words will demand more edits, each distance is divided by the length of the translated word. This normalization yields a percentage estimate of dissimilarity, which is measured across the unit interval. The average distance of a language pair is calculated by averaging across the distance estimates of all 40 words. By this procedure I estimate the linguistic distance of a language pair vis-à-vis the vocabulary dimension.

A second normalization procedure is used to adjust for the accidental similarity of two languages (Wichmann et al., 2010). This normalization accounts for similar ordering and frequency of characters that are the result of chance and independent of a word's meaning. Finally, I define the lexicostatistical similarity of a language pair as one minus this normalized distance. For a formal definition of this measure, I direct to reader to Appendix B.

The main advantage of the lexicostatistical approach is that it measures similarity in a more continuous way than the cladistic approach. Because the lexicostatistical method explicitly identifies the phonological differences of a language pair, there is far more observable variation in a measure of lexicostatistical similarity than cladistic similarity. The cladistic approach is a coarse measure of similarity because data dispersion is limited to 15 unique values, the maximum number of language family classifications in the Ethnologue.

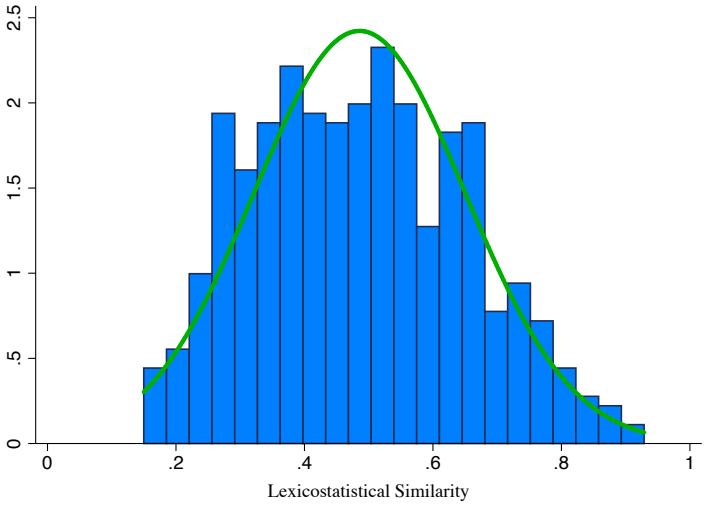
To illustrate this point, consider language pairs that share a common parent language on the Ethnologue language tree. Let these language pairs be known as siblings. All sibling pairs share the maximum number of tree nodes, and have no differences in cladistic similarity between them, but they do exhibit substantial variation in lexicostatistical similarity. To make this point clear, I plot the distribution of lexicostatistical similarities among all African sibling language pairs in Figure 4. This highlights the sizeable dispersion in lexicostatistical similarities among sibling language pairs.

2.4.1 Linguistic Similarity of Leaders and Language Groups

My independent variable of interest is a measure of bilateral linguistic similarity between each country-language group partition and the ethnolinguistic identity of the country's national

[14]For example, the French word for *you* is *vous*, and is encoded using ASJPcode as *vu* to reflect its pronunciation.

Figure 4: Lexicostatistical Similarities Among Sibling Language Pairs



This figure establishes the additional variation introduced by a lexicostatistical measure of linguistic similarity that is not observable with a cladistic measure of similarity. The histogram plots the estimates of lexicostatistical similarity among sibling language pairs for all of Africa ($n = 1,241$). Sibling language pairs are those that share a parent language on the Ethnologue language tree, which by definition implies that among sibling language pairs there is no observable variation in cladistic similarity.

leader. Because the computerized lexicostatistical method requires a word list for each language of interest, I am limited to working with languages that have lists made available by the ASJP research team. Of the 227 language groups in the full set of partitions I match 163 in the benchmark regression (72%), failing the rest either because the leader's birth language list is unavailable or the partition language list is unavailable. Furthermore, 11 out of the 102 leaders ethnolinguistic identities lack an ASJP language list and are excluded from the analysis. I address the possibility of sample selection in Section 4. The result is an (unbalanced) panel of lexicostatistical similarity between partitioned language groups and their national leader for the years 1992-2013.[15] Figure 3 colour codes these groups.[16]

2.5 Patterns in the Data

Table 1 reports descriptive statistics for the night lights and language data. For completeness, I have included a cladistic measure of similarity and a binary measure of coethnicity.[17] The

[15]See Appendix A for a complete list of included countries and language groups.

[16]The only other lexicostatistical data available for a large number of languages is from Dyen et al. (1992), which is restricted to Indo-European languages only – none of which are native to Africa.

[17]I use the term coethnicity to be consistent with the literature, but a better name would be coethnolinguists since I define coethnicity equal to one when a leader's ethnolinguistic identity is the same as a partitioned *language* group.

Table 1: Descriptive Statistics

	Obs.	Mean	Std. Dev.	Min	Max
ln(0.01 + night lights)	6,610	-3.496	1.423	-4.605	1.515
Lexicostatistical similarity	6,610	0.193	0.230	0.000	1.000
Cladistic similarity	6,610	0.409	0.330	0.000	1.000
Coethnicity	6,610	0.047	0.212	0.000	1.000

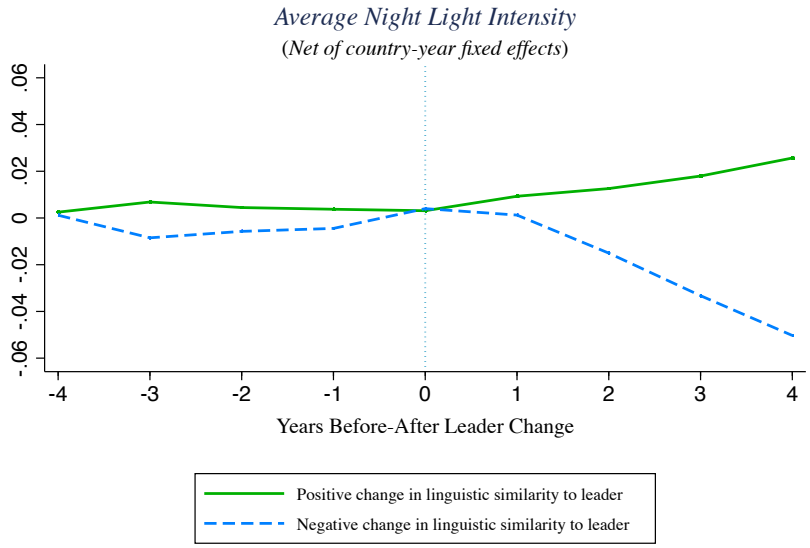
This table reports descriptive statistics for the main variables of interest used in the benchmark empirical analysis of partitioned language groups in Africa. The unit of observation is a language group l that resides in country c in year t. See Appendix A for a description of the data and sources.

Table 2: Means of Linguistic Similarity Above-Below Median Night Lights

	Observations	Above Median Luminosity	Below Median Luminosity	Difference
Panel A: Full Sample				
Lexicostatistical similarity	6,610	0.245	0.142	0.104***
		(0.005)	(0.003)	(0.006)
Cladistic similarity	6,610	0.478	0.341	0.137***
		(0.006)	(0.005)	(0.008)
Coethnicity	6,610	0.078	0.016	0.062***
		(0.005)	(0.002)	(0.005)
Panel B: Non-Coethnic Regions Only				
Lexicostatistical similarity	6,298	0.182	0.125	0.057***
		(0.003)	(0.002)	(0.004)
Cladistic similarity	6,298	0.435	0.325	0.110***
		(0.006)	(0.005)	(0.008)

This table reports differences in means for various measures of linguistic similarity. Language groups are separated by the median value of night lights into "above" and "below" groups for each sample. The full sample consists of 6,610 observations and the non-coethnic subsample consists of 6,298 observations. Standard errors are reported in parentheses.

Figure 5: Pre-Post Leadership Change



This figure plots the before and after effects of a change in leadership on average night light luminosity. The green solid line depicts luminosity in the 4 years leading up to a change in leadership and the 4 years following an increase in linguistic similarity. The blue dashed line depicts the same for country-language groups that experienced a decrease in similarity after a change in leadership. Average night light luminosity is the residual light variation net of country-year effects to account for different years of leadership change across countries.

mean value of lexicostatistical linguistic similarity says that country-language groups are 19.3 percent similar to their national leader on average, and the mean value of cladistic similarity implies 40.9 percent similarity. The mean value of coethnicity says that 4.7 percent of the benchmark sample is coethnic to their national leader.[18]

In Table 2 I preview the empirical results by splitting the sample by the median value of night lights and test for differences in average linguistic similarity. Panel A reports mean differences in the benchmark sample for all three similarity measures. Take, for example, the mean difference in lexicostatistical similarity: language groups who emit night light above the median value are on average 10.4 percent more similar to their national leader than those below the median value. This difference is highly significant, with a reported p-value of 0.000. The same pattern is true irrespective of the measure of linguistic similarity. These findings are consistent with my proposed hypothesis of ethnolinguistic favoritism, where language groups are better off the more similar they are to their national leader.

Panel B repeats this exercise in all non-coethnic sample observations. As stated in the introduction, if relative groups differences matter outside of coethnic relationships, then the

[18]Table A6 reports a complete set of descriptive statistics used throughout this analysis.

data should tell me that similarity matters among non-coethnics. This is exactly what I find: the average similarity among non-coethnic language groups above and below the median night lights value is significantly different than zero. While I reserve more conclusive statements for the regression analysis, this suggests that linguistic similarity provides significant variation that is unobservable in the conventional binary framework. Together these results show that night lights and linguistic similarity are positively related, or that on average a language group is increasingly better off the more linguistically similar they are to the birth language of their national leader. The significant pairwise correlation of 0.30 between light intensity and lexicostatistical similarity is also suggestive of this positive relationship (correlation not shown here).

I also plot average luminosity before and after a leadership change in Figure 5, separating groups who experience an increase in lexicostatistical similarity from those that experience a decrease. I construct a "treatment" time scale that takes a value of 0 in the year of a leadership change, and plot the residual light variation net of country-year effects to account for different years of leadership changes. I plot these data for the 4 years leading up to a change and the 4 years following. It is reassuring for identification that there is little observed change in night light activity in the years leading up to a change in leadership. Yet shortly after a leadership change there is a noticeable increase in night lights in regions that experienced an increase in linguistic similarity to the leader (solid green line), and a large drop in average night lights in regions that experienced a decrease in similarity (dashed blue line). Hence, Figure 5 is a clean visualization of favoritism across linguistic lines.[19]

3 Empirical Model

The main objective of this empirical analysis is to test the hypothesis that a language group that is linguistically similar to the ethnolinguistic identity of the national leader will be better off than a group whose language is relatively more distant. To do this I use a triple difference-in-differences estimator:

$$y_{c,l,t} = \gamma_{c,l} + \lambda_{c,t} + \theta_{l,t} + x'_{c,l,t}\,\Phi + \beta LS_{c,l,t-1} + \varepsilon_{c,l,t}. \tag{1}$$

The dependent variable $y_{c,l,t}$ is the night lights measure of economic activity for language

[19]The number of observations used to calculate the average night lights in either group varies by years. The nature of the data presents two challenges in constructing a standard treatment time scale. First, in some instances there is more than one leadership change in the shown 8-year interval. Second, and in consequence of the first point, two leadership changes over the 8-year interval do not always result in consistent positive or negative changes of similarity.

group l in country c in year t. As the dependent variable I follow the literature and take the aforementioned log transformation of night lights such that $y_{c,l,t} \equiv \ln(0.01 + NightLights_{c,l,t})$.

$LS_{c,l,t-1}$, the variable of interest, measures the linguistic similarity between language group l in country c and the ethnolinguistic identity of country c's political leader in year $t-1$. I lag linguistic similarity because of an expected delay between the decision to allocate public funds to a region and the actual allocation of those goods (Hodler and Raschky, 2014), and an expected delay between the actual allocation of public funds and the resulting regional increase in night light production.

$X_{c,l,t}$ is a vector of controls including the (logged) average of population density for each country-language, and the (logged) geodesic distance between language group l and the language group associated with the leader of country c.[20] I also include a variety of geographic endowment controls in $x_{c,l,t}$: two indicator variables for the presence of oil and diamond reserves in both the leader and language group regions, as well as the absolute difference in elevation, ruggedness, precipitation, average temperature and the caloric suitability index (agricultural quality). These additional controls account for the possibility that national projects that are beneficial to the leader's region because of a particular geographic characteristic might also benefit other regions of similar character.[21] $\gamma_{c,l}$ are country-language group fixed effects, $\lambda_{c,t}$ are country-year fixed effects and $\theta_{l,t}$ are language-year fixed effects.[22] In all specifications I adjust standard errors for clustering in country-language groups.[23]

3.1 Identification of Linguistic Similarity

In order to identify the effect of linguistic similarity it is necessary that the placement of national borders are not the result of local economic conditions or any factor that reflects the well-being of a language group. Indeed, national borders are a historical by-product of the Scramble for Africa. The use of straight lines prevailed when drawing borders in Africa because the Berlin Conference of 1884-85 legitimized claims of colonial sovereignty without pre-existing territorial occupation, rendering knowledge of pre-colonial boundaries inconsequential (Englebert et al., 2002). The result was a reluctance by colonialists to

[20] Population density data comes from the Gridded Population of the World. Because population density data is only available in 5-year intervals (i.e., 1990, 1995, 2000, 2005 and 2010), I assume the density to be constant throughout the unobserved intermediate years.

[21] See Appendix A for more details on data definitions and sources.

[22] In my benchmark sample $\gamma_{c,l}$ represents 355 fixed effects, $\lambda_{c,t}$ represents 691 fixed effects and $\theta_{l,t}$ represents 3044 fixed effects.

[23] Given that the benchmark sample has only 35 countries, I choose not to adjust standard errors for two-dimensional clustering within language groups and countries (Cameron et al., 2011). While the benchmark results are qualitatively similar when two-way clustering, I follow Kezdi's (2004) rule of thumb that at least 50 clusters are needed for accurate inference.

respect traditional boundaries when drawing borders (Herbst, 2000). Evidence of this is still seen today, where group partitions do not correlate with geography and natural resources (Michalopoulos and Papaioannou, 2016) and nearly 80% of all African borders follow lines of latitude and longitude – an amount larger than any other continent in the world (Alesina et al., 2011).[24]

It is the arbitrary design of African political borders that forms the basis of my identification strategy. The ethnolinguistic identity of a national leader varies by country, so group partitioning generates exogenous within-group variation in terms of that group's linguistic similarity to their leader. This strategy is similar to Michalopoulos and Papaioannou (2014), though a key difference is that I construct a panel of partitioned groups rather than a cross-section, so the relative similarity within a partitioned group also varies over time as new leaders come to power. This is instrumental to identification: by including the three sets of fixed effects discussed in the previous section, I absorb all the variation in the data with the exception of time-variation at the country-language group level. $\gamma_{c,l}$ and $\lambda_{c,t}$ respectively difference out time-invariant country-group trends and country-time trends that are differentially affecting the same group on each side of the border. The inclusion of $\theta_{l,t}$ only allows for within-group time-variation that comes from changes in leadership. Hence, with my set-up in equation (1), I am estimating the effect of linguistic similarity off of changes in the incoming leader's ethnolinguistic identity.

In my benchmark sample this variation comes from 35 leadership changes: the within transformation of $\theta_{l,t}$ implies that a leadership change in one country varies the mean similarity of a partition in that country and all other fragments of that partition in neighbouring countries. In other words, the relative similarity within a partitioned group varies with a leadership change on either side of the border. This amounts to 485 unique relative similarities observed between 1992-2013 in my data.

4 Benchmark Results

Table 3 reports nine different estimates: three versions of equation (1) for each of the three linguistic similarity measures. For each measure of similarity, I report estimates (i) without any covariates (columns 1-3), (ii) estimates that control for log population density and the logged geodesic distance between each partitioned group and the corresponding leader's group (columns 4-6), and (iii) the full set of covariates I outlined in Section 3 (columns 7-9).

[24]See Englebert et al. (2002) and Michalopoulos and Papaioannou (2014, 2016) for a detailed discussion on the arbitrary design of African borders.

Table 3: Benchmark Regressions Using Various Measures of Linguistic Similarity

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.244**			0.297**			0.305***		
	(0.112)			(0.120)			(0.116)		
Cladistic similarity$_{t-1}$		0.221**			0.219**			0.185*	
		(0.104)			(0.102)			(0.103)	
Coethnic$_{t-1}$			0.130			0.139			0.168*
			(0.099)			(0.098)			(0.094)
Geographic controls	No	No	No	No	No	No	Yes	Yes	Yes
Distance & population density	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Adjusted R^2	0.925	0.925	0.925	0.925	0.925	0.925	0.926	0.925	0.925
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports benchmark estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t - 1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if both region contains diamond and oil deposits. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Hereafter I will refer to columns 7-9 as my benchmark specification.[25]

Consistent with my hypothesis of ethnolinguistic favoritism, all nine coefficients are positive and my preferred measure of lexicostatistical similarity is always statistically significant. Because variation is coming from changes in the ethnic identity of a leader, the interpretation of these findings is that a group's well-being is increasing in their ethnic similarity to the leader. To give economic meaning to these estimates, consider the benchmark estimate of lexicostatistical similarity in column 7. Using the rule of thumb that the estimated elasticity of GDP per capita with respect to night lights is 0.3 (Henderson et al., 2012), the point estimate of 0.305 implies that a standard deviation increase in linguistic similarity (23 percent change) yields a 2.1 percent increase in regional GDP per capita, an economically significant effect.[26]

I also provide estimates for cladistic similarity and coethnicity to see how these alternative measures compare to lexicostatistical similarity. For my benchmark estimates both coefficients are positive and statistically significant, albeit only at the 10 percent level. Notice that in all iterations of equation (1), the magnitude and precision of the estimate is monotonically increasing in the measured continuity of linguistic similarity. This suggests that the observable variation among non-coethnic groups assists in identifying patterns of ethnic favoritism in Africa, and thus speaks the virtue of the lexicostatistical measure.

In Table 4 I report estimates from a series of horse race regressions. With these estimates I show that the lexicostatistical measure is better at identifying patterns of favoritism than the alternative measures of similarity. In columns 1-4, I report estimates for all possible pairings of the three measures of similarity. Because all three measures of similarity are highly correlated with each other, and for coethnic observations are equivalent, the effect of lexicostatistical and cladistic similarity are estimated off of the additional variation these measures provide among non-coethnics. In all pairings the additional lexicostatistical variation is estimated to be statistically significant, despite the fact that the effect of coethnicity is not identifiable in these regressions. In column 3, cladistic similarity outperforms coethnicity in magnitude and precision, reaffirming the value of the addition variation it provides over a coethnic indicator, but is not estimated to be significantly different than zero.

To disentangle the effect of coethnicity from the benefits of similarity among non-coethnics, I define non-coethnic lexicostatistical similarity as $(1-\text{coethnic}_{t-1})\times$ lexicostatistical similarity, and equivalently for non-coethnic cladistic similarity. In other words, these non-coethnic similarity measures are equal to zero when the observed language group is coethnic to their

[25] See Table D1 for various other combinations of fixed effects specifications.

[26] The percentage change in GDP per capita $\approx$ percentage change in night lights $\times 0.3 = (\beta \times \Delta LS_{c,l,t-j}) \times 0.3 = 0.305 \times 0.230 \times 0.3 = 2.1\%$, assuming that $\ln(0.01 + \text{nightLights}_{c,l,t}) \approx \ln(\text{nightLights}_{c,l,t})$.

Table 4: Horse Race Regressions: Contrasting the Different Measures of Linguistic Similarity

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$						
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.345**	0.473**		0.591**		
	(0.165)	(0.227)		(0.291)		
Cladistic similarity$_{t-1}$	-0.046		0.151	-0.102		
	(0.146)		(0.125)	(0.150)		
Coethnic$_{t-1}$		-0.213	0.080	-0.249	0.260**	0.230**
		(0.202)	(0.114)	(0.211)	(0.106)	(0.110)
Non-coethnic lexicostatistical similarity$_{t-1}$					0.473**	
					(0.227)	
Non-coethnic cladistic similarity$_{t-1}$						0.151
						(0.125)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355
Countries	35	35	35	35	35	35
Language groups	163	163	163	163	163	163
Adjusted R^2	0.926	0.926	0.925	0.926	0.926	0.925
Observations	6,610	6,610	6,610	6,610	6,610	6,610

This table reports horse race regressions comparing each measure of linguistic similarity. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Non-coethnic lexicostatistical similarity and Non-coethnic cladistic similarity are constructed by interacting a dummy variable for non-coethnicity with Lexicostatistical similarity and Cladistic similarity, respectively. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

national leader, and otherwise equivalent to the respective measure of similarity. Combined with the coethnic measure, I can exploit the same variation I identify off of in columns 2 and 3 but load the effect of coethnicity onto the coethnic dummy variable.

Because it is intuitive that a leader is more inclined to favor her coethnics, I expect to see a strong significant effect of coethnicity beyond the effect found among non-coethnic groups. Indeed, column 5 indicates that coethnics are most favored with an estimated increase of 0.260 in average night light luminosity. While there is still an observable benefit from similarity among non-coethnics, the magnitude of the effect is roughly one quarter the size of the coethnic effect on average. With a sample mean of 0.146, non-coethnic lexicostatistical similarity yields an average increase of 0.069 ($= 0.146 \times 0.473$) in night light luminosity.[27]

I repeat this exercise with non-coethnic cladistic similarity and report the estimates in column (6). Once again I find the corresponding estimate for cladistic similarity from column (3) but can now identify the effect of coethnicity. The estimated coefficient for coethnicity is quite similar to the coethnic effect found in column (5), only now the additional variation coming from the cladistic measure is not enough to identify the effect of similarity among non-coethnic groups.

Taken together the results of Table 3 and Table 4 indicate that favoritism is most prominent among coethnics but also to a lesser extent among non-coethnics. These results also indicate that a continuous measure of lexicostatistical similarity provides valuable information that is not observable with a coethnic indicator variable. For the remainder of this section I proceed to test the robustness of the benchmark lexicostatistical estimate.

Anticipatory Effects

In this section I run of a series of tests of the identifying assumptions underlying my benchmark estimates. Column (1) of Table 5 reproduces the benchmark estimate of lexicostatistical similarity for comparison. In column (2) I show that the lagged measure of lexicostatistical similarity is not essential to my findings; contemporaneous lexicostatistical similarity is estimated to be positive and significant at the 5 percent level.

In column (3) I report an estimate of lexicostatistical similarity measured in period $t+1$. In this specification I'm estimating the effect of linguistic similarity off of the change in an incoming leader's ethnolinguistic group in the period before that leader comes to power. Should there be any pre-trends in the incoming leader's group, then this lead measure of

[27] By these estimates the threshold value of non-coethnic similarity is 0.550, above which would imply non-coethnics are better off than coethnics. The likelihood of measurement error in linguistic similarity implies this is a rather "fuzzy" threshold, and with only 2 percent of the benchmark sample above this threshold I find this result to be reassuring.

Table 5: Testing for Anticipatory Effects: Estimates Using Leads and Lags

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lexicostatistical similarity$_{t-1}$	0.305***			0.299***		0.249**	0.170**	0.131**
	(0.116)			(0.110)		(0.101)	(0.067)	(0.062)
Lexicostatistical similarity$_{t}$		0.495**			0.406**	0.242		0.214
		(0.204)			(0.205)	(0.183)		(0.134)
Lexicostatistical similarity$_{t+1}$			0.170	0.134	0.059	0.067		0.021
			(0.117)	(0.107)	(0.096)	(0.096)		(0.070)
Night lights$_{t-1}$							0.521***	0.506***
							(0.050)	(0.055)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	351
Countries	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	161
Adjusted R^2	0.926	0.926	0.930	0.930	0.930	0.930	0.947	0.950
Observations	6,610	6,474	6,121	6,121	6,084	6,084	6,315	5,785

This table reports a series of tests for anticipatory effects in the benchmark estimates. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. The same log transformation of the dependent variable is used for the lagged value of night lights, i.e., $\ln(0.01 + NightLights_{c,l,t-1})$. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

lexicostatistical similarity should be estimated significantly different than zero. I find no evidence of a pre-trend, which is reassuring for identification that the common trends assumption is satisfied. In column (4)-(6) I report estimates from horse race regressions between lead, contemporaneous and lagged lexicostatistical similarity. Again I find no evidence of a pre-trend in the lead variable. Together these findings confirm there are no anticipatory changes in night lights preceding a change in leadership, an observation consistent with Figure 5. Column (6) also indicates that lagged lexicostatistical similarity is a better predictor of favoritism than contemporaneous similarity, a finding that supports my decision to lag lexicostatistical similarity.

Next I re-estimate equation (1) with a lagged dependent variable. Identification rests on the assumption that leaders are not endogenously elected because of the economic success of their ethnolinguistic group prior to an election. I find no evidence of this as indicated by column (7) and (8). Lexicostatistical similarity is estimated to be positive and significant at the 5 percent level, albeit with a reduced magnitude. Hence, these results are reassuring that my benchmark estimates are not an outcome of any pre-transition changes in economic activity in a leader's ethnolinguistic group.

Migration

One additional concern with my identification strategy is cross-border migration. Suppose individuals who live near the border become coethnics of the neighboring country's leader. These individuals may choose to migrate in response to this spatial disequilibrium of similarity. While the cultural affinity of partitioned groups might ease the migration process, Oucho (2006) points out that migration restrictions throughout sub-Saharan Africa make this unlikely in a formal capacity, so this might only be an issue among undocumented migrants. Not only do undocumented migrants make up a small percentage of total migrants but those that do migrate tend to do so to trade and are temporary by definition (Oucho, 2006). To corroborate this anecdotal evidence, I also regress log population density on linguistic similarity in period $t-1$ and report the estimates in Table 6. If people are in fact migrating in response to leadership changes, I should observe corresponding changes in population density. These estimates also account for the possibility of within-country migration. In all specifications, the various measures of similarity are insignificant. Overall these estimates imply that changes in night lights within a partitioned group cannot be explained by movements of people to regions that are similar to the leader in terms of ethnolinguistic identity.

Table 6: Test for Migration Following Leadership Changes

Dependent Variable: ln(Population Density$_{c,l,t}$)			
	(1)	(2)	(3)
Lexicostatistical similarity$_{t-1}$	0.001		
	(0.019)		
Cladistic similarity$_{t-1}$		-0.027	
		(0.031)	
Coethnic$_{t-1}$			0.010
			(0.016)
Country-language fixed effects	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes
Clusters	355	355	355
Countries	35	35	35
Language groups	163	163	163
Adjusted R^2	0.999	0.999	0.999
Observations	6,610	6,610	6,610

This table reports estimates associating population density with linguistics similarity as a test for changes in population density following a change in a leader's ethnolinguistic identity. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Sample Selection

My inability to observe the lexicostatistical similarity of the 64 language groups without an ASJP language list raises the question whether these unobserved groups are systematically different than those in my benchmark sample. To address this concern I test for mean differences in key observables and report these differences in Table 7.

First I show that there is no difference in the average night light luminosity between in- and out-of-sample partitioned language groups. I also show that there is no difference between the cladistic similarity of in- and out-of-sample groups. These two results are reassuring that both sets of partitioned groups are comparable in terms of economic activity and proximity to their leader.

To the contrary, I show that in-sample groups reside in countries that are, on average, more democratic, more competitive politically, have more constraints on the executive, and

Table 7: Selection into Lexicostatistical Language Lists

		Partitioned Language Groups		
	Observations	Benchmark Sample Mean	Out of Sample Mean	Difference
ln(0.01 + night lights)	11,869	-3.487	-3.505	0.018
		(0.018)	(0.022)	(0.028)
Cladistic similarity	11,869	0.276	0.272	0.004
		(0.004)	(0.004)	(0.005)
Level of democracy (Polity2)	11,822	0.677	0.319	0.358***
		(0.059)	(0.062)	(0.086)
Political competition	10,854	6.180	5.940	0.239***
		(0.032)	(0.033)	(0.046)
Executive constraints	10,854	3.634	3.368	0.266***
		(0.022)	(0.022)	(0.032)
Openness of executive recruitment	10,854	2.756	2.556	0.200***
		(0.024)	(0.028)	(0.036)
Competitiveness of executive recruitment	10,854	1.283	1.208	0.075***
		(0.014)	(0.015)	(0.021)

This table tests for selection into the available language lists in the ASJP database. The full sample of partitioned language groups are separated by those that I observe in my benchmark dataset and those that I do not because of missing ASJP language lists. Standard errors are reported in parentheses.

are more open and competitive in the recruitment of executives. Should there be an in-sample selection bias, these institutional mean differences suggest that my estimates would be biased towards zero, given the evidence that a well-functioning democracy mitigates the extent of ethnic favoritism (Burgess et al., 2015) and regional favoritism (Hodler and Raschky, 2014)

Robustness Checks

I also show that the results are robust to a variety of specifications and estimators. I report and discuss each robustness check in Appendix D. In particular, I show that the results are similar when:

- I reproduce my benchmark estimates with additional controls for malaria and land suitability for agriculture. Because these data are only available at a $0.5^\circ \times 0.5^\circ$ spatial resolution (approx. 111 km $\times$ 111 km), I exclude them from my benchmark estimates to avoid losing observations where a pixel is larger than a language group partition (Table D2).

- I check that measurement error coming from ambiguous assignment of a leader's ethnolinguistic identity does not explain my benchmark results, particularly the finding that favoritism exists among non-coethnics (Table D3).

- I reproduce my benchmark estimates on a balanced panel of 84 ethnolinguistic groups partitioned across 23 countries (Table D4).

- I re-estimate equation (1) and weight the estimates by the Ethnologue population of each language group (Table D5). The idea here is to correct for possible heteroskedasticity: the measure of night light intensity is an average within each country-language group, and it is likely to have more variance in places where the population is small.

- I also provide estimates with two alternative transformations of the night lights data to show that my benchmark lexicostatistical estimate is not an outcome of the aforementioned log transformation (Table D6).

4.1 What Drives Favoritism?

In this section I test for heterogeneity across a variety of potential channels to better understand what drives favoritism. In Table 8 I study the dynamics of my benchmark findings by account for the possibility that the extent of favoritism is a function of the time a leader has held office. In column 1, I report estimates of an augmented equation (1) that includes an interaction between linguistic similarity and a count of the years a leader has held office. The interaction term enters positive and statistically significant, indicating that favoritism is an increasing function of the years a leader has held office.

I also construct a set of indicator variables at 5-year intervals to explore the non-linearities further. Column 2 reports these estimates. All coefficients are positive and the magnitude of effect is increasing in the length of tenure, however there is no significant effect associated with the first five leaders of leadership. Taken together, Table 8 indicates that the extent of ethnolinguistic favoritism is an increasing function of a leader's incumbency. In a continent where multi-decade presidencies are not uncommon (e.g., Jose Eduardo dos Santos in Angola or Robert Mugabe in Zimbabwe), it should come as no surprise that favoritism is so rampant.

I also check for heterogeneous effects across seven other measures: the level of democracy (Padro i Miquel, 2007; Burgess et al., 2015), language group Ethnologue population shares (Francois et al., 2015), distance to the capital from a group's centroid (Michalopoulos and Papaioannou, 2014), distance to the nearest coast from a group's centroid (Nunn, 2008; Nunn and Wantchekon, 2011), presence of an oil reserve and diamond mine within the territory of the country-language group (Jensen and Wantchekon, 2004). Table 9 reports these estimates.

Table 8: The Dynamics of Ethnolinguistic Favoritism

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$		
	(1)	(2)
Lexicostatistical similarity$_{t-1}$	0.072	
	(0.160)	
Lexicostatistical similarity$_{t-1}$	0.027*	
$\times$ Years in office$_{t-1}$	(0.016)	
Lexicostatistical similarity$_{t-1}$		0.118
$\times$ 1(Years in office$_{t-1} \leq 5$)		(0.129)
Lexicostatistical similarity$_{t-1}$		0.325*
$\times$ 1($5 <$ Years in office$_{t-1} \leq 10$)		(0.170)
Lexicostatistical similarity$_{t-1}$		0.561***
$\times$ 1($10 <$ Years in office$_{t-1} \leq 15$)		(0.197)
Lexicostatistical similarity$_{t-1}$		0.555**
$\times$ 1($15 <$ Years in office$_{t-1} \leq 20$)		(0.233)
Lexicostatistical similarity$_{t-1}$		0.689**
$\times$ 1($20 <$ Years in office$_{t-1}$)		(0.347)
Geographic controls	Yes	Yes
Distance & population density	Yes	Yes
Language-year fixed effects	Yes	Yes
Country-language fixed effects	Yes	Yes
Country-year fixed effects	Yes	Yes
Clusters	355	355
Countries	35	35
Language groups	163	163
Adjusted R^2	0.926	0.926
Observations	6,610	6,610

This table reports estimates of the dynamics of ethnolinguistic favoritism. The unit of observation is a language group l in country c in the specified year. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the ethnolinguistic identity of the national leader. Current years in office is a count variable of the years the incumbent leader has been in power, and total years in office measures the total years the incumbent leader will remain in power. Quartile measures relate to current years in office. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 9: Benchmark Regressions with Heterogeneous Effects

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$						
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.298**	0.407**	0.379**	0.327*	0.305***	0.397***
	(0.127)	(0.161)	(0.174)	(0.190)	(0.116)	(0.135)
Lexicostatistical similarity$_{t-1}$	-0.005					
× Democracy$_{t-1}$	(0.020)					
Lexicostatistical similarity$_{t-1}$		-0.610				
× Population share		(0.533)				
Lexicostatistical similarity$_{t-1}$			-0.000			
× Distance to the capital			(0.000)			
Lexicostatistical similarity$_{t-1}$				-0.000		
× Distance to the coast				(0.000)		
Lexicostatistical similarity$_{t-1}$					0.232	
× Oil reserve					(1.140)	
Lexicostatistical similarity$_{t-1}$						-0.336*
× Diamond mine						(0.190)
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355
Countries	35	35	35	35	35	35
Language groups	163	163	163	163	163	163
Adjusted R^2	0.927	0.926	0.926	0.926	0.926	0.926
Observations	6,540	6,610	6,610	6,610	6,610	6,610

This table reports a series of tests for heterogeneous effects in the benchmark estimates. Average night light intensity is measured in language group l of country c in year t, and lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. All control variables are described in Table 3. Democracy is the polity2 score of democracy for the country in which a group resides, geodesic distances are measured in kilometres from a group's centroid to the capital city and the nearest coast, oil reserve and diamond mine represent indicators variables at the group level. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

The analysis reveals little evidence of heterogeneity. One explanation for a lack of heterogeneity is that these different channels are only relevant in some countries and do not generalize to the 35-country sample I use here. Another possible explanation is that the rich set of fixed effects in each regression absorb much of the important variation. For example, in column (1), I find that democracy has a mitigating effect on the extent of observed favoritism, but this effect is not statistically significant. While the intuition is consistent with Burgess et al. (2015), the lack of precision likely comes from the fact that country-year fixed effects account for the level effect of democracy, and the residual variation is not significant enough to identify any meaningful effect. A similar explanation applies to the remaining variables, where country-language fixed effects absorb the level effect for each because of the time invariance of these group-level measures.

However, there is some evidence of heterogeneity in terms of a diamond mine being present within a country-language group. The negative coefficient implies favoritism is less prevalent in regions where diamond mines exist. On interpretation is that the presence of diamonds creates wealth, and the resulting development may reduce the material importance of patronage to the region. Yet the lack of heterogeneity in oil reserves does not corroborate this story, so I leave a more concrete analysis of why diamond mines might constrain favoritism to future research.

5 How Is Patronage Distributed?

In this section I develop a within-group set-up similar to the previous section using individual-level data from the Demographic and Health Survey (DHS). Exploiting the same source of variation with different data serves as an additional robustness check of my benchmark analysis. The DHS data also allows me to explore how patronage is distributed. In particular, I use data on an individual's location and ethnolinguistic identity to construct two measures of lexicostatistical similarity: locational and individual similarity. I define individual similarity as the lexicostatistical similarity of the leader to each respondent's ethnolinguistic identity. To assign a locational language I use the Ethnologue language map and individual location coordinates to determine the language group associated with an individual's location. I define locational similarity as the lexicostatistical similarity of a leader to the respondent's locational language. Because these measures do not always coincide, I can jointly estimate both channels to determine the relative importance of being similar to the leader versus living in a location with an attached identity similar to the leader.

5.1 DHS Individual-Level Data

I collect data from the Demographic and Health Surveys (DHS) for 13 African countries.[28] For each country I pool both the male and female samples for each wave, and when separately provided, I merge the wealth index dataset for that year. To replicate the same variation I use in my benchmark estimates, I choose DHS countries and survey waves as follows:

(1) I identify all DHS country-waves that include latitude and longitude coordinates for each survey cluster as well as information on a respondent's home language and/or ethnic identity.

(2) I identify all language groups that are partitioned across contiguous country pairs in the DHS database that also possess the necessary information noted in (1).

(3) For each partitioned language group identified in (2) I keep those that possess at least 2 consecutive surveys from the same set of DHS waves.

Next I project the latitude and longitude coordinates for each survey cluster onto the Ethnologue language map and back out the language group associated with that location.[29] I assign this language as the locational language for that cluster and construct a measure of locational similarity as the lexicostatistical similarity of that region to the incumbent leader.

To measure individual similarity I use data on the language a respondent speaks at home, and when not available data on their ethnicity. I describe the mapping between ethnicity and language in detail in Appendix C. I construct a measure of individual similarity as the lexicostatistical similarity between the home language of an individual and the ethnolinguistic identity of their national leader. To be consistent with my benchmark model, I measure locational and individual linguistic similarity to the national leader in year $t-1$.

The result is 33 DHS country-waves, 13 countries and 11 country pairs, with 20 partitioned language groups. Having at least 2 consecutive survey waves for each partitioned group allows for a set-up similar to my benchmark model, where within-group time variation comes from leadership changes across waves. One important difference from my benchmark set-up is that for 3 countries I only observe a single partitioned language group, meaning that country-location-language fixed effects are not applicable in this context.

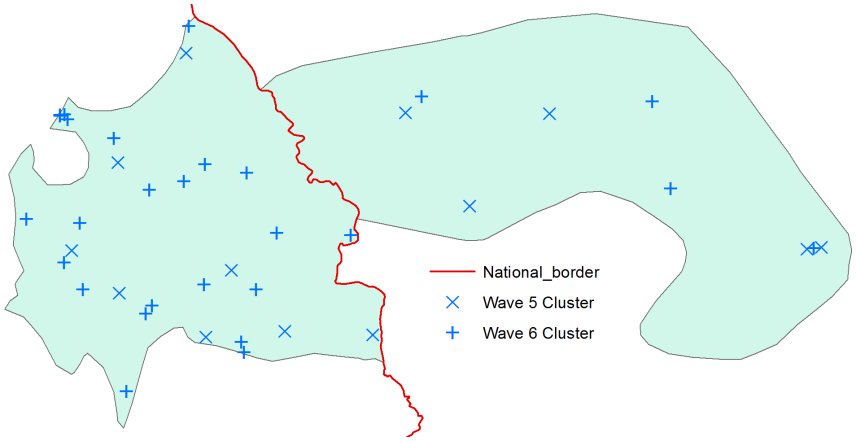
Among the 56,455 respondents for whom I successfully match both locational and individual languages, I find that 55.9 percent reside in their ethnolinguistic homeland.[30] This

[28] See Appendix A for a list of countries and a detailed discussion of all DHS data.

[29] In instances of overlapping language groups, I assign the largest group in terms of population

[30] Nunn and Wantchekon (2011) also find that 55 percent of respondents in the 2005 Afrobarometer reside in their ethnolinguistic homeland. The consistency across datasets is quite remarkable since only 7 out of the 13 countries used in this paper overlap with the Afrobarometer data in Nunn and Wantchekon (2011).

Figure 6: DHS Clusters Across Waves in the Kuranko Language Group Partition



This figure documents the spatial distribution of DHS enumeration clusters in the partitioned Kuranko language group in Sierra Leone (west of the border) and Guinea (east of the border). Variation between individual and locational lexicostatistical similarity comes from the 40 percent of respondents who identify with an ethnolinguistic group different than the Kuranko.

finding corroborates the implicit assumption made in the regional-level analysis that the majority of a language region's inhabitants are native to that region. At the same time, having 44.1 percent of respondents be non-native to their location allows me to exploit variation in individual and location similarity to separately estimate the two effects off of leadership changes.[31]

Consider, as an example, the Kuranko language group partitioned across Guinea and Sierra Leone. Figure 6 depicts the spatial distribution of Kuranko survey clusters by wave. For each survey respondent living in one of these clusters I assign the Kuranko language as their locational language, despite the fact that only 60.1 percent of respondents report Kuranko as their ethnolinguistic identity. Among the remaining 39.9 percent of respondents in the Kuranko region there are 9 other reported ethnolinguistic identities. Take the 117 Sierra Leoneans living in the Kuranko region who identify as Themne – the ethnicity/language of president Ernest Bai Koroma. The inclusion of individual similarity allows me to ask if Themne respondents benefit from coethnicity – and similarity more generally – irrespective of where they live.

[31]The use of non-natives in this way is methodologically similar to Nunn and Wantchekon (2011) and Michalopoulos et al. (2016), who also use variation within non-native Africans to disentangle regional effects from individual-level effects.

5.2 Locational and Individual Similarity Estimates

I test the general importance of locational and individual similarity vis-à-vis changes in the DHS wealth index – a composite measure of cumulative living conditions for a household. The index is constructed using data on a household ownership of assets (e.g., television, refrigerator, telephone, etc.) and access to public resources (e.g., water, electricity, sanitation facility, etc.). Since variation in linguistic similarity comes from leadership changes, a positive estimate for either measure implies better access to public resources and more acquired assets because of an individual's similarity across the significant dimension.

In every specification I include country-wave fixed effects, locational language-wave fixed effects and individual language-wave fixed effects. As previously mentioned I do not include country-language fixed effects because in some instances I only observe a single language for a country. Unlike estimating equation (1), I include individual language-wave fixed effects because 45 percent of respondents' home language is different than their locational language.

I report these estimates in Table 10. In column 1 the estimate for lexicostatistical locational similarity is positive and significant at the 1 percent level. This point estimate of 0.540 is equivalent to 0.35 of a standard deviation increase in the wealth index. In column 2 I report the estimate for individual similarity. While the estimate has the expected positive sign, the coefficient is not precisely estimated. This suggests that changes in the individual-level wealth index are coming from transfers based on the ethnic identity of a region. Indeed, when I run a horse race between the two, I find that locational similarity is significantly different than zero while individual similarity remains insignificant. [32]

Overall, these estimates indicate that favoritism operates through regional transfers, which suggests that favoritism is beneficial to all inhabits of a region regardless of their background. This finding is consistent with the evidence that Kenyan leaders invest twice as much in roads (Burgess et al., 2015), and disproportionately target school construction in their coethnic districts (Kramon and Posner, 2016). In a case study of Congo-Brazzaville, Franck and Rainer (2012) similarly find that ethnic divisions impact the patterns of regional school construction. However, this case study also points to anecdotal evidence of the individual-level channel, where coethnic individuals benefit from preferential access to education and civil servant jobs irrespective of where they live. Kramon and Posner (2016) similarity posit the existence of this preferential access channel. To the contrary, I find that an individual's similarity to her leader does not afford her any luxuries beyond the location effect.

Finally, to show that the locational mechanism is not only driven by the coethnic effect, I

[32] See Appendix D for the unconditional estimates.

Table 10: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index				
	(1)	(2)	(3)	(4)
Locational similarity$_{t-1}$	0.540***		0.541***	
	(0.128)		(0.128)	
Individual similarity$_{t-1}$		0.239	0.240	
		(0.216)	(0.216)	
Locational coethnicity$_{t-1}$				0.501***
				(0.133)
Non-coethnic locational similarity$_{t-1}$				0.742***
				(0.148)
Individual controls	Yes	Yes	Yes	Yes
Distance controls	Yes	Yes	Yes	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes
Clusters	88	88	88	88
Countries	13	13	13	13
Language groups	20	20	20	20
Adjusted R^2	0.605	0.605	0.605	0.605
Observations	56,455	56,455	56,455	56,455

This table reports estimates that test for favoritism outside of coethnic language partitions. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, rural indicator variable, a gender indicator variable and an indicator for respondents living in the capital city. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

separately estimate locational coethnicity and non-coethnic locational similarity. I do this in the same way I did in the regional-level analysis: I define non-coethnic locational similarity as $(1-\text{coethnicity}) \times \text{locational similarity}$. Column 4 of Table 10 reports this estimate. I find that both the coethnic and non-coethnic effect are positive and strongly significant. These estimates suggest that the average level of non-coethnic locational similarity (0.164) yields

an increase of 0.122 ($= 0.164 \times 0.742$) in the wealth index – roughly one fourth the coethnic effect.

6 Interpretation

The results of this paper indicate that ethnic favoritism is widespread throughout Africa, and that patronage is distributed to both the ethnic region of the leader and to related but non-coethnic regions. But what mechanism drives these regional transfers? Why might we expect to see favoritism outside of the leader's ethnic region?

I offer an explanation that relates to the literature on coalition building. Central to this literature is the idea that leaders respond to political instability by co-opting elites from outside of their ethnic group into high-level government positions to pacify unrest and to maintain control of the state (Joseph, 1987; Arriola, 2009; Francois et al., 2015). The fact that similar but not identical ethnic regions benefits from patronage suggests that ethnicity is more than just a marker of identity: similarity may capture affinity between related non-coethnic groups. It is intuitive that the "closeness" of a group to the leader would predict their share in the governing coalition for reasons related to trust (Habyarimana et al., 2009), reduced costs of coordination (Miguel and Gugerty, 2005), clientelistic networks (Wantchekon, 2003) and more. Because leaders share power with ethnic groups other than their own to make credible their promise of patronage (Arriola, 2009), any evidence that leaders appoint closely related groups is an indication that coalition building is one mechanism underlying this paper's findings.

An insightful paper by Francois et al. (2015) provides theoretical and empirical support for the claim that ethnic group representation in the governing coalition is proportional to a group's share of the national population. The logic of this theory runs contrary to ethnic favoritism: their proposed mechanism underlying coalition building is group size. Yet these authors still find that a leader's ethnic group receives a premium in government appointments over and above the effect of group size. While it is beyond the scope of this paper to take a stance on the relative importance of these channels, what is important is that they are not mutually exclusive to each other.

To shed light on this interesting area of research I document that the similarity of an ethnic group to the leader correlates with an ethnic group's representation in government *conditional* on group size. I use yearly data from Francois et al. (2015) on the share of an ethnic group's representation in the governing coalition for 15 African countries between 1992 and 2004.[33] The majority of Ethnologue groups are defined as Others in Francois et al.'s

[33]See Appendix D for details on the construction of this dataset.

(2015) data, which severely limits the observable number of group partitions. Consequently, it is not possible to use the same source of within-group variation employed elsewhere in this paper. Instead I use an identical set-up to Francois et al. (2015), but augment their empirical model with an indicator variable for similar but not identical ethnic groups:

$$y_{c,e,t} = \alpha \, \text{coethnic}_{c,e,t} + \beta \, \text{similar}_{c,e,t} + f\left(\text{groupsize}_{c,e}\right) + \delta_c + \gamma_t + \epsilon_{c,e,t}. \tag{2}$$

The outcome variable $y_{c,e,t}$ is ethnic group e's share of cabinet positions in country c in year t. In addition to the usual $\text{coethnic}_{c,e,t}$ indicator variable, I include an indicator equal to one when a non-coethnic group's linguistic similarity is greater than a defined threshold of "closeness" (i.e., $\text{similar}_{c,e,t}$). Francois et al. (2015) find that $\text{groupsize}_{c,e}$ – the population share of ethnic group e in country c – is concave in its relationship with a group's share of cabinet positions, so I include $\text{groupsize}_{c,e}$ and its polynomial in all regressions. δ_c and γ_t capture unobserved time-invariant country effects and time trends. I follow Francois et al. (2015) and cluster standard errors at the country level.

I exploit a range of thresholds to let the data inform me of the relevant threshold of closeness. My preferred approach is to split the distribution of linguistic similarity for non-coethnics into deciles. I assign the threshold for closeness as any non-coethnic observation with similarity to the leader that is equal to or greater than a defined decile of the distribution. Hence, these thresholds are cumulative, where $\text{similar}_{c,e,t}$ is equal to one when a non-coethnic group's linguistic similarity is equal to or greater than the decile cut-off.[34]

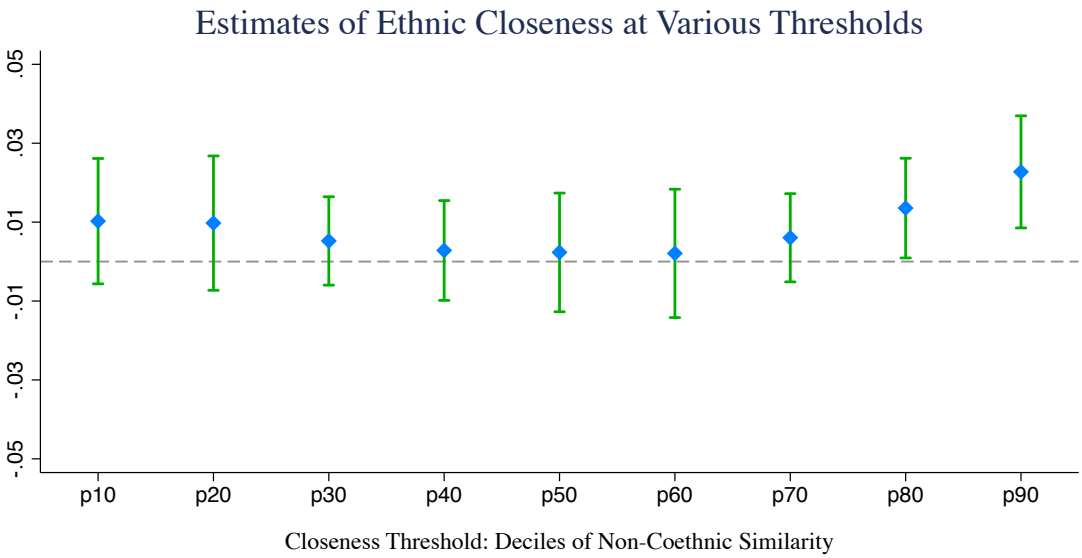
I plot the point estimates of β in Figure 7 for various thresholds, where the intervals reflect 99 percent confidence levels. The figure clearly documents non-coethnic favoritism in coalition building, at least among the stricter definitions of closeness. Because I include $\text{coethnic}_{c,e,t}$ and $\text{similar}_{c,e,t}$ in each regression, the estimate of β reflects the additional share of cabinet positions a similar non-coethnic group receives *relative* to other non-coethnics that do not satisfy the threshold level of similarity.

To assess the economic meaning of these estimates, I can compare the difference in a group's predicted outcome conditional on *mean* group size, after turning β on and off. Let the closeness threshold be the most stringent threshold at the 90^{th} percentile of the distribution. I find that a non-coethnic group's share in the governing coalition jumps from 5.1 to 7.4 percent when β is included – a 45 percent increase.[35] The resulting share is almost 2 percentage points larger than the sample average share of 5.6 percent.

[34] In Appendix D, I replicate Table III from Francois et al. (2015) and include the more general specification of lexicostatistical similarity in place of coethnicity.

[35] The point estimate for $\text{similar}_{c,e,t} = 0.023$, while the point estimates for group size and its polynomial are 1.225 and -1.795. For the mean non-coethnic group, the effect of group size is $5.1 = 1.225 \times 0.052 - 1.795 \times 0.007$, and 7.4 when adding the $\text{similar}_{c,e,t}$ premium.

Figure 7: Ethnic Favoritism and Coalition Building



This figure plots point estimates of the $\text{similar}_{c,e,t}$ indicator variable in equation (2) at nine different closeness thresholds. Thresholds are set according to deciles of the distribution for non-coethnic similarity. These thresholds are cumulative, where $\text{similar}_{c,e,t}$ is equal to one when a non-coethnic group's linguistic similarity is equal to or greater than the decile threshold. Each estimate reflects the additional share of cabinet positions a similar non-coethnic group receives relative to non-similar non-coethnics. Intervals reflect 99% confidence levels.

But how similar are these "close" groups? Consider the Gbe ethnolinguistic family, where three of the most widely spoken languages include Fon, Ewe and Gen. For the three possible pairings of these languages, the average lexicostatistical similarity is 46.8 percent. The mean similarity among non-coethnic groups in the 90^{th} percentile of the distribution is 45.7 percent. This suggests that leaders appoint elites from outside of their immediate ethnic group that are part of the same family cluster. In other words, the affinity that similarity captures is reflective of the shared ancestry in a group's larger ethnic network.

Overall, these findings suggest that leaders are inclined to make ethnicity-based decisions when appointing ministers from outside their own ethnic group. While the estimates of equation (2) cannot necessarily be taken as causal, they are informative of the mechanism through which public resources are allocated to non-coethnic regions. The tendency of ministers to redirect funds to their coethnics explains why non-coethnic groups with representation in government receive patronage (Arriola, 2009).

7 Concluding Remarks

Ethnic favoritism is often thought to be endemic to African politics, yet the empirical basis for this claim is largely founded on single-country case studies. In this paper, I document evidence that ethnic favoritism is widespread throughout Africa using data for 35 sub-Saharan countries. I also introduce a novel measure of linguistic similarity that contributes to this literature in three ways: (i) it better predicts patterns of ethnic favoritism with added variation in measured similarity, (ii) the continuity of the measure enables the study of favoritism among groups that are not coethnic to the leader, and (iii) it informs our understanding of a new mechanism related to the ethnic affinity between similar but non-coethnic groups. This deepens our understanding of the *extent* of favoritism – evidence of favoritism among non-coethnics normally goes undetected when using a coethnic dummy variable. I also show that patronage tends to be distributed at a regional level rather than as targeted transfers towards individuals. I interpret these results through the lens of coalition building and find that ethnicity is one of the guiding principles behind high-level government appointments.

These observations inform policy in a number of new ways. In particular, my findings are informative of both the extent of favoritism and where it is expected to take place. This can be used for many purposes, one of which is to enhance monitoring of foreign aid. There is a growing body of evidence that links the misuse of foreign aid to ethnic patronage in Africa (Briggs, 2014; Jablonski, 2014). Greater oversight is achieved through a deeper understanding of where patronage is expected to flow. My findings suggest aid donors should not only worry about patronage directed toward the leader's ethnic group but also toward other related groups. The benefits of oversight are evident when comparing the non-interference aid policy of China with conditional transfers from the World Bank. Dreher et al. (2015) find little evidence that World Bank aid is used for patronage purposes in contrast to the evidence that China's non-interference policy engenders resource allocation across ethnic lines rather than on a basis of need.

More generally, my findings speak to the value of nation-building policies that promote diversity in a Pan-Africanist tradition. Tanzania is a good example of a country that has stressed a sense of unity and shared history in its national policies. One nation-building tool of this type that is particularly relevant to this paper is Tanzania's national language policy (Miguel, 2004). In the mid-1960s, the Tanzanian government changed the official language of the country to Swahili. The extent to which Swahili is found in other countries and commonly used as a lingua franca speaks to the ethnic neutrality of the language. Within only a few years of its implementation, the official status granted to Swahili was described as a "linguistic revolution" for its ability to help shape a national consciousness that runs

contrary to ethnic identity (Harries, 1969, p. 277). The Tanzania example is a model to be replicated elsewhere, given the evidence that ethnic favoritism is so widespread throughout Africa. This is not to imply that national language policies are the only means to pacify existing ethnic tensions: the lesson here is that national policies must be designed to engender acceptance of diversity through unity. For example, education is an effective way to build a national culture that actively values diversity and differences in experience and background.

Lastly, the findings of this paper call for future work. The evidence that favoritism is not simply a coethnic phenomenon demands a deeper understanding of what it means to be "close" to the ruling ethnic group. The taxonomy of linguistic and ethnic clusters provide an opportunity to study this notion of closeness in the same vein as Desmet et al. (2012). Linking the extent of favoritism to the impact it has on ethnic inequality is an important next step in this line of research. The Tanzania example also suggests ethnic favoritism is not an unavoidable consequence of a country's high level of diversity, an observation that is consistent with the literature on ethnic inequality. Why then do we observe favoritism in some countries and not others? Geographic segregation is linked to ethnic favoritism in Africa (Ejdemyr et al., 2014), while geographic endowments are linked to ethnic inequality (Alesina et al., 2016), which suggests an answer to this question lies at the intersection of these two areas of research.

References

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, 8(2):155–194.

Alesina, A., Easterly, W., and Matuszeski, J. (2011). Artificial States. *Journal of the European Economic Association*, 9(2):246–277.

Alesina, A. and La Ferrara, E. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A., Michalopoulos, S., and Papaioannou, E. (2016). Ethnic Inequality. *Journal of Political Economy*, 124(2):428–488.

Arriola, L. R. (2009). Patronage and Political Stability in Africa. *Comparative Political Studies*, 42(10):1339–1362.

Bakker, D., Brown, C. H., Brown, P., Egorov, D., Grant, A., Holman, E. W., Mailhammer, R., Müller, A., Velupillai, V., and Wichmann, S. (2009). Add Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology*, 13:167–179.

Baldwin, K. and Huber, J. D. (2010). Economic Versus Cultural Differences: Forms of Ethnic Diversity and Public Goods Provision. *American Political Science Review*, 104(4):644–662.

Bates, R. H. (1974). Ethnic Competition and Modernization in Contemporary Africa. *Comparative Political Studies*, 6(4):457–484.

Batibo, H. M. (2005). *Language Decline and Death in Africa: Causes, Consequences and Challenges*. Multilingual Matters, Tonawanda.

Bloemen, H. G. (2013). Language Proficiency of Migrants: The Relation with Job Satisfaction and Matching. *IZA Discussion Paper 7366*.

Bowles, S. and Gintis, H. (2004). Persistent Parochialism: Trust and Exclusion in Ethnic Networks. *Journal of Economic Behavior & Organization*, 55:1–23.

Briggs, R. C. (2014). Aiding and Abetting: Project Aid and Ethnic Politics in Kenya. *World Development*, 64:194–205.

Burgess, R., Miguel, E., Jedwab, R., Morjaria, A., and Padró i Miquel, G. (2015). The Value of Democracy: Evidence from Road Building in Kenya. *American Economic Review*, 105(6):1817–1851.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Caselli, F. and Coleman, W. J. (2013). On the Theory of Ethnic Conflict. *Journal of the European Economic Association*, 11(S1):161–192.

Collier, P. and Gunning, J. W. (1999). Explaining African Economic Performance. *Journal of Economic Literature*, 37(1):64–111.

De Luca, G., Hodler, R., Raschky, P. A., and Valsecchi, M. (2015). Ethnic Favoritism: An Axiom of Politics? *CESifo Working Paper 5209*, pages 1–35.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2012). The Political Economy of Linguistic Cleavages. *Journal of Development Economics*, 97(2):322–338.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2015). Culture, Ethnicity and Diversity. *NBER Working Paper 20989*.

Desmet, K., Ortuño-Ortín, I., and Weber, S. (2009). Linguistic Diversity and Redistribution. *Journal of the European Economic Association*, 7(6):1291–1318.

Dickens, A. (2016). Population Relatedness and Cross-Country Idea Flows: Evidence from Book Translations. *York University, mimeo*.

Dreher, A., Fuchs, A., Parks, B. C., Raschky, P. A., and Tierney, M. J. (2015). Aid on Demand: African Leaders and the Geography of China's Foreign Assistance. *CESifo Working Paper 5439*.

Dyen, I., Kruskal, J. B., and Black, P. (1992). An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society*, 82(5):1–132.

Easterly, W. and Levine, R. (1997). Africa's Growth Tragedy: Policies and Ethnic Divisions. *The Quarterly Journal of Economics*, 112(4):1203–1250.

Ejdemyr, S., Kramon, E., and Robinson, A. L. (2014). Segregation, Ethnic Favoritism, and the Strategic Targeting of Local Public Goods. *Stanford University, mimeo*.

Englebert, P., Tarango, S., and Carter, M. (2002). Dismemberment and Suffocation: A Contribution to the Debate on African Boundaries. *Comparative Political Studies*, 35(10):1093–1118.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and Conflict: An Empirical Study. *American Economic Review*, 102(4):1310–1342.

Esteban, J. and Ray, D. (2011). A Model on Ethnic Conflict. *Journal of the European Economic Association*, 9(3):496–521.

Fearon, J. D. (2003). Ethnic and Cultural Diversity by Country. *Journal of Economic Growth*, 8(2):195–222.

Fearon, J. D. and Laitin, D. D. (1999). Weak States, Rough Terrain, and Large-Scale Ethnic Violence Since 1945. *Paper presented at the 1999 Annual Meetings of the American Political Science Association*.

Fenske, J. (2013). Does Land Abundance Explain African Institutions? *The Economic Journal*, 123(573):1363–1390.

Franck, R. and Rainer, I. (2012). Does the Leader's Ethnicity Matter? Ethnic Favoritism, Education, and Health in Sub-Saharan Africa. *American Political Science Review*, 106(2):294–325.

Francois, P., Rainer, I., and Trebbi, F. (2015). How Is Power Shared in Africa? *Econometrica*, 83(2):465–503.

Galor, O. and Ozak, O. (2016). The Agricultural Origins of Time Preference. *American Economic Review*, 106(10):3064–3103.

Gennaioli, N. and Rainer, I. (2007). The Modern Impact of Precolonial Centralization in Africa. *Journal of Economic Growth*, 12(3):185–234.

Ginsburgh, V. A. and Weber, S. (2016). Linguistic Distances and Ethnolinguistic Fractionalization and Disenfranchisement Indices. In Ginsburgh, V. A. and Weber, S., editors, *The Palgrave Handbook of Economics and Language*, pages 137–173. Palgrave Macmillan UK, London.

Goemans, H. E., Gleditsch, K. S., and Chiozza, G. (2009). Introducting Archigos: A Data Set of Political Leaders. *Journal of Peace Research*, 46(2):269–283.

Gomes, J. F. (2014). The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa. pages 1–48.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural Biases in Economic Exchange? *The Quarterly Journal of Economics*, 124(3):1095–1131.

Habyarimana, J., Humphreys, M., Posner, D. N., and Weinstein, J. M. (2009). Coethnicity and Trust. In Cook, K., Levi, M., and Hardin, R., editors, *Whom Can We Trust?*, pages 42–64. Russell Sage Foundation, New York.

Harries, L. (1969). Language Policy in Tanzania. *Journal of the International African Institute*, 39(3):275–280.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring Economic Growth From Outer Space. *The American Economic Review*, 102(2):994–1028.

Herbst, J. (2000). *State and Power in Africa*. Princeton University Press, Princeton.

Hodler, R. and Raschky, P. A. (2014). Regional Favoritism. *The Quarterly Journal of Economics*, 129(2):995–1033.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2009). Explorations in Automated Language Classification. *Folia Linguistica*, 42(3-4):331–354.

Huber, J. D. and Suryanarayan, P. (2014). Ethnic Inequality and the Ethnification of Political Parties: Evidence from India. *Columbia University, mimeo.*

Isphording, I. E. and Otten, S. (2013). The Costs of Babylon: Linguistic Distance in Applied Economics. *Review of International Economics*, 21(2):354–369.

Isphording, I. E. and Otten, S. (2014). Linguistic Barriers in the Destination Language Acquisition of Immigrants. *Journal of Economic Behavior and Organization*, 105(5):30–50.

Jablonski, R. S. (2014). How Aid Targets Votes: The Impact of Electoral Incentives on Foreign Aid Distribution. *World Politics*, 66(2):293–330.

Jensen, N. and Wantchekon, L. (2004). Resource Wealth and Political Regimes in Africa. *Comparative Political Studies*, 37(7):816–841.

Joseph, R. A. (1987). *Democracy and Prebendalism in Nigeria*. Cambridge University Press, Cambridge.

Kasara, K. (2007). Tax Me If You Can: Ethnic Geography, Democracy, and the Taxation of Agriculture in Africa. *The American Political Science Review*, 101(1):159–172.

Kezdi, G. (2004). Robust Standard Error Estimation in Fixed-Effects Panel Models. *Hungarian Statistical Review*, 9:95–116.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A Global Index Representing the Stability of Malaria Transmission. *American Journal of Tropical Medicine and Hygiene*, 70(5):486–498.

Kramon, E. and Posner, D. N. (2016). Ethnic Favoritism in Primary Education in Kenya. *Quarterly Journal of Political Science*, 11(1):1–58.

Kyriacou, A. P. (2013). Ethnic Group Inequalities and Governance : Evidence from Developing Countries. *Kyklos*, 66(1):78–101.

Lewis, P. M. (2009). *Ethnologue: Languages of the World*. SIL International, Dallas, 16 edition.

Marx, B., Stoker, T. M., and Suri, T. (2015). There is No Free House: Ethnic Patronage in a Kenyan Slum. *MIT, mimeo.*

Michalopoulos, S. and Papaioannou, E. (2013). Pre-colonial Ethnic Institutions and Contemporary African Development. *Econometrica*, 81(1):113–152.

Michalopoulos, S. and Papaioannou, E. (2014). National Institutions and Subnational Development in Africa. *The Quarterly Journal of Economics*, 29(1):151–213.

Michalopoulos, S. and Papaioannou, E. (2016). The Long-Run Effects of the Scramble for Africa. *American Economic Review*, 106(7):1802–1848.

Michalopoulos, S., Putterman, L., and Weil, D. N. (2016). The Influence of Ancestral Lifeways on Individual Economic Outcomes in Sub-Saharan Africa. *NBER Working Paper 21907*.

Miguel, E. (2004). Tribe or Nation?: Nation Building and Public Goods in Kenya versus Tanzania. *World Politics*, 56(3):327–362.

Miguel, E. and Gugerty, M. K. (2005). Ethnic Diversity, Social Sanctions and Public Goods in Kenya. *Journal of Public Economics*, 89(11-12):2325–2368.

Mwakikagile, G. (2010). *Ethnic Diversity and Integration in The Gambia: The Land, The People and The Culture*. Continental Press, Dar es Salaam.

Nunn, N. (2008). The Long-Term Effects of Africa's Slave Trades. *The Quarterly Journal of Economics*, 123(1):139–176.

Nunn, N. and Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American Economic Review*, 101(7):3221–3252.

Oucho, J. (2006). Cross-Border Migration and Regional Initiatives in Managing Migration in Southern Africa. In Kok, P., Gelderblom, D., Oucho, J., and van Zyl, J., editors, *Migration in South and Southern Africa: Dynamics and Determinants*, pages 47–70. HSRC Press, Cape Town.

Padro i Miquel, G. (2007). The Control of Politicians in Divided Societies: The Politics of Fear. *The Review of Economic Studies*, 74(2007):1259–1274.

Posner, D. N. (2004). Measuring ethnic fractionalization in Africa. *American Journal of Political Science*, 48(4):849–863.

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). A Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Changes. *Global Ecology and Biogeography*, 11:377–392.

Solon, G., Haider, S. J., and Wooldridge, J. (2015). What Are We Weighting For? *Journal of Human Resources*, 50(2):301–316.

Spolaore, E. and Wacziarg, R. (2009). The Diffusion of Development. *The Quarterly Journal of Economics*, 124(2):469–529.

Swadesh, M. (1952). Lexicostatistical Dating of Prehistoric Ethnic Contracts. *Proceedings of the American Philosophical Society*, 96:121–137.

Swadesh, M. (1955). Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21:121–137.

Wantchekon, L. (2003). Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin. *World Politics*, 55(3):399–422.

Wesseling, H. (1996). *Divide and Rule: The Partition of Africa, 1880-1914*. Praeger, Westport.

Wichmann, S., Holman, E. W., Bakker, D., and Brown, C. H. (2010). Evaluating Linguistic Distance Measures. *Physica A*, 389(17):3632–3639.

Wright, D. R. (2015). *The World and a Very Small Place in Africa: A History of Globalization in Niumi, The Gambia*. Routledge, New York.

Young, A. (2013). Inequality, the Urban-Rural Gap, and Migration. *Quarterly Journal of Economics*, 128(4):1727–1785.

FOR ONLINE PUBLICATION

A Data Descriptions, Sources and Summary Statistics

A.1 Regional-Level Data Description and Sources

Country-language groups: Geo-referenced country-language group data comes from the World Language Mapping System (WLMS). These data map information from each language in the Ethnologue to the corresponding polygon. When calculating averages within these language group polygons, I use the Africa Albers Equal Area Conic projection.
Source: http://www.worldgeodatasets.com/language/

Linguistic similarity: I construct two measures of linguistic similarity: lexicostatistical similarity from the Automatic Similarity Judgement Program (ASJP), and cladistic similarity using Ethnologue data from the WLMS. I use these to measure the similarity between each language group and the ethnolinguistic identity of that country's national leader. I discuss how I assign a leader's ethnolinguistic identity in Section 2.3.
Source: http://asjp.clld.org and http://www.worldgeodatasets.com/language/

Night lights: Night light intensity comes from the Defense Meteorological Satellite Program (DMSP). My measure of night lights is calculated by averaging across pixels that fall within each WLMS country-language group polygon for each year the night light data is available (1992-2013). To minimize area distortions I use the Africa Albers Equal Area Conic projection. In some years data is available for two separate satellites, and in all such cases the correlation between the two is greater than 99% in my sample. To remove choice on the matter I use an average of both. The dependent variable used in the benchmark analysis is $\ln(0.01 + \text{average night lights})$.
Source: http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Population density: Population density is calculated by averaging across pixels that fall within each country-language group polygon. To minimize area distortions I use the Africa Albers Equal Area Conic projection. Data comes from the Gridded Population of the World, which is available in 5-year intervals: 1990, 1995, 2000, 2005, 2010. For intermediate years I assume population density is constant; e.g., the 1995 population density is assigned to years 1995-1999. Throughout the regression analysis I use log population density.
Source: http://sedac.ciesin.columbia.edu/data/collection/gpw-v3

National leaders: I collected birthplace locations of all African leaders between 1991-2013. Names of African leaders and years entered and exited office comes from the Archigos Database on Leaders 1875-2004 (Goemans et al., 2009), which I extended to 2011 using data from Dreher et al. (2015), and 2012-2013 using a country's Historical Dictionary and other secondary sources.
Source: http://www.rochester.edu/college/faculty/hgoemans/data.htm

National leader birthplace coordinates: Birthplace locations are confirmed using Wikipedia, and entered into www.latlong.com to collect latitude and longitude coordinates.
Source: http://www.latlong.net

Years in office: To calculate each leader's current years in office and total years in office I use the entry and exit data described above.
Source: Calculated using Stata.

Distance to leader's birth region: Country-language group centroids calculated in ArcGIS, and the distance between each centroid and the national leader's birthplace coordinates is calculated in Stata using the `globdist` command. Throughout the regression analysis I use log leader birthplace distance.
Source: Calculated using ArcGIS and Stata.

Absolute difference in elevation: I collect elevation data from the National Geophysical Data Centre (NGDC) at the National Oceanic and Atmospheric Administration (NOAA). I measure average elevation of each partitioned language group and leader's ethnolinguistic group. To minimize area distortions I use the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Absolute difference in ruggedness: As a measure of ruggedness I use the standard deviation of the NGDC elevation data. I use Stata to calculate the absolute difference between the two.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Absolute difference in precipitation: Precipitation data comes from the WorldClim – Global Climate Database. I measure average precipitation within each partitioned lan-

guage group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://www.worldclim.org/current

Absolute difference in temperature: Temperature data comes from the WorldClim – Global Climate Database. I measure the average temperature within each partitioned language group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://www.worldclim.org/current

Absolute difference in caloric suitability index: I sourced the caloric suitability index (CSI) data from Galor and Ozak (2016). CSI is a measure of agricultural productivity that reflects the caloric potential in a grid cell. It's based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). A variety of related measures are available: in the reported estimates I use the pre-1500 average CSI measure that includes cells with zero productivity. The results are not sensitive to which measure I use. I measure average CSI within each partitioned language group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://omerozak.com/csi

Oil reserve: I construct an indicator variable equal to one if an oil field is found in both the partitioned language group and leader's ethnolinguistic group. Version 1.2 of the Petroleum Dataset contains geo-referenced point data indicating the presence of on-shore oil and gas deposits from around the world.
Source: https://www.prio.org/Data/Geographical-and-Resource-Datasets/Petroleum-Dataset/

Diamond reserve: I construct an indicator variable equal to one if a known diamond deposit is found in both the partitioned language group and leader's ethnolinguistic group. Version 1.2 of the Petroleum Dataset contains geo-referenced point data indicating the presence of on-shore oil and gas deposits from around the world.
Source: https://www.prio.org/Data/Geographical-and-Resource-Datasets/Diamond-Resources/

A.2 Individual-Level Data Description and Sources

Unless otherwise stated, all individual-level data comes from the Demographic and Health Surveys (DHS). Source: http://dhsprogram.com/

Individual linguistic similarity: To assign an individual a home language I assign the reported language a respondent speaks at home when this data is available (59 percent availability). For surveys when this data isn't available or the reported language is "other", I map the respondent's home language from their reported ethnicity. To do this I use the following assignment rule:

1. Direct match: the DHS ethnicity name is the same as an Ethnologue language name for the respondent's country of residence.

2. Alternative name: the unmatched DHS ethnicity is an unambiguous alternative name for a language in the Ethnologue or Glottolog database.

3. Macrolanguage: if the ethnicity corresponds to a macrolanguage in the Ethnologue, then I assign the most populated sub-language of that macrolanguage.

4. Population size: if the unmatched ethnicity maps to numerous languages, I choose the language with the largest Ethnologue population.

I also cross-reference the Wikipedia page for each ethnic group to corroborate that the assigned language maps into the reported ethnicity. Then using the same data on leaders as in the regional-analysis, I match the lexicostatistical similarity of the respondent's home language to the leader's ethnolinguistic identity.
Source: http://asjp.clld.org

Locational linguistic similarity: I project DHS cluster latitude and longitude coordinates onto the Ethnologue language map and assign the associated language as the regional language group to that respondent. In instances of overlapping language groups, I assign the largest group in terms of population. Then using the same data on leaders as in the regional-analysis, I match the lexicostatistical similarity of the respondent's home language to the leader's ethnolinguistic identity.
Source: http://asjp.clld.org

Wealth Index: I use the quantile DHS wealth index. The quantile index is derived from a composite measure of a household's assets (e.g., television, refrigerator, telephone, etc.)

and access to public resources (e.g., water, electricity, sanitation facility, etc.), in addition to data indicating if a household owns agricultural land and if they employ a domestic servant. Principal component analysis is used to construct the original index, then respondents are order by score and sorted into quintiles. Read the DHS Comparative Report: The DHS Wealth Index for more details.

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.

Education: The 10 education fixed effects are from question 90.

Religion: The 18 fixed effects for the religion of a respondent come from question 91.

Distance to the capital: I use the World Cities layer available on the ArcGIS website, which includes latitude-longitude coordinates and indicators for capital cities. I calculate language group centroids coordinates using ArcGIS, and measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: `http://www.arcgis.com/home/`

Distance to the coast: I use the coastline shapefile from Natural Earth, calculate the nearest coastline from a language groups centroid using the Near tool in ArcGIS. I measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: `http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/`

Distance to the border: I use country boundaries from the Digital Chart of the World (5^{th} edition) that's complimentary to the Ethnologue data from the WLMS, and calculate the nearest border from a language groups centroid using the Near tool in ArcGIS. I measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: `http://www.worldgeodatasets.com/language/`

A.3 Summary Statistics and Additional Details

Table A1: Language Groups Included in Regional-Level Analysis

Sample	**Language Groups**
Regional-Level Analysis	Acholi, Adamawa Fulfulde, Adele, Afade, Afrikaans, Alur, Anuak, Anufo, Anyin, Baatonum, Badyara, Baka, Bari, Bata, Bayot, Bedawiyet, Bemba, Berta, Bissa, Boko, Bokyi, Bomwali, Borana-Arsi-Guji Oromo, Buduma, Central Kanuri, Chadian Arabic, Chidigo, Cokwe, Daasanach, Dan, Dazaga, Dendi, Dholuo, Diriku, Ditammari, Ejagham, Ewe, Fur, Gbanziri, Gidar, Glavda, Gola, Gourmanchema, Gude, Gumuz, Hausa, Herero, Holu, Jola-Fonyi, Juhoan, Jukun Takum, Jula, Kaba, Kacipo-Balesi, Kako, Kakwa, Kalanga, Kaliko, Kaonde, Kasem, Khwe, Kikongo, Kisikongo, Kiswahili, Komo, Konkomba, Koromfe, Kuhane, Kunama, Kunda, Kuo, Kuranko, Kusaal, Kwangali, Kxauein, Langbashe, Lozi, Lugbara, Lunda, Lutos, Luvale, Maasai, Madi, Makonde, Mambwe-Lungu, Mandinka, Mandjak, Manga Kanuri, Mann, Manyika, Masana, Mashi, Mbandja, Mbay, Mbukushu, Mende, Monzombo, Moore, Mpiemo, Mundang, Mundu, Musey, Musgu, Nalu, Naro, Ndali, Ndau, Ngangam, Ngbaka Mabo, Ninkare, Northern Kissi, Northwest Gbaya, Nsenga, Ntcham, Nuer, Nyakyusa-Ngonde, Nyanja, Nzakambay, Nzanyi, Nzema, Oshiwambo, Pana, Peve, Pokoot, Psikye, Pulaar, Pular, Runga, Rwanda, Saho, Shona, Shuwa Arabic, Somali, Soninke, Southern Birifor, Southern Kisi, Southern Sotho, Susu, Swati, Taabwa, Talinga-Bwisi, Tamajaq, Tedaga, Teso, Tigrigna, Tonga, Tswana, Tumbuka, Tupuri, Vai, Venda, Wandala, Western Maninkakan, Xhosa, Xoo, Yaka, Yaka, Yalunka, Yao, Yeyi, Zaghawa, Zande, Zarma, Zemba, Zulu

Table A2: Language Groups Included in DHS Individual-Level Analysis

Sample	**Language Groups**
Individual-Level Analysis (Locational)	Alur, Bemba, Borana, Kaonde, Kasem, Kisi (Southern), Kissi (Northern), Kuhane, Kuranko, Lamba, Lugbara, Lunda, Maninkakan (Western), Mann, Oromo (Borana-Arsi-Guji), Pular, Somali, Soninke, Susu, Taabwa, Teso
Individual-Level Analysis (Individual)	Afar, Amharic, Aushi, Bamanankan, Bandi, Bemba, Berta, Bissa, Bobo Madare (Southern), Bwile, Cokwe, Dagaare (Southern), Dagbani, Dan, Dholuo, Ekegusii, Farefare, Ganda, Gedeo, Gikuyu, Gola, Gourmanchema, Gwere, Hadiyya, Harari, Hausa, Ila, Jola-Fonyi, Kamba, Kambaata, Kaonde, Kigiryama, Kipsigis, Kisi (Southern), Kissi (Northern), Kono, Koongo, Kpelle (Guinea), Kpelle (Liberia), Krio, Kuhane, Kunda, Kuranko, Lala-Bisa, Lamba, Lendu, Lenje, Limba (East), Lozi, Luba-Kasai, Lugbara, Lunda, Luvale, Maasai, Madi, Mambwe-Lungu, Mandinka, Maninkakan (Kita), Mann, Mbunda, Mende, Moore, Ngombe, Nkoya, Nsenga, Nyanja, Oromo (Borana-Arsi-Guji), Oromo (West Central), Oyda, Pulaar, Pular, Rendille, Samburu, Sebat Bet Gurage, Senoufo (Mamara), Serer-Sine, Sherbro, Sidamo, Soli, Somali, Songhay (Koyra Chiini), Soninke, Susu, Swahili, Taabwa, Tamasheq, Teso, Themne, Tigrigna, Tonga, Tumbuka, Turkana, Wolaytta, Wolof

Table A3: Leaders Included in Regional-Level Analysis

Sample	Leaders
Regional-Level Analysis	**Angola:** José Eduardo dos Santos; **Benin:** Thomas Yayi Boni, Mathieu Kérékou; **Botswana:** Quett Masire, Festus Mogae; **Burkina Faso:** Blaise Compaoré; **Cameroon:** Paul Biya; **Central African Republic:** Ange-Félix Patassé, André-Dieudonné Kolingba; **Chad:** Idriss Déby; **Congo:** Pascal Lissouba, Denis Sassou Nguesso; **Côte d'Ivoire:** Konan Bedie, Laurent Gbagbo, Robert Guéï, Félix Houphouët-Boigny, Alassane Ouattara; **DRC:** Joseph Kabila, Laurent-Désiré Kabila, Mobutu Sese Seko; **Eritrea:** Isaias Afewerki; **Ethiopia:** Hailemariam Desalegn, Meles Zenawi; **Gambia:** Yahya Jammeh, Dawda Jawara; **Ghana:** John Evans Atta-Mills, John Agyekum Kufuor, John Dramani Mahama, Jerry Rawlings; **Guinea:** Moussa Dadis Camara, Alpha Condé, Lansana Conté, Sékouba Konaté; **Guinea-Bissau:** Kumba Ialá, Manuel Serifo Nhamadjo, Henrique Periera Rosa, Malam Bacai Sanhé, João Bernardo Vieira; **Kenya:** Daniel arap Moi; Mwai Kibaki; **Lesotho:** Elias Phisoana Ramaema, Ntsu Mokhehle, Pakalithal Mosisili, Tom Thabane; **Liberia:** Gyude Bryant, Ruth Perry, Wilton G. S. Sankawulo, Ellen Johnson Sirleaf, Charles Taylor; **Malawi:** Hastings Kamuzu Banda, Joyce Banda, Bakili Muluzi, Bungu wa Mutharika; **Mali:** Alpha Oumar Konaré, Amadou Toumani Touré, Dioncounda Traoré; **Mozambique:** Armando Guebuza, Joaquim Chissano; **Namibia:** Sam Nujoma, Hifikepunye Pohamba; **Niger:** Mahamadou Issoufou, Ibrahim Baré Maïnassara, Mahamane Ousmane, Ali Saibou, Mamadou Tandja; **Nigeria:** Sani Abacha, Abdulsalami Abubakar, Goodluck Jonathan, Olusegun Obasanjo, Umaru Musa Yar'Adua; **Senegal:** Abdou Diouf, Macky Sall, Abdoulaye Wade; **Sierra Leone:** Ahmad Tejan Kabbah, Ernest Bai Koroma, Johnny Paul Koroma, Valentine Strasser; **Somalia:** Abdullahi Yusuf Ahmed, Sharif Sheikh Ahmed, Abdiqasim Salad Hassan, Hassan Sheikh Mohamud, Ali Mahdi Muhammad; **South Africa:** Frederik Willem de Klerk, Nelson Mandela, Thabo Mbeki, Jacob Zuma; **Sudan:** Omar Hassan Ahmad al-Bashir; **Tanzania:** Jakaya Kikwete, Benjamin Mkapa, Ali Hassan Mwinyi; **Togo:** Gnassingbé Eyadéma, Faure Gnassingbé; **Uganda:** Yoweri Museveni; **Zambia:** Frederick Chiluba, Levy Mwanawasa, Michael Sata; **Zimbabwe:** Robert Mugabe

Table A4: Leaders Included in DHS Individual-Level Analysis

Sample	Leaders
Individual-Level Analysis	**Burkina Faso:** Blaise Compaoré **Democratic Republic of Congo:** Joseph Kabila **Ethiopia:** Meles Zenawi **Ghana:** Jerry Rawlings; John Agyekum Kufuor **Guinea:** Alpha Condé; Lansana Conté **Kenya:** Mwai Kibaki **Liberia:** Ellen Johnson Sirleaf **Mali:** Alpha Oumar Konaré; Amadou Toumani Touré **Namibia:** Hifikepunye Pohamba **Senegal:** Abdou Diouf; Abdoulaye Wade **Sierra Leone:** Ernest Bai Koroma **Uganda:** Yoweri Museveni **Zambia:** Levy Mwanawasa; Michael Sata

Table A5: Countries Included in Regional- and Individual-Level Analysis

Sample	Countries
Regional-Level Analysis	Angola, Benin, Botswana, Burkina Faso, Cameroon, Central African Republic, Chad, Congo, Cote d'Ivoire, Democratic Republic of Congo, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Malawi, Mali, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Somalia, South Africa, Sudan, Tanzania, Togo, Uganda, Zambia, Zimbabwe
Individual-Level Analysis	Burkina Faso, Democratic Republic of Congo, Ethiopia, Ghana, Guinea, Kenya, Liberia, Mali, Namibia, Senegal, Sierra Leone, Uganda, Zambia

Table A6: Summary Statistics – Regional-Level Dataset

	Mean	Std dev.	Min	Max	N
Night lights$_t$	0.123	0.387	0.000	4.540	6,610
ln(0.01 + night lights$_t$)	-3.487	1.427	-4.605	1.515	6,610
ln(0.01 + night lights$_{t-1}$)	-3.507	1.415	-4.605	1.515	6,315
$\sqrt{\text{night lights}_t}$	0.187	0.297	0.000	2.131	6,610
ln(night lights$_t$)	-3.370	2.049	-10.60	1.513	4,069
Lexicostatistical similarity$_{t-1}$	0.193	0.230	0.000	1.000	6,610
Cladistic similarity$_{t-1}$	0.409	0.330	0.000	1.000	6,610
Coethnicity$_{t-1}$	0.047	0.212	0.000	1.000	6,610
Non-coethnic cladistic similarity$_{t-1}$	0.362	0.313	0.000	0.966	6,610
Non-coethnic lexicostatistical similarity$_{t-1}$	0.146	0.148	0.000	0.960	6,610
Lexicostatistical similarity$_{t+1}$	0.194	0.230	0.000	1.00	6228
Current years in office$_{t-1}$	11.44	8.680	1.000	38.00	6,610
Total years in office$_{t-1}$	18.50	10.19	1.000	38.00	6,610
Log distance (km) to leader's group$_{t-1}$	5.844	1.485	0.000	7.419	6,610
Log population density$_t$	2.886	1.529	-2.169	6.116	6,610
Absolute difference in elevation$_t$	250.5	296.1	0.000	2,021	6,610
Absolute difference in ruggedness$_t$	101.5	105.5	0.000	542.4	6,610
Absolute difference in precipitation$_t$	30.20	28.90	0.00	230.7	6,610
Absolute difference in mean temperature$_t$	16.81	17.09	0.000	120.2	6,610
Absolute difference in caloric suitability index$_t$	298.0	310.1	0.000	1711	6,610
Oil reserve in both leader and language group$_t$	0.018	0.131	0.000	1.000	6,610
Diamond mine in both leader and language group$_t$	0.079	0.269	0.000	1.000	6,610
Absolute difference in malaria suitability$_t$	4.951	5.635	0.000	29.30	5,111
Absolute difference in land suitability$_t$	0.178	0.184	0	0.777	5111
Democracy$_{t-1}$	0.435	4.877	-9.000	9.000	6,573
Language group population share	0.045	0.113	0	0.851	6610
Distance (km) to capital city	559.7	397.7	26.58	1922	6,610
Distance (km) to the coast	677.9	408.4	10.52	1743	6,610

Table A7: Summary Statistics – DHS Individual-Level Dataset

	Mean	Std Dev.	Min	Max	N
Wealth index	2.974	1.468	1.000	5.000	56,455
Locational similarity	0.350	0.380	0.025	1.000	56,455
Individual similarity	0.363	0.387	0.021	1.000	56,455
Age	29.36	10.51	15.00	78.00	56,455
Female indicator	0.663	0.473	0.000	1.000	56,455
Rural indicator	0.635	0.482	0.000	1.000	56,455
Education	4.721	1.520	1.000	6.000	56,455
Religion	4.912	2.032	1.000	8.000	56,455
Log distance to the coast (km)	6.059	0.910	1.654	7.238	56,455
Log distance to the border (km)	4.948	0.887	0.920	6.801	56,455
Log distance to the capital (km)	5.676	0.727	2.070	7.548	56,455

Table A8: Summary Statistics – Power Sharing Dataset

	Mean	Std dev.	Min	Max	N
Share of cabinet positions$_t$	0.056	0.078	0.000	0.471	2,539
Share of top cabinet positions$_t$	0.057	0.108	0.000	0.643	2,539
Share of low cabinet positions$_t$	0.055	0.078	0.000	0.450	2,539
Coethnicity$_t$	0.077	0.266	0.000	1.000	2,539
Lexicostatistical similarity$_t$	0.196	0.267	0.000	1.000	2,539
Non-coethnic lexicostatistical similarity$_t$	0.114	0.122	0.000	0.659	2,539
Ethnic group population share$_t$	0.057	0.065	0.005	0.390	2,539

B Measures of Linguistic Similarity

B.1 Computerized Lexicostatistical Similarity

The computerized approach to estimating lexicostatistical distances was developed as part of the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology. To begin a list of 40 implied meanings (i.e., words) are compiled for each language to compare the lexical similarity of any language pair. Swadesh (1952) first introduced the notion of a basic list of words believed to be universal across nearly all world languages. When a word is universal across world languages, its implied meaning, and therefore any estimate of linguistic distance, is independent of culture and geography. From here on I refer to this 40-word list as a Swadesh list, as it is commonly called.[36]

For each language the 40 words are transcribed into a standardized orthography called ASJPcode, a phonetic ASCII alphabet consisting of 34 consonants and 7 vowels. A standardized alphabet restricts variation across languages to phonological differences only. Meanings are then transcribed according to pronunciation before language distances are estimated.

I use a variant of the Levenshtein distance algorithm, which in its simplest form calculates the minimum number of edits necessary to translate the spelling of a word from one language to another. In particular, I use the normalized and divided Levenshtein distance estimator proposed by Bakker et al. (2009).[37] Denote $LD(\alpha_i, \beta_i)$ as the raw Levenshtein distance for word i of languages α and β. Each word i comes from the aforementioned Swadesh list. Define the length of this list be M, so $1 \leq i \leq M$.[38] The algorithm is run to calculate $LD(\alpha_i, \beta_i)$ for each word in the M-word Swadesh list across each language pair. To correct for the fact that longer words will often demand more edits, the distance is normalized according to word length:

$$LDN(\alpha_i, \beta_i) = \frac{LD(\alpha_i, \beta_i)}{L(\alpha_i, \beta_i)} \tag{3}$$

where $L(\alpha_i, \beta_i)$ is the length of the longer of the two spellings α_i and β_i of word i. $LDN(\alpha_i, \beta_i)$ is the normalized Levenshtein distance, which represents a percentage estimate of dissimilarity between languages α and β for word i. For each language pair, $LDN(\alpha_i, \beta_i)$ is calculated

[36] A recent paper by Holman et al. (2009) shows that the 40-item list employed here, deduced from rigorous testing for word stability across all languages, yields results at least as good as those of the commonly used 100-item list proposed by Swadesh (1955).

[37] I use Taraka Rama's (2013) Python program for string distance calculations.

[38] Wichmann et al. (2010) point out that in some instances not every word on the 40-word list exists for a language, but in all cases a minimum of 70 percent of the 40-word list exist.

for each word of the M-word Swadesh list. Then the average lexical distance for each language pair is calculated by averaging across all M words for those two languages. The average distance between two languages is then

$$LDN(\alpha, \beta) = \frac{1}{M} \sum_{i=1}^{M} LDN(\alpha_i, \beta_i). \tag{4}$$

A second normalization procedure is then adopted to account for phonological similarity that is the result of coincidence. This adjustment is done to correct for accidental similarity in sound structure of two languages that is unrelated to their historical relationship. The motivation for this step is that no prior assumptions need to be made about historical versus chance relationship. To implement this normalization the defined distance $LDN(\alpha, \beta)$ is divided by the global distance between two language. To see this, first denote the global distance between languages α and β as

$$GD(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j}^{M} LD(\alpha_i, \beta_j), \tag{5}$$

where $GD(\alpha, \beta)$ is the global (average) distance between two languages excluding all word comparisons of the same meaning. This estimates the similarity of languages α and β only in terms of the ordering and frequency of characters, and is independent of meaning. The second normalization procedure is then implemented by weighting equation (4) with equation (5) as follows:

$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{GD(\alpha, \beta)}. \tag{6}$$

$LDND(\alpha, \beta)$ is the final measure of linguistic distance, referred to as the normalized and divided Levenshtein distance (LDND). This measure yields a percentage estimate of the language dissimilarity between α and β. In instances where two languages have many accidental similarities in terms of ordering and frequency of characters, the second normalization procedure can yield percentage estimates larger than 100 percent by construction, so I divide $LDND(\alpha, \beta)$ by its maximum value to normalize the measure as a continuous $[0, 1]$ variable. Finally, I construct a measure of lexicostatistical linguistic similarity as follows:

$$LS(\alpha, \beta) = 1 - LDND(\alpha, \beta). \tag{7}$$

B.2 Cladistic Similarity

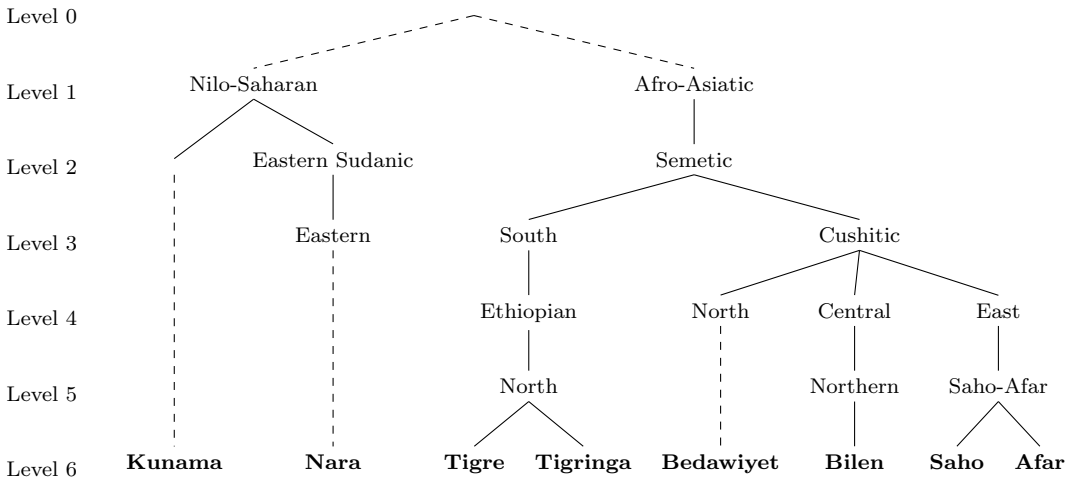
To construct a measure cladistic similarity I first calculate the number of shared branches between language α and β on the Ethnologue language tree, denoted $s(\alpha, \beta)$. Let M be the maximum number of tree branches between any two languages. I then construct cladistic linguistic similarity as follows:

$$CS(\alpha, \beta) = \left(\frac{s(\alpha, \beta)}{M} \right)^{\delta}, \tag{8}$$

where δ is an arbitrarily assigned weight used to discount more recent linguistic cleavages relative to deep cleavages. I describe this weight as arbitrary because there is no consensus on the appropriate weight to be assumed. Fearon (2003) argues the true function is probably concave and assumes a value of $\delta = 0.5$, which has since become the convention. Desmet et al. (2009) experiment with a range of values between $\delta \in [0.04, 0.10]$, but settle on a value of $\delta = 0.05$. In all reported estimates I assume $\delta = 0.5$, though the estimates are robust to alternative weighting assumptions (not shown here).

One issue with calculating cladistic similarity is the asymmetrical nature of historical language splitting. Because the number of branches varies among language families and subfamilies, the maximum number of branches between any two languages is not constant. To overcome this challenge I assume that all current languages are of equal distance from the proto-language at the root of the Ethnologue language tree. I visualize this assumption in Figure B1, where I have constructed a phylogenetic language tree for the 8 distinct languages of Eritrea. The dashed lines represent this assumed historical relationship, so in all cases the contemporary Eritrean languages possess an equal number of branches to the proto-language at Level 0. Although $M = 6$ in Figure B1, in the Ethnologue language tree the highest number of classifications for any language is $M = 15$, which I abstract from here for simplicity.

Figure B1: Phylogenetic Tree of Eritrean Languages



This figure depicts the language tree for the 8 major languages of Eritrea. Because of the asymmetrical nature of language splitting, the number of branches varies among language families. To measure cladistic similarity it is necessary that all branches be extended to the lowest level of aggregation. To do this I assume all languages are of equal distance from the proto-language at Level 0. Hence, the dashed lines represent the assumed relationship between the proto-language (Level 0) and the set of current Eritrean languages (Level 6).

C Mapping Ethnicity to Language

There is mostly agreement between ethnographers that language is a suitable marker of ethnicity in Africa (Batibo, 2005; Desmet et al., 2015). The challenge of mapping ethnicity to language is that, in some instances, a single ethnic group speaks many languages. In such instances it's not obvious what language is the appropriate language to match to a leader's ethnicity. As a solution to this problem I use the following three-step assignment rule to construct a mapping between ethnicity and language in Africa.

Step 1: For each ethnic group, I refer to the Ethnologue list of languages for the country to which they belong. If a language name is identical to the ethnic name then I assign the corresponding language to that ethnicity.

Step 2: If there is no language name identical to the ethnicity then I check the alternate names for a language. If an ethnic name matches an alternate language name, I assign the corresponding language to that ethnicity.

Step 3: If a set of potential language matches still exist, I assign the largest language group (in terms of population) to the ethnic group.

D Supplementary Material

This section presents results referenced but not presented in the main body of the paper.

D.1 Various Fixed Effects Specifications

Table D1 reports 27 different estimates: 9 versions of equation (1) for each of the 3 linguistic similarity measures. Columns 1-3 report between-group estimates with country-year fixed effects, the estimates in columns 4-6 add country-language fixed effects, and columns 7-9 report estimates for the triple-difference estimator. For each set of three regressions I report estimates (i) without any covariates, (ii) estimates that only control for log population density and the logged geodesic distance between each partitioned group and the corresponding leader's group, and (iii) the full set of covariates I outlined in Section 3.

Consistent with my hypothesis of ethnolinguistic favoritism, all 27 coefficients are positive and the majority are statistically significant. In all cases my preferred measure of lexicostatistical similarity is significant with the exception of column 4, where lexicostatistical similarity has a reported p-value of 0.127. However, in this instance, the estimator lacks language-year fixed effects and thus does not exploit the counterfactual comparison of the same language group on the other side of the border.

Indeed, the addition of language-year fixed effects in 7-9 adds considerable precision to the estimates relative to columns 4-6. The allowance of a within-group estimator that comes from having a panel of partitioned language groups substantially improves my ability to identify ethnolinguistic favoritism.

I also provide estimates for cladistic similarity and coethnicity to see how these alternative measures compare to lexicostatistical similarity. For my benchmark estimates both coefficients are positive and statistically significant, albeit only at the 10 percent level. Not only does the estimated coefficient monotonically increase in the measured continuity of linguistic similarity, but lexicostatistical similarity is also more precisely estimated than both alternative measures. This suggests that the observable variation among non-coethnic groups assists in identifying patterns of ethnic favoritism in Africa.

D.2. Additional Controls

In this section I reproduce the benchmark estimates with two additional control variables: the Malaria Ecology Index (Kiszewski et al., 2004) and the Agricultural Suitability Index (Ramankutty et al., 2002). The trouble with these data is that in a number of instances a single raster cell covers an area larger than a country-language group partition because these

data are only available at a spatial resolution of $0.5° \times 0.5°$ (approximately 111 km $\times$ 111 km). These partitions are dropped from group average calculations, resulting in a sample 61.5 percent of the benchmark sample size.

Table D2 reports these subsample estimates that include the additional control variables. For each of the three measures of similarity I report estimates that include the absolute difference in the Malaria Ecology Index, the absolute difference in the Agricultural Suitability Index and estimates that include both measures, in addition to benchmark set of controls. The results are unchanged by including these controls.

D.3 Measurement Error

When an unambiguous assignment of a leader's ethnolinguistic identity cannot be made, I assign the group with the largest population among the set of potential matches. The finding that favoritism exists among groups that are not coethnic to the leader might be driven by the measurement error introduced by this approach.

In this section I report estimates on a subsample of my benchmark dataset that excludes the 4 leaders I could not unambiguously match.[39] Table D3 reports these results. Overall little is changed from my benchmark estimates, with the exception that coethnicity is no longer significant at standard levels of confidence. However, lexicostatistical similarity is roust to these excluded leaders, and most importantly, column (4) of Table D3 makes clear that the significance of non-coethnic similarity is not a consequence of the possible measurement error introduced when assigning an ethnolinguistic identity to the aforementioned leaders.

D.4 Balanced Panel

In this section I test the robustness of the benchmark estimates using a balanced panel of country-language groups between 1992 and 2013. My benchmark panel was unbalanced because of missing data on language lists used to estimate lexciostatistical similarity. This is problematic if these lists are missing for non-random reasons (Cameron and Trivedi, 2005). To check this I limit the analysis to a balanced sample of 84 language groups partitioned across 23 countries. Table D4 reports these estimates.

In all 27 reported regressions the measure of linguistic similarity takes the expected positive sign positive. For my preferred measure of lexicostatistical similarity the coefficients are statistically significant in all but one regression. The magnitudes of the estimates are

[39] Mobutu Sese Seko (DRC), Joseph Kabila (DRC), Laurent-Desire Kabila (DRC) and Goodluck Jonathan (Nigeria).

also relatively similar to my benchmark estimates. To the contrary cladistic similarity seems to be quite sensitive to this subsample and in only significant in a single instance. The coethnic results are similar to those in Table 3.

D.5 Weighted Regressions

In this section I test for heteroskedasticity in my benchmark estimates by weighting regressions by the Ethnologue population of each language group. The idea is that the measure of night light intensity is an average within each country-language group, and it is likely to have more variance in places where the population is small (Solon et al., 2015). Table D5 reports these estimates.

The lexicostatistical estimates are less sensitive to weighting than the cladistic and coethnic estimates. While a few lexicostatistical estimates lose their significance in columns (4)-(6), these estimates do not exploit language-year fixed effects, and hence are not identified off the exogenous within-group variation. In my benchmark specification in column (9), the effect of lexicostatistical similarity is significant at the 5 percent level and very similar to the benchmark estimate in terms of magnitude.

D.6 Alternative Night Light Transformations

The log transformation used throughout the regional analysis is without a doubt arbitrary. The use of this transformation has become the convention when using these night lights data so I follow the literature in my chose to add 0.01 to the log transformation. Nonetheless, I experiment with two alternative transformations in Table D6.

In columns (1)-(3) I report estimates where the dependent variable is defined as the square root of the raw night lights data. In columns (4)-(6) I log the night lights data without adding a constant. The latter results in a substantial loss of observations due to the fact that 40 percent of the observations exhibit zero night light activity. Because I must observe a partitioned group on both sides of the border for any year, I lose nearly 60 percent of my benchmark sample using this log transformation.

I find that the lexicostatistical estimate is robust to both transformations, while the cladistic is only robust to the square root transformation. Coethnicity remains positive but loses its statistical significance is both instances.

D.7 DHS Additional Tables

Table D9 reports 15 estimates: 5 separate specifications for both locational and individual similarity, and the same five specifications for the joint similarity estimates. In all specifications I adjust standard errors for clustering in country-wave-locational-language areas.

The top panel reports estimates for locational similarity. In column (1) the coefficient takes the expected positive sign, but is insignificant because the standard error is estimated to be quite large. However, in this specification I do not account for any individual characteristics, including whether a respondent lives in a rural location. Young (2013) shows that the urban-rural income gap accounts for 40 percent of mean country inequality in a sample of 65 DHS countries. In column (2) I report an estimate that includes a rural indicator variable. Indeed, the inclusion of this indicator substantially improves the precision of estimation, where locational similarity is now significant at the 1 percent level. In column (3) I add a set of individual controls.[40] The magnitude of locational similarity increases slightly and maintains its strong significant effect on individual wealth. In Table D9 I add each individual control variable one at a time. While I account for capital city effects with an indicator variable, I also account for additional spatial effects in columns (4) and (5) by separately adding the geodesic distance to the nearest coast and border.[41]

The middle panel of Table D7 reports estimates for individual similarity. While all coefficients take the expected positive sign, only a single estimate of individual similarity is statistically significant. When I do not control for any covariates the effect of individual similarity is very precisely estimated. To the contrary, the effect goes away once I account for respondents living in rural locations. The same is true when including the full set of controls.

Next I jointly estimate both channels using the aforementioned variation among individuals non-native to the region in which they reside. The results are consistent with the rest of the table and reported in the bottom panel of Table D7. In column (1) the estimate for individual similarity outperforms locational similarity when no individual characteristics are accounted for, however the reverse is true in columns (2)-(5) as covariates are incrementally added – in particular the rural indicator.

To show that the locational mechanism is not only driven by the coethnic effect, I separately estimate locational coethnicity and non-coethnic locational similarity. I do this in the same way I did in the regional-level analysis: I define non-coethnic locational similar-

[40] The set of individual controls include age, age squared, a female indicator, a rural indicator, a capital city indicator, 5 education fixed effects and 7 religion fixed effects. See Appendix A for variable definitions.

[41] I include distances separately because language areas tend to be fairly small, so location clusters in a partition are usually very close together and distance measures are highly collinear.

ity as (1 − coethnicity) × locational similarity. Table D8 reports these estimates. While non-coethnic locational similarity is estimated to be no different than zero in the most basic regression, once again after the baseline set of controls are added both the coethnic and non-coethnic effect are positive and strongly significant. Using the more conservative estimates of column (5), this suggests that the average level of non-coethnic locational similarity (0.164) yields an increase of 0.094 (= 0.164 × 0.573) in the wealth index – roughly one fourth the coethnic effect.

Finally, I also report the DHS estimates for locational similarity and include each baseline covariate one at a time. The idea here is to highlight the relative importance of controlling for the urban-rural inequality gap when using the DHS wealth index (Young, 2013). Table D9 reports these estimates.

Indeed I find that the precision of the locational similarity estimate is substantially improved by including an indicator variable for respondents living in rural regions. While many of the other covariates are themselves positive, no other variable have such a large confounding effect on locational similarity in its absence.

D.8 Coalition Building

Data

I use data from Francois et al. (2015) on the share of an ethnic group's representation in the governing coalition for 15 African countries.[42] These data are available at a yearly interval until 2004 for the ethnic groups listed in Alesina et al. (2003) and Fearon (2003). Because the unit of observation is an ethnic group, I assign an Ethnologue language group to each ethnicity using the assignment strategy outlined in Appendix C.[43] I measure the lexicostatistical similarity of these groups to the ethnolinguistic identity of the national leader between 1992 and 2004 using the leader data described in Section 2.3. In each country a residual ethnic categorization named Other is assigned to capture all groups outside of a country's major ethnic groups. Because Others lack a single ethnolinguistic identity, I assign Other groups a value of zero percent similarity to their leader.

[42]Benin, Cameroon, Cote dIvoire, Democratic Republic of Congo, Gabon, Ghana, Guinea, Liberia, Nigeria, Republic of Congo, Sierra Leone, Tanzania, Togo, Kenya, and Uganda.

[43]For 87.5 percent of the 264 ethnic groups not listed as "Other", the name of the ethnic group unambiguously corresponds to an Ethnologue name or alternative name in the country in which the group resides. Only 12.5 percent of groups require I use population as a tie breaker when multiple languages can be mapped to an ethnicity. 51 of the assigned languages do not possess an ASJP language list and thus are dropped from the analysis.

Results

I report estimates of equation (2) in Table D10. Column 1 replicates the main estimate of Francois et al. (2015) on the subset of data that I observe lexicostatistical similarity. The coefficient for coethnicity takes the expected positive sign, implying there is a 9 percent increase in the leader's group share of the governing coalition over and above the ministerial appointments made in accordance with the leader's group size. The magnitude of this coefficient is slightly smaller than the comparable coefficient in Francois et al.'s (2015) Table III. This suggests that, if anything, this subsample biases the coefficient downward. Column 2 corroborates this result using lexicostatistical similarity in place of coethnicity. In column 3, I separate the effect of coethnicity from lexicostatistical similarity using the same approach I used in Section 4; i.e., non-coethnic similarity = (1−coethnicity)×lexicostatistical similarity. The reported estimates in column 3 confirm that linguistic similarity predicts a group's representation in the governing coalition even among non-coethnic groups.

In columns 4-6 I explore the allocation of top positions in the governing coalition, and in columns 7-9 the allocation of positions outside of the top.[44] In all cases the variables of interest are positive and statistically significant. The most notable observation in this table is remarkable consistency in the magnitude of non-coethnic similarity across specifications. Related groups outside of the leader's ethnic group benefit from receiving positions both low and high in the hierarchy of government.[45]

[44]Top positions include the president and deputies, as well as ministers of defence, budget, commerce, finance, treasury, economy, agriculture, justice, and state/foreign affairs.

[45]Though not reported here, the estimates for group size are statistically significant in all instances. The estimates are also comparable in magnitude to those in Table 3 of Francois et al. (2015), and similarity show evidence of concavity in the effect of group size.

Table D1: Benchmark Regressions Using Various Combinations of Fixed Effects

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	1.292***	0.806***	0.936***	0.115	0.200**	0.213**	0.244**	0.297**	0.305***
	(0.255)	(0.306)	(0.318)	(0.075)	(0.087)	(0.088)	(0.112)	(0.120)	(0.116)
Adjusted R^2	0.342	0.428	0.452	0.921	0.921	0.922	0.925	0.925	0.926
Cladistic similarity$_{t-1}$	0.835***	0.488**	0.446**	0.044	0.065	0.058	0.221**	0.219**	0.185*
	(0.199)	(0.205)	(0.203)	(0.064)	(0.066)	(0.068)	(0.104)	(0.102)	(0.103)
Adjusted R^2	0.331	0.428	0.449	0.921	0.921	0.921	0.925	0.925	0.925
Coethnic$_{t-1}$	1.058***	0.386	0.648**	0.092	0.193**	0.202**	0.130	0.139	0.168*
	(0.244)	(0.325)	(0.314)	(0.064)	(0.084)	(0.082)	(0.099)	(0.098)	(0.094)
Adjusted R^2	0.332	0.423	0.447	0.921	0.921	0.922	0.925	0.925	0.925
Geographic controls	No	No	Yes	No	No	Yes	No	No	Yes
Distance & population density	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Language-year fixed effects	No	No	No	No	No	No	Yes	Yes	Yes
Country-language fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports benchmark estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t - 1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if either region contains diamond and oil deposits. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D2: Robustness Check: Benchmark Regressions with Additional Control Variables

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.384***	0.368***	0.380***						
	(0.120)	(0.122)	(0.120)						
Cladistic similarity$_{t-1}$				0.255**	0.242**	0.256**			
				(0.114)	(0.111)	(0.111)			
Coethnic$_{t-1}$							0.271**	0.257**	0.269**
							(0.108)	(0.109)	(0.109)
Malaria control	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Land suitability control	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	228	228	228	228	228	228	228	228	228
Countries	33	33	33	33	33	33	33	33	33
Language groups	105	105	105	105	105	105	105	105	105
Adjusted R^2	0.950	0.949	0.950	0.949	0.949	0.949	0.949	0.949	0.949
Observations	4,065	4,065	4,065	4,065	4,065	4,065	4,065	4,065	4,065

This table reports estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if either region contains diamond and oil deposits. The malaria controls measures the absolute difference in the Malaria Ecology Index between leader and country-language groups, while the land suitability control measures the absolute difference in Ramankutty et al.'s (2002) Agricultural Suitability Index. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D3: Robustness Check: Excluding Leaders with Ambiguous Ethnolinguistic Identities

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$					
	(1)	(2)	(3)	(4)	(5)
Lexicostatistical similarity$_{t-1}$	0.278**				
	(0.116)				
Cladistic similarity$_{t-1}$		0.199*			
		(0.108)			
Coethnicity$_{t-1}$			0.145	0.229**	0.218*
			(0.095)	(0.104)	(0.112)
Non-coethnic lexicostatistical similarity$_{t-1}$				0.480**	
				(0.237)	
Non-coethnic cladistic similarity$_{t-1}$					0.185
					(0.130)
Geographic controls	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	314	314	314	314	314
Countries	34	34	34	34	34
Language groups	144	144	144	144	144
Adjusted R^2	0.922	0.922	0.922	0.922	0.922
Observations	5,745	5,745	5,745	5,745	5,745

This table reports estimates from a subsample that excludes all ambiguous leadership assignments. Because these problematic assignments introduce measurement error, excluding them from the analysis ensures that the results are not a consequence of measurement. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. The same log transformation of the dependent variable is used for the lagged value of night lights, i.e., $\ln(0.01 + NightLights_{c,l,t-1})$. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D4: Robustness Check: Benchmark Regressions on a Balanced Panel

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.500**	0.563**	0.542***						
	(0.200)	(0.222)	(0.206)						
Cladistic similarity$_{t-1}$				0.491**	0.460*	0.407*			
				(0.231)	(0.238)	(0.238)			
Coethnic$_{t-1}$							0.328	0.337	0.338*
							(0.198)	(0.209)	(0.185)
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	177	177	177	177	177	177	177	177	177
Countries	23	23	23	23	23	23	23	23	23
Language groups	84	84	84	84	84	84	84	84	84
Adjusted R^2	0.921	0.921	0.921	0.921	0.920	0.921	0.920	0.920	0.921
Observations	3,894	3,894	3,894	3,894	3,894	3,894	3,894	3,894	3,894

This table reproduces benchmark estimates on a balanced subset of the panel dataset. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D5: Robustness Check: Benchmark Regressions Weighted by Language Group Population

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.231**	0.329**	0.308**						
	(0.105)	(0.141)	(0.124)						
Cladistic similarity$_{t-1}$				0.202*	0.213**	0.190*			
				(0.103)	(0.108)	(0.107)			
Coethnic$_{t-1}$							0.161*	0.234**	0.260***
							(0.090)	(0.095)	(0.094)
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Adjusted R^2	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.990
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports the benchmark estimates weighted by Ethnologue language group population. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D6: Robustness Check: Benchmark Regressions with Alternative Dependent Variables

	$\sqrt{\text{NightLights}_{c,l,t}}$			$\ln(NightLights_{c,l,t})$		
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.038**			0.396**		
	(0.018)			(0.191)		
Cladistic similarity$_{t-1}$		0.029*			0.189	
		(0.016)			(0.163)	
Coethnic$_{t-1}$			0.012			0.258*
			(0.014)			(0.138)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	214	214	214
Countries	35	35	35	33	33	33
Language groups	164	164	164	98	98	98
Adjusted R^2	0.952	0.952	0.952	0.935	0.935	0.935
Observations	6,610	6,610	6,610	2,921	2,921	2,921

This table tests the robustness of the dependent variable using two alternative transformations: a square root of the raw night lights data ($\sqrt{\text{NightLights}_{c,l,t}}$) and the natural log of the raw night lights data without a constant term ($\ln(NightLights_{c,l,t})$). Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D7: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index					
	(1)	(2)	(3)	(4)	(5)
Locational similarity$_{t-1}$	0.594	0.463***	0.479***	0.643***	0.365**
	(0.613)	(0.152)	(0.119)	(0.153)	(0.140)
Adjusted R^2	0.312	0.574	0.603	0.603	0.604
Individual similarity$_{t-1}$	1.260***	0.123	0.211	0.228	0.219
	(0.359)	(0.220)	(0.219)	(0.219)	(0.215)
Adjusted R^2	0.313	0.574	0.602	0.603	0.604
Locational similarity$_{t-1}$	0.592	0.463***	0.479***	0.643***	0.364**
	(0.613)	(0.153)	(0.119)	(0.153)	(0.140)
Individual similarity$_{t-1}$	1.259***	0.122	0.211	0.230	0.218
	(0.359)	(0.220)	(0.219)	(0.219)	(0.215)
Adjusted R^2	0.313	0.574	0.603	0.603	0.604
Rural indicator	No	Yes	Yes	Yes	Yes
Individual controls	No	No	Yes	Yes	Yes
Distance to border	No	No	No	Yes	No
Distance to coast	No	No	No	No	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88
Countries	13	13	13	13	13
Language groups	20	20	20	20	20
Observations	56,455	56,455	56,455	56,455	56,455

This table provides estimates for two channels: the effect of individual and locational similarity on the DHS wealth index. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, a gender indicator variable, an indicator for respondents living in the capital city, 5 education fixed effects and 7 religion fixed effects. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D8: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index					
	(1)	(2)	(3)	(4)	(5)
Locational coethnicity$_{t-1}$	0.838*	0.485***	0.437***	0.324**	0.601***
	(0.430)	(0.139)	(0.116)	(0.134)	(0.160)
Non-coethnic locational similarity$_{t-1}$	-0.692	0.348*	0.697***	0.573***	0.854***
	(0.556)	(0.205)	(0.148)	(0.167)	(0.173)
Rural indicator	No	Yes	Yes	Yes	Yes
Individual controls	No	No	Yes	Yes	Yes
Distance to coast	No	No	No	Yes	No
Distance to border	No	No	No	No	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88
Countries	13	13	13	13	13
Language groups	20	20	20	20	20
Adjusted R^2	0.314	0.574	0.603	0.604	0.603
Observations	56,455	56,455	56,455	56,455	56,455

This table reports estimates that test for favoritism outside of coethnic language partitions. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, a gender indicator variable and an indicator for respondents living in the capital city. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D9: Individual-Level Regressions: Baseline Covariates

Dependent Variable: DHS Wealth Index									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Locational similarity$_{t-1}$	0.585	0.594	0.463***	0.636	0.490	1.024*	0.518	0.608	0.479***
	(0.604)	(0.613)	(0.152)	(0.398)	(0.637)	(0.592)	(0.587)	(0.399)	(0.119)
Age	-0.021***								-0.008
	(0.005)								(0.006)
Age squared	0.000***								0.000
	(0.000)								(0.000)
Female indicator		-0.010							0.112***
		(0.013)							(0.013)
Rural indicator			-1.846***						-1.606***
			(0.072)						(0.079)
Capital city indicator				1.502***					0.238***
				(0.053)					(0.053)
Distance to the coast					-0.001				
					(0.000)				
Distance to the border						-0.001*			
						(0.001)			
Religion FE	No	No	No	No	No	No	Yes	No	Yes
Education FE	No	No	No	No	No	No	No	Yes	Yes
Country-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location-language-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual-language-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88	88	88	88	88
Countries	13	13	13	13	13	13	13	13	13
Language groups	20	20	20	20	20	20	20	20	20
Adjusted R^2	0.316	0.312	0.574	0.342	0.314	0.317	0.317	0.416	0.603
Observations	56,455	56,455	56,455	56,455	56,455	56,455	56,455	56,455	56,455

This table establishes the impact of each baseline covariate used in Section 5. The unit of observation is an individual. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D10: Ethnic Favoritism and Coalition Power Sharing

	Share of cabinet positions			Share of top cabinet positions			Share of low cabinet positions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coethnicity$_t$	0.093***		0.100***	0.179***		0.185***	0.050***		0.057***
	(0.012)		(0.013)	(0.019)		(0.020)	(0.013)		(0.014)
Lexicostatistical similarity$_t$		0.095***			0.172***			0.057***	
		(0.013)			(0.021)			(0.013)	
Non-coethnic similarity$_t$			0.047**			0.047*			0.048**
			(0.018)			(0.022)			(0.019)
Group size controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Countries	15	15	15	15	15	15	15	15	15
Ethnic groups	187	187	187	187	187	187	187	187	187
Adjusted R^2	0.665	0.664	0.668	0.539	0.521	0.541	0.544	0.549	0.548
Observations	2,539	2,539	2,539	2,539	2,539	2,539	2,539	2,539	2,539

This table establishes that linguistic similarity predicts an ethnic group's share in the governing coalition of a country. The unit of observation is an ethnic group. The dependent variable in columns (1)-(3) is the share of cabinet positions of an ethnic group in the governing coalition, whereas in columns (4)-(6) and (7)-(9) the dependent variable measures the cabinet share of top positions and low positions. The group size controls include a time-invariant measure of an ethnic group's share of the national population and its polynomial. Standard errors are in parentheses and adjusted for clustering at the country level. * p < 0.10, ** p < 0.05, *** p < 0.01.