NBER WORKING PAPER SERIES

THE POLITICAL ECONOMY OF ETHNOLINGUISTIC CLEAVAGES

Klaus Desmet
Ignacio Ortuño-Ortín
Romain Wacziarg

Working Paper 15360
http://www.nber.org/papers/w15360

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2009

We thank Jim Fearon for helpful comments. Desmet gratefully acknowledges financial support from the Comunidad de Madrid (PROCIUDAD-CM), and the Spanish Ministry of Science (ECO2008-01300). Ortuño-Ortín gratefully acknowledges financial support from the Spanish Ministry of Science (SEJ2007-67135). Wacziarg gratefully acknowledges financial support from Stanford University's Presidential Fund for Innovation in International Studies and from UCLA's Center for International Business Education and Research. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Political Economy of Ethnolinguistic Cleavages
Klaus Desmet, Ignacio Ortuño-Ortín, and Romain Wacziarg
NBER Working Paper No. 15360
September 2009
JEL No. H1,N4,O4,O5

ABSTRACT

This paper proposes a new method to measure ethnolinguistic diversity and offers new results linking such diversity with a range of political economy outcomes -- civil conflict, redistribution, economic growth and the provision of public goods. We use linguistic trees, describing the genealogical relationship between the entire set of 6,912 world languages, to compute measures of fractionalization and polarization at different levels of linguistic aggregation. By doing so, we let the data inform us on which linguistic cleavages are most relevant, rather than making ad hoc choices of linguistic classifications. We find drastically different effects of linguistic diversity at different levels of aggregation: deep cleavages, originating thousands of years ago, lead to measures of diversity that are better predictors of civil conflict and redistribution than those that account for more recent and superficial divisions. The opposite pattern holds when it comes to the impact of linguistic diversity on growth and public goods provision, where finer distinctions between languages matter.

Klaus Desmet
Department of Economics
Universidad Carlos III
28903 Getafe (Madrid)
SPAIN
klaus.desmet@uc3m.es

Ignacio Ortuño-Ortín
Department of Economics
Universidad Carlos III
28903 Getafe (Madrid)
SPAIN
iortuno@eco.uc3m.es

Romain Wacziarg
Anderson School of Management at UCLA
C-510 Entrepreneurs Hall
110 Westwood Plaza
Los Angeles, CA 90095-1481
and NBER
wacziarg@ucla.edu

1 Introduction

How does ethnolinguistic diversity affect political and economic outcomes? In recent years, a vast literature has argued that such cultural heterogeneity impacts a wide range of outcomes, fostering civil war, undermining growth, hindering redistribution and the provision of public goods. However, evidence on this point remains subject to some disagreement. For instance, there is a vibrant debate on the role of ethnolinguistic divisions as determinants of civil wars.[1] Econometric results on growth, redistribution and public goods provision also vary widely across studies, raising issues of robustness.[2]

These inconclusive results may stem in part from the inability to convincingly define the ethnolinguistic groups used as primitives to construct measures of heterogeneity. When faced with the issue of how to define groups, researchers have either relied on readily available classifications, such as the ones based on the *Atlas Narodov Mira* or the *Encyclopedia Britannica*, or have carefully constructed their own classifications.[3] Both approaches are problematic: the former runs the risk of missing the relevant cleavages, whereas the latter is subject to the criticism that groups are defined based on how important they are expected to be for the problem at hand. In this paper, we propose a methodology that addresses both criticisms, and argue that the degree of coarseness of ethnolinguistic classifications has profound implications for inference on the role of diversity.

The methodology we propose computes diversity measures at different levels of aggregation. We do so by exploiting the information of language trees. We refer to this as a phylogenetic approach, since tree diagrams describe the family structure of world languages. Depending on how finely or coarsely groups are defined, the measure of ethnolinguistic diversity will be different. For example, if one takes the different dialects of Italian to constitute different groups, then Italy appears to be very diverse. However, if one considers these different dialects to be only minor variations of Italian, then Italy looks homogeneous. Apart from allowing us to classify languages at

[1] Fearon and Laitin (2003) show that ethnic fractionalization is not an important determinant of the onset of civil wars. Montalvo and Reynal-Querol (2005), in contrast, argue that ethnic polarization is a significant determinant of the incidence of civil conflict.

[2] Alesina et al. (2003) argue that while ethnic and linguistic fractionalization are usually negatively related to growth and the quality of government, the significance of these partial correlations is sensitive to the specification.

[3] For an excellent discussion of the difficulties raised by the issue of defining relevant or salient ethnolinguistic groups, see Alesina and La Ferrara (2005), section 5.2.1, page 792.

different levels of aggregation, this approach has the advantage of giving a historical dimension to our analysis. Coarse linguistic divisions, obtained at high levels of aggregation, describe cleavages that go back thousands of years. In contrast, finer divisions, obtained at low levels of aggregation, are the result of more recent cleavages. Since we rely on data that cover the entire set of $6,912$ world languages, and examine effects of heterogeneity measures computed at all possible levels of aggregation, we are able to capture a wide range of linguistic classifications. Rather than choosing the "correct" classification ourselves, we let the data inform us as to which linguistic cleavages are most relevant for different outcomes of interest.[4]

Our empirical analysis reveals drastically different effects of linguistic heterogeneity at different levels of aggregation. We also find that the relevant cleavages differ greatly across political economy outcomes. Starting from the data, specifications and estimation methods from major contributions to the literature on the political economy of ethnolinguistic diversity, we substitute our new measures of diversity for those commonly used. For civil conflict and the extent of redistribution, issues that inherently involve conflicts of interest, coarse divisions seem to matter most. While we find only weak evidence that diversity (whether measured by fractionalization or polarization) affects the onset of civil wars at any level of linguistic aggregation, the estimated effects do tend to be larger and more significant when considering a coarse classification. This finding is consistent with existing conflicts in African countries, such as Chad and Sudan, on the border between the Afro-Asiatic family and the Nilo-Saharan family. It may also help explain conflict in certain Latin American countries, such as Mexico and Bolivia, where the Indo-European family coexists with different Amerindian languages. For redistribution, the results are more robust, and suggest once again that measures based on a high level of aggregation matter most. In contrast, for economic growth, where coordination between individuals or groups is essential and market integration is important, we find that finer divisions lead to heterogeneity measures that matter more. The same pattern holds across a wide array of measures of public goods provision.

Thus, when the main issue involves conflicts of interest (as for the onset of civil wars and

[4]Our approach is related to existing work arguing that people identify with different groups in different contexts (particularly the work of Crawford Young on situational identity - see Young, 1976). For instance, ethnolinguistic cleavages that matter for voting behavior in local elections may differ from those that matter for national elections. For a related point, see Posner's 2005 book on ethnic politics in Zambia. More generally, cleavages that matter for some outcomes may not matter for others. There is no such thing as a "correct" classification of languages or ethnicities - this depends on the context.

the extent of redistribution), deep differences originating thousands of years ago matter most: different groups' interests differ more when cleavages are more deeply rooted. In contrast, more superficial and recent divisions appear sufficient to hinder growth, an outcome related to the ease of coordination. For instance, to the extent that clusters of economic activity form around language lines, linguistic divisions may limit the integration of markets, and prevent economic growth. Even though Hindi and Gujarati are not so different, this linguistic cleavage may hinder the integration of the corresponding regions of India. What matters here is whether two individuals or groups can interact effectively. In fact, finer linguistic classifications deliver heterogeneity measures that matter more for outcomes such as economic growth, which is hindered by lack of coordination and integration. As for public goods, they fall somewhere inbetween both cases: although they have a redistributive aspect, their effective provision also requires coordination between groups or individuals. Empirically, we find that fine linguistic divisions, based on more superficial cleavages, hinder public goods provision across a wide array of indicators.

This paper is related to a vast literature in political economy. Various authors have studied how ethnolinguistic diversity affects redistribution, growth and civil conflict (Easterly and Levine, 1997; La Porta et al., 1999; Alesina et al., 2003; Fearon and Laitin, 2003; Alesina and La Ferrara, 2005, Alesina, Baqir and Easterly, 1999, among many others). Measurement issues are central to recent research on these topics. One issue is that standard indices of diversity do not take into account the distance between groups (Fearon, 2003; Desmet et al., 2009; Spolaore and Wacziarg, 2009). Another possibility is that for certain issues, such as civil conflict, polarization may be more relevant than fractionalization (Esteban and Ray, 1994; Montalvo and Reynal-Querol, 2005), an issue we revisit below. A third problem is the difficulty of determining the right level of aggregation when computing heterogeneity measures, i.e., identifying the relevant ethnolinguistic cleavages. This issue has received little attention, and it is the main focus of the present study.[5]

This paper is organized as follows. Section 2 describes conceptual issues related to the measurement of heterogeneity based on language trees, and describes the data. Section 3 discusses the effects of diversity on civil conflict and redistribution. Section 4 covers the effects on public goods

[5]Fearon (2003) does discuss at lengths the issue of how to define the "right list" of ethnic groups serving as the basis for computing heterogeneity measures, and recognizes explicitly that not all cleavages may be relevant for a given outcome. However, he presents data on ethnic groups based on a single classification. Scarritt and Mozaffar (1999) present data on ethnic groups for Sub-Saharan countries using three different classifications, but do not examine the effects of using these different classifications on political and economic outcomes.

provision and economic growth. Section 5 concludes.

2 Aggregation and Linguistic Diversity

2.1 A Tale of Two Countries

To illustrate our approach, we start with a comparative case study. Over the period 1965-2000, Chad and Zambia experienced some of the lowest growth rates on the globe, their income per capita shrinking by an average of 1 percentage point per year (Table 1). The 2005 Human Development Index ranked Chad 170 and Zambia 165 out of a total of 177 countries. It has long been argued that low growth may be related to high ethnolinguistic diversity. With 135 languages spoken in Chad, and between 40 and 70 in Zambia, these countries certainly are very diverse: taking the commonly used fractionalization index as a measure of diversity, the *Ethnologue* database on languages gives a value of 0.95 for Chad and 0.85 for Zambia, putting both countries in the top decile. As highlighted by Easterly and Levine (1997), data for a broad cross-section of countries point more formally to a general negative relationship between ethnic heterogeneity and economic performance. In our data, the 10% most diverse countries had an average per capita growth rate of a meager 0.54% over the period 1960-2004, whereas the 10% least diverse countries posted a much more sturdy figure of 2.59% (linguistic diversity here is measured using the most disaggregated classification of languages).

In spite of their high ethnolinguistic fractionalization, in terms of conflict and civil war Chad and Zambia have been at opposite sides of the spectrum. Chad has been at war almost continuously since independence, whereas Zambia has not witnessed any civil conflict worth speaking of. In Chad, during colonization, and after independence in 1960, the Christian South was privileged, and formed the political elite, to the detriment of the Islamic and partly Arab-speaking North. Dissatisfaction by the North led to a civil war, which started in 1965, and lasted for about a decade and a half, culminating in the rebels taking over the capital and ending Southern dominance. Since then the country has remained unstable, partly because of the inverted power relation, with the North now dominating the South, but also because of power struggles within these regions. In recent years, for example, there has been increasing ethnic tension between the Zaghawa and Tama, two non-Arab groups. Zambia, in contrast, has had a history of peaceful coexistence between the many groups and tribes. Although voting behavior in Zambia tends to run along language groups (Posner, 2003), it has not led to the violence seen in countries such as Chad. Income redistribution,

which is an issue involving divergence of interests, is often interpreted as related to conflict. Data on redistribution confirm the contrast between both countries: figures on transfers and subsidies as a share of GDP reveal that on average between 1985 and 1995 Chad redistributed 0.9% of GDP, compared to 3.8% in Zambia.

This example illustrates the main point of this paper: although commonly used measures of diversity make Chad and Zambia look very similar, those measures mask one important difference between these countries in terms of diversity. Of the total population in Chad, one third speaks an Afro-Asiatic languages, a little over half a Nilo-Saharan languages, and the rest a language of the Niger-Congo family. In contrast, in Zambia, 99.5% of the population speaks a language from the Niger-Congo family. This raises an important point: whereas Chad and Zambia are amongst the most diverse countries on the globe, when considering language families rather than individual languages, we obtain a very different picture. While Chad continues to be one of the most diverse countries, ranking 7 out of 225, Zambia now looks very homogeneous, ranking 176 out of 225, similar to Portugal. In other words, when taking every language as being different, Zambia is very diverse, similar to Chad, whereas when aggregating into language families, Zambia no longer appears to be quite so heterogeneous.

The experience of Chad and Zambia suggests that the type of diversity that matters for economic growth is different from the type of diversity that matters for civil conflict and redistribution. The essential difference between the two types of diversity is the degree of aggregation. The relevant degree of aggregation, and thus the relevant definition of a group, depends on the problem at hand. This case study suggests that, for economic growth, fine differences between languages may matter, whereas for civil conflict and redistribution, only coarse differences may play a role - as is confirmed below in large samples.[6]

[6] The difference in the experience of Chad and Zambia with conflict and redistribution is not related to the use of measures of linguistic fractionalization rather than polarization, but to the issue of aggregation. As Table 1 reveals, using a standard measure of polarization instead of fractionalization leads to the same conclusion: the difference in polarization between Zambia and Chad is much more pronounced for highly aggregated linguistic classifications than for disaggregated ones. Correspondingly, conflict and war has been continuous in Chad, but absent in Zambia. We discuss the important issue of how the distinction between polarization and fractionalization (which has to do with the functional form used to calculate measures of diversity) relates to the level aggregation (which has to do with the definition of relevant groups) in Section 2.3.

2.2 Language Trees and Linguistic Diversity

2.2.1 The Construction of Language Trees

This paper seeks to measure linguistic diversity at different levels of aggregation. To do so, we use language trees. We refer to this as a phylogenetic approach (as the linguistics literature does), referring to the fact that tree diagrams capture the genealogy of languages, classified in terms of their family structure.[7] Using language trees gives a historical dimension to our analysis. Coarse linguistic divisions, such as that between Indo-European and non Indo-European languages, describe cleavages that originate thousands of years ago. In contrast, finer divisions, such as that between Dutch and German, tend to be the result of more recent splits. For instance, Gray and Atkinson (2003) estimate separation times between language groups within the Indo-European family. While the separation between Indo-European languages and all others is estimated to have occurred prior to $8,700$ years ago, the separation time between different dialects of Greek is estimated to have occurred only 800 years ago. There are differences of opinion between linguists on the precise dates, but the general point of an association between tree structure and separation times remains. We do not require that there be a strict association between the coarseness of the linguistic classification and the time since the linguistic split between groups occurred - we only point out that coarse classifications capture cleavages that tend to go back deeper in the past.

Linguistic differentiation occurs because specific human populations become relatively isolated from each other and, as a result, develop specific languages over time. In general three major factors can affect the degree to which languages differ. The first factor is the time since the populations speaking these languages have split from each other. As noted, populations speaking French and Spanish have split from each other much more recently than populations speaking, say, Swahili

[7] This point was recognized going at least as far back as Charles Darwin, who wrote: "If we possessed a perfect pedigree of the mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient language had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilization of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock, would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue." (Darwin, 1902, p. 380).

and Tibetan. The second factor, known by linguists as *Sprachbund* (or language union), results from interactions between populations that are already linguistically distinct (Emeneau and Anwar, 1980). For example, historically the spread of Latin words likely had a homogenizing influence on European languages, keeping Romance and Germanic languages more similar than would have been the case without commercial and political interactions. The third factor is the size of the population. Linguistic drift tends to be faster in smaller populations. For instance, Lithgow (1973) studies the Muyuw language, spoken on Woodlark Island (New Guinea): 13% of the Muyuw vocabulary was replaced in a period of 50 years during the middle of the 20^{th} Century (see also Dixon, 1997, for a discussion). This language is spoken by only $6,000$ individuals, according to *Ethnologue*. Empirically, this determinant of linguistic differentiation does not greatly affect our measures of diversity, as it only affects very small linguistic groups.

Linguistic trees such as those from *Ethnologue*, which we use in our empirical analysis, are constructed by linguists to capture the first factor.[8] That is, Spanish is a closer cousin of French than Swahili is of Tibetan. Higher levels of aggregation describe deeper ethnolinguistic cleavages, while differences between languages that are noticeable at lower levels of aggregation only reflect more superficial cleavages. The degree of linguistic diversity considered at these different levels of aggregation differs, and this is the variation we exploit in our empirical work.

We emphasize that the issue of aggregation is separate from (although related to) the issue of how to capture the distance separating languages when computing measures of diversity (for a paper that accomplishes the latter goal, see Desmet et al., 2009; for indices of fractionalization that take into account distances, see Greenberg, 1956 and Bossert, D'Ambrosio and La Ferrara, 2009). We are after identifying the level of aggregation that corresponds to the most relevant cleavages for the various dependent variables we examine. A focus on the level of aggregation that captures the relevant cleavages retains a strong focus on ethnolinguistic groups as the basis for individuals' identification with, or alienation from, a given ethnolinguistic identity or group (we borrow the identification/alienation terminology from Esteban and Ray, 1994). In contrast, distance-weighted measures of diversity (such as the measure proposed by Greenberg, 1956), capture the expected

[8] There are controversies among linguists on the right classification of languages. For example, Greenberg (1987) considers that all Native American languages can be classified into three groups (Eskimo-Aleut, Na-Dene, Amerindian) whereas the *Ethnologue* contemplates dozens of unrelated families. However, the classification provided in the *Ethnologue* is the most widely used and, to the best of our knowledge, the only one available in electronic format covering all of the languages of the world.

distance between individuals, and relegates the group structure to the background. Our approach is therefore distinct from approaches that make use of distances between groups: we are interested in identifying the group structures (or classifications) that matter most for political economy outcomes. At the same time, by construction, more aggregated classifications retain groups that tend to me more distant from each other (in terms, say, of separation times, or in terms of how different the languages are), compared to more disaggregated classifications.

To illustrate the discussion above, Figure 1 displays the tree for the major languages in Pakistan. On the left side of the figure, we list the level of aggregation. At level 7, the most disaggregated level, there are seven main languages: Panjabi, Pashto, Sindhi, Seraiki, Urdu, Balochi, and Brahui. Going up the tree, the number of groups declines, as the level of aggregation rises. For instance, at level 4, there are only 5 linguistic groups - at that level, Panjabi, Seraiki and Sindhi are classified as one and the same. At level 3, only linguistic groups are left (Iranian, Indo-Aryan and Dravidian). Finally, at aggregation level one, there are two groups: Dravidian (Brahui) and Indo-European (all others). These classifications allow us to compute measures of diversity at each level of aggregation.

2.2.2 Measuring Diversity at Different Levels of Aggregation

How precisely are the measures of diversity computed? A typical tree is represented in Figure 2. The *root* of the tree is represented by the upper-case letter O, whereas the *leafs* of the tree are represented by lower-case letters a through g. In Figure 2, all leafs have a common root, so that the tree is rooted (this terminology is borrowed from the field of linguistics). As can be seen, the tree has three different levels. Each of the seven leafs at level 2 represent a living language. The three nodes at level 1 represent the (extinguished) mother languages of the existing languages. The node at level 0 represents the common ancestor language of the three mother languages. The number below each living language at level 2 indicates the assumed shares of the population speaking the corresponding language. The numbers below the (extinguished) mother languages at level 1 are the aggregated population shares of their corresponding daughter languages.

To compute diversity at different levels, we require that the tree be rooted, and that the number of *branches* (or *edges*) between any leaf and the root be identical. In this subsection, we focus on the widely used index of ethnolinguistic fractionalization (or ELF), the probability that two randomly picked individuals belong to different groups (in our empirical work we also consider measures of polarization). The diversity measure at a given level of aggregation is the ELF index for the

linguistic groups as they appear at that level. For example, diversity at level 2 is given by the ELF index, taking the seven living languages as the relevant groups. Thus, $ELF(2) = 1 - 3 \times (0.2^2) - 4 \times (0.1^2) = 0.84$. To calculate diversity at level 1, the seven living languages are aggregated into 3 distinct groups A, B and C, resulting in an ELF index $ELF(1) = 1 - 0.4^2 - 2 \times (0.1^2) = 0.66$.

One difficulty remains. The linguistic trees from *Ethnologue* are all rooted trees, but the number of branches varies among linguistic families and subfamilies. Figure 3 represents a typical language tree from *Ethnologue*. As can be seen, language A has more descendent generations than languages B or C. As before, the leafs of the tree represent the existing languages. They are denoted by the letters $a11$, $a12$, $a21$, $a22$, $a31$, $a32$, $b1$, $b2$ and c. It is clear that for this type of tree we cannot use the method applied in Figure 2, because at level 3 we would be ignoring 3 of the 7 languages. The branches in the trees need to be extended, and there are two main ways to do this, as displayed in the two panels of Figure 3. This ensures that all the existing languages are represented as leafs at the lowest level of aggregation.

The first approach, displayed in Panel I of Figure 4, assumes that all living languages are equally distant from the root, where the distance between languages is defined by the number of branches or nodes separating them (in technical terms, this assumes that the tree is *ultrametric*). Take, for example, language C. We insert two fictitious languages, $c1$ and $c11$, at levels 1 and 2, so that the total number of branches between C and O is the same as for all other leafs. The second approach, displayed in Panel II of Figure 4, assumes that C is only one branch removed from the root O. In this case, Figure 4 shows that to have all living languages at the same level, we move C down to level 3, but assume that its mother, grandmother and great-grandmother have all remained the same as the origin language O. In our empirical work, we use measures based on the first approach, as it seems natural to assume that languages went through intermediate states between their origin languages and their current form. We also think it unlikely that origin languages remained unchanged until a recent date. However, for the sake of robustness we also computed and used measures based on the second approach, and using either approach did not make much difference for our results (estimates are available upon request).

Completing the tree under either approach, we can apply the method used in the example of Figure 2 to compute the degree of diversity at the different levels of the tree. Notice that at the lowest level (level 3) both approaches yield the same degree of diversity since the different groups are just the existing languages. It is easy to see, however, that the two approaches in general do

not yield the same degree of diversity at other levels of the tree.

2.3 Measures of Diversity and Summary Statistics

We consider two sets of commonly-used measures of diversity: fractionalization and polarization. For $i(j) = 1....N(j)$ groups of size $s_{i(j)}$, where $j = 1...J$ denotes the level of aggregation at which the group shares are considered, fractionalization is just the probability that two individuals chosen at random, will belong to different groups:

$$ELF(j) = 1 - \sum_{i(j)=1}^{N(j)} \left[s_{i(j)}\right]^2$$

This measure is maximized when each individual belongs to a different group. Polarization, in contrast, is maximized when there are two groups of equal size. We use the polarization measure from Montalvo and Reynal-Querol (2005). This index satisfies the conditions for a desirable index of polarization in the axiomatic approach of Esteban and Ray (1994):

$$POL(j) = 4 \sum_{i(j)=1}^{N(j)} \left[s_{i(j)}\right]^2 \left[1 - s_{i(j)}\right]$$

We compute these measures at each of the 15 levels of aggregation available in the linguistic classification in the 15^{th} edition of *Ethnologue*, the source for our linguistic data (Ethnologue, 2005). The sample contains 226 observations which include countries and their dependencies (due to data availability, our regression results are based on a smaller set of countries). To simplify the presentation, we focus on only 5 levels of aggregation (those are levels 1, 3, 6, 10 and 15, with higher numbers denoting a lower degree of aggregation). All our empirical results are also available at the intermediate levels. Table 2 presents summary statistics for the diversity measures at these 5 levels of aggregation, and Appendix 1 contains the corresponding data series by country. To facilitate the quantitative assessment of the regression results, Panel A displays means and standard deviations. When measured using the ELF index, the average degree of diversity rises as the level of aggregation falls, as expected. When measured using a polarization index, diversity falls at high levels of aggregation, and plateaus as aggregation falls further.

Interesting information can also be gleaned from Panel B of Table 2, displaying correlations.[9] First, changing the level of aggregation greatly affects the measures of diversity: the correlation

[9] We also investigated the pairwise correlations between our measures of diversity and measures commonly used in the literature. These correlations are maximized when using our most disaggregated measure. For instance, the

between $ELF(1)$ and $ELF(15)$ is only 0.526. Second, the correlation between polarization and fractionalization, at the same levels of aggregation, rises as the level of aggregation increases (the correlation between $POL(15)$ and $ELF(15)$ is only 0.555, while the correlation between $ELF(1)$ and $POL(1)$ is 0.988). This is intuitive as, when aggregating, fewer groups remain, and the distinction between polarization and fractionalization fades. Third, aggregating up is not the same as switching from a measure of fractionalization to a measure of polarization: the correlation between $ELF(1)$ and $POL(15)$ is only 0.391. This last observation indicates that the issue of aggregation is very different from the choice of functional form to compute diversity measures. In our empirical work, we show that switching from fractionalization to polarization measures has relatively benign effects on the substantive results, while changing the level of aggregation to compute either measure delivers vastly different estimates of the effect of diversity on political economy outcome.

Finally, Figures 5 and 6 display the full distributions of $ELF(1)$, $ELF(15)$, $POL(1)$ and $POL(15)$. As can be seen, at high levels of aggregation the distributions of both fractionalization ($ELF(1)$) and polarization ($POL(1)$) have a strong positive skew. This makes sense: when classifying languages to be different only when they pertain to entirely different families, most countries display low levels of diversity, and only a few exhibit high diversity. In contrast, at low levels of aggregation the distributions of fractionalization ($ELF(15)$) and polarization ($POL(15)$) are much more uniform. That is, many of the countries that were not diverse when only looking at language families are now much more diverse. This is the example of Zambia mentioned above: it is highly diverse if each of the 46 languages are taken to be different, and it is not very diverse when one considers that only 2 out of the 46 languages do not belong to the Niger-Congo family.

correlation of $ELF(15)$ with ethnic fractionalization from the *Atlas Narodov Mira* is 0.82, with the Alesina et al. (2003) measure of ethnic fractionalization, it is 0.67, with the Alesina et al. (2003) measure of linguistic fractionalization, it is 0.84, and with the Fearon (2003) measure of linguistic fractionalization, it is 0.75. These correlations fall to the $0.35 - 0.4$ range when using $ELF(1)$, the measure based on the most aggregated linguistic classification. Turning to religious fractionalization, the correlation between ELF(15) and religious fractionalization from the Alesina et al. (2003) dataset is 0.195, and at the level of ELF(1) this correlation drops to 0.098.

3 Linguistic Diversity, Civil Conflict and Redistribution

3.1 Civil Conflict

There is an ongoing academic debate on the relationship between ethnolinguistic diversity and the onset of civil conflict. In a seminal paper, Fearon and Laitin (2003) argued that once measures of income per capita are controlled for, measures of ethnic and religious fractionalization are unrelated to the *onset* of civil conflict. We reexamine this issue using the baseline specification in Fearon and Laitin's study (column 1 of their Table 1, page 84). Using exactly their data, their estimation method and their dependent variable (the onset of civil conflict), we simply substitute our measures of linguistic heterogeneity for their measure of ethnic fractionalization. Results are presented in Table 3 (for fractionalization) and Table 4 (for polarization).[10]

The first and most important observation is that the effect of fractionalization and the corresponding level of statistical significance both fall dramatically and monotonically when the level of aggregation falls. At level 1 (the most aggregated level), linguistic fractionalization has a coefficient of 1.06 with a t-statistic of 2.02, and the coefficient falls to -0.051 with a t-statistic of 0.14 at level 15. This pattern is robust to using polarization instead of fractionalization. The second observation is that the coefficient on linguistic diversity is only positive and significant when considering the most aggregated classification of languages - whether for polarization or for fractionalization. The coefficient remains significant at least at the 10% level for most of the robustness tests we conducted - but since the level of significance sometimes falls below 5% we want to be cautious in claiming that there exists a robust relationship even at this level of aggregation. A conservative reading of our results suggests that, to the extent there is a statistically significant link between diversity and civil conflict, it only appears when the relevant cleavages are the deepest (aggregation level 1). In terms of economic magnitude, the estimated effects are far from trivial at aggregation level 1. When evaluating marginal effects at the mean of all the independent variables, a one standard deviation change in linguistic fractionalization (0.173) is associated with an increase in the probability of conflict equal to roughly 11% of this variable's mean (the mean probability of civil war onset is 1.725% in the sample). This effect quickly fades to zero as the level of aggregation falls, as displayed graphically in Figure 7. The standardized magnitude is the same for polarization at aggregation level 1, and fades to zero even faster.

[10] The tables presents results at selected levels of aggregation, namely levels 1, 3, 6, 10 and 15 (results for all other levels are available upon request, but do not add much to the picture).

The pattern of coefficients across levels of aggregation is robust to a wide range of modifications of the baseline specification: 1) adding continent-level dummy variables, 2) substituting a dichotomous measure of democracy for the continuous one, 3) controlling for intermediate levels of democracy (anocracy), 4) redefining civil wars to only include "ethnic" civil wars (as defined in Fearon in Laitin, 2003), 5) using the Correlates of War definition of civil wars instead of Fearon and Laitin's, 6) controlling for GDP growth and lagged growth and 7) using the *incidence* of conflict rather than the onset, as Montalvo and Reynal-Querol (2005) did in their study. All these robustness tests are available upon request.

As shown in Figures 5 and 6, most countries in the world appear very homogeneous at level 1. Countries that do feature such cleavages tend to coincide with the geographic breakpoints of major linguistic groups, such as in Chad. Our results indicate that ethnolinguistic divisions of this nature may matter for civil conflict, but that more superficial divisions do not. Since there are few countries that feature high levels of diversity at the very aggregated level of linguistic families, civil conflict affected by this type of cleavage must be relatively rare.

Where does this leave us in the debate about the role of ethnolinguistic diversity as a determinant of civil wars? On the one hand, for all but one level of aggregation, ethnic diversity does not matter. As was recognized in the past literature, this does not imply that civil conflicts do not often have an ethnic dimension - conditional on having a civil conflict, it may very well be waged along ethnic or linguistic lines (for instance ethnolinguistic differences may help identify combatants, as in the famous Biblical example of the shibboleth). This is compatible with a finding that linguistic diversity does not affect the probability of conflict onset. On the other hand, we did find that the significance and magnitude of diversity rises as the level of aggregation increases. To the extent that civil conflict is caused by the "us" versus "them" divide, this result helps clarify that "us" and "them" need to be separated by deep historical and cultural cleavages for these divides to have any claim of affecting the onset of civil conflict.

3.2 Redistribution

A vast literature examines the role of ethnic and linguistic differences as a determinant of the extent of income redistribution. At the microeconomic level, several authors have examined the propensity to redistribute. For instance, Luttmer and Fong (2009) find in an experimental setting that people donate more money to Hurricane Katrina victims when the victims are perceived to be of the same

race as the donor. In another study, Luttmer (2001) reports that "individuals increase their support for welfare spending as the share of local recipients from their own racial group rises", using data from the United States, also suggesting a preference channel. These results are in line with those of Alesina, Glaeser and Sacerdote (2001), as well as Alesina and Glaeser (2004), arguing that the U.S. redistributes less than Europe in part because of its greater degree of racial heterogeneity.

At the cross-country level, results are more mixed. While the preponderance of evidence points to a negative association between ethnolinguistic fractionalization and redistribution, this finding is not always robust to the use of alternative measures of diversity and to the inclusion of controls. For instance, in Alesina et al. (2003), the effect of ethnolinguistic fractionalization on the share of transfers and subsidies to GDP appears sensitive, in terms of statistical significance, to the inclusion of several control variables. This study measures fractionalization using a rather disaggregated classification of ethnic and linguistic groups. In a broad cross-country sample, Desmet et al. (2009) find that linguistic diversity, measured to account for the distance between groups, is negatively associated with redistribution, measured by the share of transfers and subsidies in GDP. However, this result does not hold when measures of diversity do not account for the degree of linguistic distance between groups, suggesting that the depth of linguistic cleavages matters. In a wide variety of settings, ethnolinguistic diversity seems associated with lower redistribution, but what cleavages are more or less relevant to account for these findings has not been determined.

We use exactly the specification and data in Desmet et al. (2009) to examine what level of linguistic aggregation matters for redistribution, i.e. what are the relevant cleavages. The dependent variable is the average share of transfers and subsidies in GDP between 1985 and 1995. The specification is the one that involves the broadest set of control variables - including GDP per capita, country size, the percentage of the population over 65, legal origins and a variety of geographic variables (Table 2, column 8 in Desmet et al., 2009). Tables 5 and 6 present the results for, respectively, fractionalization and polarization. The results for both measures are similar, and reveal a striking pattern: linguistic diversity negatively affects redistribution at high levels of aggregation, but the effect declines in magnitude as the level of aggregation falls, and ceases to be statistically significant at the 5% after aggregation level 5. Figure 8 displays this pattern, plotting the standardized beta on fractionalization (i.e. the effect of a one standard deviation increase in fractionalization as a fraction of a one standard deviation change in the dependent variable) against the level of aggregation. The effect of $ELF(1)$ is substantial in magnitude, as it equals -8.7% and

is significant at the 5% level. It falls to -4.2% for $ELF(6)$ and ceases to be statistically significant. These results are robust to considering alternative sets of controls, as in Desmet et al. (2009), with the caveat that with a sufficiently restricted set of control variables, the effect of linguistic diversity remains statistically significant even at low levels of aggregation.

To summarize, we find that for redistribution, as for conflict, the relevant cleavages are those that capture deep ethnolinguistic splits, rather than divisions that are more recent and superficial. Commentators often point out that solidarity does not travel well across groups. We find that solidarity travels without trouble across groups that are separated by shallow gullies, but not across those separated by deep canyons. This is consistent with aforementioned studies arguing that racial animosity has negative effects on redistribution in the U.S., as those studies focus almost exclusively on the arguably deep cleavage between blacks and whites.

4 Linguistic Diversity, Public Goods and Growth

4.1 Public Goods

The effect of ethnolinguistic diversity on the provision of public goods raises interesting conceptual issues. On the one hand, public goods entail a dimension of redistribution, and differences in preferences may hinder their provision. In this sense, there is an element of conflict of interest when it comes to public goods. On the other hand, free rider problems and coordination failures need to be overcome for the effective provision of public goods. Linguistic diversity may work to affect public goods through both channels.

Several studies have explored the relationship between public goods provision and ethnolinguistic diversity, both across and within countries. In their important study of the cross-national determinants of the quality of government, La Porta, Lopez de Silanes, Shleifer and Vishny (1999, henceforth LLSV) showed that ethnolinguistic fractionalization, measured by an average of five existing indices of fractionalization, generally had a negative impact on several measures of public goods, such as literacy rates, infant mortality, school attainment and infrastructure. Alesina et al. (2003) broadly confirmed these results using new data on ethnic, linguistic and religious fractionalization and polarization, although the results were somewhat sensitive to the chosen measure of diversity and specification. In a within country context, Alesina, Baqir and Easterly (1999) showed that across cities, metropolitan areas and urban counties of the United States, greater

ethnic diversity was associated with lower provision of education, roads and sewers.

In a more microeconomic context, Habyarimana et al. (2007) report that in a variety of games, co-ethnic participants from a sample of slum dwellers in Kampala, Uganda, play cooperative strategies more so than players from different ethnic groups. This is consistent with findings in Miguel and Gugerty (2005), suggesting that public goods provision is lower in more ethnically diverse locations in Kenya. Other studies include Vigdor (2004) who shows that higher racial, generational and socioeconomic heterogeneity across US counties is associated with lower response rates to the 2000 Census questionnaire, and Banerjee et al. (2005) who, in the context of rural India, find that higher caste and religious fragmentation is associated with lower provision of a wide range of public goods. Although these results are compelling, it is not clear what ethnolinguistic cleavages are most relevant as determinants of public goods provision.

To analyze empirically the effects of diversity computed at different levels of aggregation on the provision of public goods, we start with the econometric specification and data in LLSV (1999). To minimize the potential for omitted variables bias, we focus on the specification that include the largest set of control variables – including legal origins, GNP per capita, latitude, and religion shares variables (this corresponds to the specification of their Table 6, pp. 256-260). Instead of focusing on a broad set of measures of the quality of government as they did, we focus on the category of dependent variables they label "output of public goods". This includes log infant mortality, log of school attainment, the illiteracy rate, and an index of infrastructure quality.

The results are presented in the top panels of Table 7 (for fractionalization) and Table 8 (for polarization). For three of the four dependent variables, the statistical significance of the coefficient on ELF rises as the level of linguistic aggregation falls. The effects are of the expected signs, namely linguistic fractionalization is negatively associated with school attainment, but positively associated with log infant mortality and the illiteracy rate. There is no significant association with the index of infrastructure quality at any level of aggregation (this was also the case in LLSV). The LLSV measure of ethnolinguistic fractionalization is most highly correlated with $ELF(15)$ - the correlation between the two measures is 0.835, and falls steadily as the level of aggregation rises. Correspondingly, in quantitative terms the magnitude of our estimates is very close to LLSV's when ELF is measured at aggregation level 15. Finally, comparing Tables 7 and Table 8, we see that linguistic fractionalization is a much better predictor of public goods than linguistic polarization, as no clear patten emerges when using the latter set of measures.

In order to investigate whether these results hold up to using a broader set of indicators of public goods provision, the bottom panels of Tables 7 and 8 consider 6 additional dependent variables, taken from the World Development Indicators (World Bank, 2008). These includes measures of health care (hospital beds per person, measles immunization rates for children), measures of access to public services (availability of sanitation services and clean water), and specific measures of infrastructure (road and rail network density).[11] The results broadly confirm the findings obtained from the LLSV measures: 1) for measures of sanitation and clean water, the effect of fractionalization rises in magnitude and statistical significance as the level of aggregation falls; 2) for measures of health services, the effect of ELF remains consistently significant for the measles immunization rate across aggregation levels, but is insignificant for hospital beds; 3) infrastructure measures are unaffected by fractionalization whatever the level of aggregation; and 4) fractionalization is a better predictor of public goods provision than polarization.

To summarize, across a wide range of measures of public goods, we broadly confirm results from the literature referenced above: ethnolinguistic diversity is bad for public goods provision. More importantly for our purposes, we also find that measures of fractionalization based on finer classifications of linguistic groups tend to matter more than those based on deep cleavages only. In contrast with redistribution, for which only deep splits were important, even relatively recent and shallow linguistic cleavages are sufficient to hinder the provision of public goods.

4.2 Growth

In recent years, scholars have focused on ethnolinguistic diversity as a determinant of economic performance. Easterly and Levine (1997) argue that ethnic diversity, measured by an index of fractionalization, may account for much of Africa's growth tragedy. These cross-country results were reinforced and extended in Alesina et al. (2003). In particular, the latter paper showed that linguistic diversity per se, not just ethnic diversity, has a significantly negative effect on per capita income growth in a panel of countries, so that both ethnic and linguistic diversity are alternative ways to capture a broader concept of cultural heterogeneity. In addition, the paper found that fractionalization measures were more robust predictors of growth than polarization measures, an

[11] We measure the latter as a ratio of kilometers per 1,000 inhabitatnts, but the results ars unchanged when using kilometers per square kilometer of land area instead. Results are available upon request.

issue we revisit below.[12]

To examine the impact of linguistic cleavages at various levels of aggregation, we start with a specification in Easterly and Levine (1997). We focus on the specification in their Table IV, column 1, which contains a number of control variables while at the same time allowing for a relatively large sample of countries over the period 1960-1990. Our results are presented in Tables 9 and 10, for fractionalization and polarization, respectively. The estimator is random effects applied to a panel of decade averages.[13] Table 9 reveals that the effect of fractionalization increases in absolute value when the level of aggregation declines - this is true in terms of the estimated coefficient, and also true in terms of the standardized beta displayed in Figure 9. ELF starts to become statistically significant at the 95% confidence level at aggregation level 9, and remains significant thereafter. In terms of magnitudes, Figure 9 indicates that the magnitudes settle, after level 9, at about 18% - namely a typical deviation in fractionalization can account for 18% of a typical variation in growth. No clear pattern emerges when it comes to polarization, which comes out negative but not statistically significant at acceptable levels of confidence (Table 10): consistent with findings in Alesina et al. (2003), linguistic polarization appears largely unrelated to economic growth, and measures of fractionalization are more robust predictors of growth. Finally, these findings are robust to considering an expanded set of control variables, as was done in Easterly and Levine (1997). While doing so results in a smaller number of observations due to data availability issues, the basic pattern outlined above holds, or is even reinforced, when: 1) controlling for political assassinations, 2) controlling for political assassinations plus financial depth, the black market premium, and the

[12] For a survey of the empirical literature on ethnolinguistic diversity and economic performance at the level of countries, cities and villages in developing countries, see Alesina and La Ferrara (2005). This is related to a more microeconomic approach highlighting the costs and benefits of cultural and linguistic diversity within teams or organizations. See for instance Lazear (1999), Prat (2002) and Cremer, Garicano and Prat (2007). While at the cross-country level the empirical results point to a negative relationship between ethnolinguistic diversity and growth, the findings are more contrasted at the within-firm level, with some studies pointing to positive effects of diversity. At the cross-city level in the U.S., Ottaviano and Peri (2006) also point to a positive effect of cultural diversity on the productivity of U.S. natives.

[13] Hauk and Wacziarg (2009), using simulations based on the Solow model, show that random effects and SUR produce very similar estimates. We use random effects because it is computationally much easier to produce estimates for an unbalanced panel, and we wish to exploit as much of the available information as possible. Our results are not materially different when using SUR on a balanced panel, despite the resulting fall in the number of available observations. These SUR results are available upon request.

fiscal surplus to GDP ratio, 3) controlling for all of the above plus the number of coups d'état, the number of revolutions, a dummy for civil wars and a measure of political rights (Gastil's index of democracy).

In a second test, we extend the time span of our data. We also include a number of additional control variables that have become commonly used in more recent empirical growth research: we start from an augmented Solow model which include the investment rate, a measure of human capital (the number of years of schooling in the adult population aged 25 and over - results do not change when using a flow measure such as the secondary school enrollment rate), and a measure of population growth. In addition, it includes measures of market size used in Ades and Glaeser (1999) and Alesina, Spolaore and Wacziarg (2000), namely the ratio of imports plus exports to GDP, the log of population, and the interaction between these two variables. Finally, the regression includes period fixed effects, and dummy variables for Sub-Saharan Africa as well as Latin America and the Caribbean. The timespan extends from 1960 to 2004, and the regressions include an unbalanced panel of 101 countries. With the wider set of control variables and the longer time span, we get results very similar to those obtained using the Easterly and Levine (1997) specification. Coefficient estimates are shown in Tables 11 and 12. The standardized betas displayed in Figure 10 are very similar to those in Figure 9. Again, the effect of ELF becomes greater in magnitude and more significant when the level of aggregation falls, and the standardized beta settles around 18% at levels 9 and higher. The level of significance is overall greater than before, with ELF becoming significant at the 95% confidence level at aggregation level 6. As before, polarization measures appear largely unrelated to growth.

To illustrate the quantitative importance of ethnolinguistic diversity for economic growth, we analyze the case of the world's two most populous countries, China and India. Both have experienced high growth rate in recent decades, although India continues to lag behind its East Asian neighbor. According to the Penn World Tables (version 6.2), over the period 1960-2004 China averaged an annual growth rate in real GDP per capita of 5.63%, compared to 2.75% in India. China is also much less linguistically diverse than India: at aggregation level 15, India's ELF index is 0.93, while China's is 0.49. Fitting the regression model in Column 5 of Table 11 to these two datapoints, we calculate that 25% of the growth difference between India and China over 1960-2004 is accounted for by differences between these two countries in $ELF(15)$. About 29% of the difference is accounted for by differences in $ELF(9)$, where the magnitude of the effect of linguistic

fractionalization is maximized. Thus, taken at face value the estimates in our model can account for a large portion of the difference in growth performance in India and China.

To summarize, these results show that to capture the relevant cleavages that affect economic growth, focusing only on deep cleavages is not sufficient. Instead, one needs to take into account finer distinctions across linguistic groups. This does not imply that deep cleavages do not contribute to negatively affecting growth, as these deep cleavages do contribute to diversity at lower levels of aggregation: fractionalization measured at low levels of aggregation is affected by both deep and shallow cleavages. The point is that fractionalization measured at high levels of aggregation ignores many of the shallower, yet relevant, cleavages, and therefore amounts to a noisy measure to predict the effect of diversity on growth.

5 Conclusion

In this paper, we have uncovered new evidence on the relationship between ethnolinguistic diversity and a range of political economy outcomes, such as the onset of civil wars, the extent of redistribution, the provision of public goods, and economic growth. We sought to identify the relevant linguistic cleavages to explain variation in these outcomes. We let the data tell us whether deep cleavages, originating at an earlier time in history, are more or less important than more superficial cleavages that have arisen more recently. Doing so, we departed from the common approach relying on arbitrary definitions of what constitutes a relevant ethnolinguistic group. Our results carry several lessons. When it comes to civil conflict and redistribution, deeper cleavages tend to matter more. In contrast, for economic growth and public goods, we found that diversity measured using only deep cleavages is not sufficient to predict significant differences in growth. Instead, measures based on more disaggregated classifications of linguistic groups, capturing finer distinctions between languages, are important predictors of growth and public goods provision both in terms of statistical significance and in terms of economic magnitude.

How should we interpret these results? We have shown that the type of cultural diversity that matters for outcomes involving conflicts of interest - civil wars, redistribution - is different from the type of diversity that matters for outcomes that entail issues of efficiency and coordination, such as growth. When it comes to conflict and redistribution, preferences are of the essence. The willingness to settle disputes or to transfer resources across a cultural divide depends on how deep the divide happens to be. Deep cleavages that go back thousands of year appear to be related with

more conflicts of interest, compared to more superficial cleavages.

In contrast, economic growth requires that groups be able to coordinate and interact, and organize in networks of production, knowledge and trade that are affected by ethnolinguistic divisions. In India, for instance, the degree of integration between regions is likely hindered by linguistic barriers - even linguistic barriers separate relatively similar languistic groups such as Hindi and Gujarati speakers. Coordination, integration and more generally the ability to form knowledge, production and trading networks is hampered as soon as linguistic differentiation prevents interactions between groups, and this can occur between groups that are relatively similar linguistically.

The case of public goods shares characteristics of both types of outcomes: public goods are inherently redistributive in nature, and their provision depends on differences in preferences among participants. At the same time, the provision of public goods requires coordination and interactions, that even superficial cleavages might hamper. We found that, much as in the case of growth, for a wide array of measures of public goods, fine distinctions between linguistic groups matter to hinder their provision. Even when cleavages are shallow, a country may fail to have well-functioning public services, not necessarily because people are unwilling to redistribute, but because of coordination failures.

Future work should seek to better understand the theoretical mechanisms that account for the contrasting findings between conflict and redistribution on the one hand, and growth and public goods on the other hand. In particular, clarifying the differing effects of diversity on efficiency and coordination (where fine distinctions seem to matter more) and preferences (where coarse distinctions seem of the essence) may help account for our results.

Finally, we have focused on linguistic diversity, as a measure of a broader concept of ethnolinguistic heterogeneity, and even more broadly as a proxy for cultural diversity. One advantage of focusing on languages is that linguistic distinctions are quite objective: it is easier to judge whether two populations speak different languages than to decide whether two populations belong to different ethnicities, a more amorphous concept (precisely for this reason, ethnic categorizations are often based on linguistic divisions, particularly for Africa). Another advantage is that data on linguistic divisions, particularly in the form of trees, is more readily available than data on the genealogical structure of ethnic groups within countries. There are, however, drawbacks to focusing on languages: to the extent that linguistic divisions are imperfect measures of the source of diversity that matters most, this should lead to downward bias on the estimates of the effect of diversity on

political economy outcomes. In principle, the methodology we have developed for linguistic trees should be applicable to other kinds of differences between populations. With advances in population genetics, population phylogenies have become more widely available. Although this data is not yet available in a single format such as the *Ethnologue* for languages, applying our method to genetic data could lead to fruitful advances in the study of the political economy of cultural diversity.

References

Ades A. and E. L. Glaeser (1999), "Evidence on Growth, Increasing Returns, and the Extent of the Market", *Quarterly Journal of Economics*, vol. 114, no. 3, August, pp. 1025-1045.

Alesina, A. and E. L. Glaeser (2004), *Fighting Poverty in the US and Europe: A World of Difference*, Oxford: Oxford University Press.

Alesina, A. and E. La Ferrara (2000), "Participation in Heterogeneous Communities", *Quarterly Journal of Economics*, vol. 115, no. 3, August, pp. 847-904.

Alesina, A. and E. La Ferrara (2005), "Ethnic Diversity and Economic Performance", *Journal of Economic Literature*, vol. 43, September, pp. 762-800.

Alesina, A., R. Baqir and W. Easterly (1999), "Public Goods and Ethnic Divisions", *Quarterly Journal of Economics*, vol. 114, no. 4, November, pp. 1243-1284.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat and R. Wacziarg (2003), "Fractionalization", *Journal of Economic Growth*, vol. 8, no. 2, June, pp. 155-194.

Alesina, A., E. Glaeser and B. Sacerdote (2001), "Why Doesn't the United States Have a European-Style Welfare State?", *Brookings Papers on Economic Activity*, vol. 2001, no. 2, pp. 187-254.

Alesina, A, E. Spolaore and R. Wacziarg (2000), "Economic Integration and Political Disintegration", *American Economic Review*, vol. 90, no. 5. December, pp. 1276-1296.

Banerjee, A., Iyer, L. and Somanathan, R. (2005), "History, Social Divisions, and Public Goods in Rural India", *Journal of the European Economic Association*, vol. 3, no. 2-3, pp. 639-647.

Barro, R. J. and J. W. Lee (2000), "International Data on Educational Attainment: Updates and Implications", Harvard CID Working Paper No. 42, April.

Bergslund, K. and H. Vogt (1962), "On the Validity of Glottochronology", *Current Anthropology*, vol. 3, pp. 115-153.

Bossert, W., C. D'Ambrosio and E. La Ferrara (2009), "A Generalized Index of Fractionalization", *Economica*, forthcoming.

Central Intelligence Agency (2009), *The World Factbook 2009*, Washington, DC: Central Intelligence Agency, 2009, https://www.cia.gov/library/publications/the-world-factbook/index.html

Cremer, J., L. Garicano and A. Prat (2007), "Language and the Theory of the Firm", *Quarterly Journal of Economics*, vol. 122, no. 1, February, pp. 373-407.

Darwin, C. (1902), *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*, 6^{th} edition, London: Grant Richards.

Desmet, K., I. Ortuño Ortín and S. Weber (2009), "Linguistic Diversity and Redistribution", *Journal of the European Economic Association*, vol. 7, no. 6, December (forthcoming).

Dixon, R. M. W. (1997), *The Rise and Fall of Languages*, Cambridge: Cambridge University Press.

Easterly, W. and R. Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, vol. 112, no. 4, November, pp. 1203-1250.

Emeneau, M. and D. Anwar (1980), *Language and Linguistic Area: Essays by Murray B. Emeneau*, Palo Alto: Stanford University Press.

Esteban, J. M., and D. Ray (1994), "On the Measurement of Polarization, *Econometrica*, vol. 62, no. 4, pp. 819-851.

Ethnologue (2005), *Ethnologue: Languages of the World*, 15^{th} Edition, SIL International, www.ethnologue.com.

Fearon, J. and D. Laitin (2003), "Ethnicity, Insurgency, and Civil War", *American Political Science Review*, vol. 97, no. 1, February, pp. 75-90.

Fearon, J. (2003), "Ethnic and Cultural Diversity by Country", *Journal of Economic Growth*, vol. 8, no. 2, June, pp. 195-222.

Gray, R. D. and Q. D. Atkinson (2003), "Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin", *Nature*, vol. 426, 27 November, pp. 435-439.

Greenberg, J. H. (1987), *Language in the Americas*, Palo Alto: Stanford University Press.

Greenberg, J. H. (1956), "The Measurement of Linguistic Diversity", *Language*, vol. 32, no. 1, Jamuary, pp. 109-115.

Habyarimana, J., M. Humphreys, D. Posner and J. Weinstein (2007), "Why Does Ethnic Diversity Undermine Public Goods Provision?", *American Political Science Review*, vol. 101, no. 4,

November, pp. 709-725.

Habyarimana, J, M. Humphreys, D. Posner and J. Weinstein (2009), *Coethnicity: Diversity and the Dilemmas of Collective Action*, New York: Russell Sage Foundation Press.

Hauk, W. and R. Wacziarg (2009), "A Monte Carlo Study of Growth Regressions", *Journal of Economic Growth*, vol. 14, no. 2, June, pp. 103-147.

Heston, A., R. Summers and B. Aten (2006), "Penn World Table Version 6.2", *Center for International Comparisons of Production, Income and Prices*, University of Pennsylvania, September.

La Porta, R., F. Lopez-de-Silanes, A. Shleifer and R. Vishny (1999), "The Quality of Government", *Journal of Law, Economics, and Organization*, vol. 15, no. 1, pp. 222-279.

Lazear, E. (1999), "Culture and Language", *Journal of Political Economy*, vol. 107, no. 6, December, pp. S95-S126.

Lithgow, D. (1973), "Language Change on Woodlark Island", *Oceania*, vol. 44, no. 2, December, pp.101-108.

Luttmer, E. F. P. (2001), "Group Loyalty and the Taste for Redistribution", *Journal of Political Economy*, vol. 109, no. 3, June, pp. 500-528.

Luttmer E. F. P. and C. Fong (2009), "What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty", *American Economic Journal: Applied Economics*, vol. 1, no. 2, April, pp. 64-87.

Miguel, E., and M. K. Gugerty (2005), "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya", *Journal of Public Economics*, vol. 89, no. 11, pp. 2325-2368.

Montalvo, J. G. and M. Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict and Civil War", *American Economic Review*, vol. 95, no. 3, June, pp. 796-816.

Nettle, D. (1998), "Explaining Global Patterns of Language Diversity", *Journal of Anthropological Archaeology*, vol. 17, pp. 354-374.

Ottaviano, G. M. and G. Peri (2006), "The Economic Value of Cultural Diversity: Evidence from U.S. Cities", *Journal of Economic Geography*, vol. 6, no. 1, January, pp. 9-44.

Posner, D. (2003), "The Colonial Origins of Ethnic Cleavages: The Case of Linguistic Divisions in Zambia", *Comparative Politics*, vol. 35, no. 2, pp. 127-146.

Posner, D. (2005), *Institutions and Ethnic Politics in Africa*, Cambridge (UK): Cambridge University Press.

Prat, A. (2002), "Should a Team Be Homogeneous?", *European Economic Review*, vol. 46, no. 7, pp. 1187-1207.

Scarritt, J. R. and S. Mozaffar (1999), "The Specification of Ethnic Cleavages and Ethnopolitical Groups for the Analysis of Democratic Competition in Contemporary Africa", *Nationalism and Ethnic Politics*, vol. 5, no. 1, March, pp. 82-117.

Spolaore E. and R. Wacziarg (2009), "The Diffusion of Development", *Quarterly Journal of Economics*, vol. 124, no. 2, May, pp. 469-529.

Vigdor, J. L. (2004), "Community Composition and Collective Action: Analyzing Initial Mail Response to the 2000 Census", *Review of Economics and Statistics*, vol. 86, no. 1, pp. 303-312.

World Bank (2008), *World Development Indicators 2008*, Washington, DC: World Bank.

Young, C. (1975), *The Politics of Cultural Pluralism*, Madison: University of Wisconsin Press.

Appendix 1 – Data on linguistic fractionalization (ELF) and polarization (POL), at selected levels of aggregation

country	ELF (1)	ELF (3)	ELF (6)	ELF (10)	ELF (15)	POL (1)	POL (3)	POL (6)	POL (10)	POL (15)
Afghanistan	0.298	0.348	0.543	0.732	0.732	0.567	0.574	0.685	0.638	0.638
Albania	0.000	0.257	0.257	0.257	0.257	0.000	0.453	0.453	0.453	0.453
Algeria	0.009	0.283	0.313	0.313	0.313	0.018	0.557	0.533	0.533	0.533
American Samoa	0.088	0.088	0.088	0.088	0.116	0.170	0.170	0.170	0.170	0.218
Andorra	0.000	0.025	0.025	0.574	0.574	0.000	0.050	0.050	0.910	0.910
Angola	0.015	0.015	0.015	0.726	0.785	0.030	0.030	0.030	0.723	0.588
Anguilla	0.140	0.140	0.140	0.140	0.140	0.281	0.281	0.281	0.281	0.281
Antigua and Barbuda	0.057	0.057	0.057	0.057	0.057	0.112	0.112	0.112	0.112	0.112
Argentina	0.120	0.142	0.213	0.213	0.213	0.225	0.261	0.371	0.371	0.371
Armenia	0.085	0.174	0.174	0.174	0.174	0.169	0.318	0.317	0.317	0.317
Aruba	0.355	0.381	0.387	0.387	0.387	0.710	0.641	0.623	0.623	0.623
Australia	0.028	0.108	0.126	0.126	0.126	0.055	0.203	0.233	0.233	0.233
Austria	0.012	0.036	0.521	0.540	0.540	0.024	0.071	0.978	0.957	0.957
Azerbaijan	0.346	0.367	0.373	0.373	0.373	0.634	0.594	0.588	0.588	0.588
Bahamas	0.282	0.386	0.386	0.386	0.386	0.564	0.663	0.663	0.663	0.663
Bahrain	0.455	0.467	0.663	0.663	0.663	0.689	0.649	0.698	0.698	0.698
Bangladesh	0.016	0.016	0.332	0.332	0.332	0.033	0.033	0.560	0.560	0.560
Barbados	0.091	0.091	0.091	0.091	0.091	0.182	0.182	0.182	0.182	0.182
Belarus	0.003	0.139	0.397	0.397	0.397	0.005	0.275	0.624	0.624	0.624
Belgium	0.038	0.518	0.663	0.675	0.734	0.076	0.973	0.817	0.797	0.708
Belize	0.614	0.693	0.693	0.693	0.693	0.838	0.752	0.752	0.752	0.752
Benin	0.014	0.126	0.614	0.901	0.901	0.029	0.241	0.729	0.314	0.314
Bermuda	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bhutan	0.418	0.418	0.720	0.846	0.846	0.837	0.837	0.742	0.492	0.492
Bolivia	0.647	0.665	0.680	0.680	0.680	0.879	0.854	0.827	0.827	0.827
Bosnia and Herzegovina	0.019	0.158	0.416	0.416	0.416	0.037	0.307	0.632	0.632	0.632
Botswana	0.067	0.067	0.068	0.364	0.444	0.132	0.131	0.129	0.582	0.617
Brazil	0.008	0.026	0.032	0.032	0.032	0.016	0.052	0.063	0.063	0.063
British Ind. Ocean Terr.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
British Virgin Islands	0.167	0.167	0.167	0.167	0.167	0.334	0.334	0.334	0.334	0.334
Brunei	0.246	0.387	0.455	0.456	0.456	0.464	0.580	0.609	0.608	0.608
Bulgaria	0.171	0.222	0.224	0.224	0.224	0.343	0.413	0.414	0.414	0.414
Burkina Faso	0.028	0.453	0.532	0.723	0.773	0.056	0.681	0.667	0.590	0.531
Burundi	0.001	0.001	0.001	0.004	0.004	0.002	0.002	0.002	0.007	0.008
Cambodia	0.088	0.143	0.157	0.157	0.157	0.170	0.265	0.286	0.286	0.286
Cameroon	0.211	0.328	0.443	0.878	0.942	0.414	0.545	0.598	0.362	0.195
Canada	0.071	0.489	0.541	0.549	0.549	0.136	0.820	0.791	0.771	0.771
Cape Verde Islands	0.070	0.070	0.070	0.070	0.070	0.140	0.140	0.140	0.140	0.140
Cayman Islands	0.499	0.547	0.547	0.547	0.547	0.999	0.956	0.956	0.956	0.956
Central African Republic	0.353	0.424	0.557	0.958	0.960	0.584	0.628	0.669	0.159	0.151
Chad	0.550	0.818	0.895	0.949	0.950	0.888	0.542	0.340	0.184	0.181
Chile	0.029	0.034	0.034	0.034	0.034	0.058	0.067	0.067	0.067	0.067
China	0.066	0.491	0.491	0.491	0.491	0.127	0.623	0.622	0.622	0.622
Colombia	0.025	0.030	0.030	0.030	0.030	0.050	0.059	0.059	0.059	0.059
Comoros	0.011	0.011	0.011	0.011	0.551	0.022	0.022	0.022	0.022	0.946
Congo	0.498	0.512	0.554	0.775	0.820	0.963	0.941	0.918	0.607	0.466
Cook Islands	0.061	0.061	0.061	0.061	0.379	0.122	0.122	0.122	0.122	0.599

country	ELF (1)	ELF (3)	ELF (6)	ELF (10)	ELF (15)	POL (1)	POL (3)	POL (6)	POL (10)	POL (15)
Costa Rica	0.049	0.050	0.050	0.050	0.050	0.097	0.097	0.097	0.097	0.097
Cote dIvoire	0.004	0.382	0.820	0.862	0.917	0.008	0.647	0.503	0.399	0.277
Croatia	0.000	0.078	0.087	0.087	0.087	0.000	0.152	0.167	0.167	0.167
Cuba	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.001	0.001
Cyprus	0.360	0.366	0.366	0.366	0.366	0.717	0.720	0.720	0.720	0.720
Czech Republic	0.000	0.049	0.069	0.069	0.069	0.000	0.097	0.134	0.134	0.134
Dem Rep Congo	0.351	0.354	0.438	0.916	0.947	0.600	0.589	0.641	0.280	0.185
Denmark	0.014	0.034	0.051	0.051	0.051	0.029	0.066	0.099	0.099	0.099
Djibouti	0.050	0.191	0.592	0.592	0.592	0.101	0.360	0.911	0.911	0.911
Dominica	0.307	0.313	0.313	0.313	0.313	0.614	0.619	0.619	0.619	0.619
Dominican Republic	0.051	0.053	0.053	0.053	0.053	0.102	0.105	0.105	0.105	0.105
East Timor	0.524	0.524	0.575	0.897	0.897	0.848	0.848	0.721	0.366	0.366
Ecuador	0.240	0.255	0.264	0.264	0.264	0.473	0.490	0.461	0.461	0.461
Egypt	0.022	0.025	0.509	0.509	0.509	0.044	0.048	0.820	0.820	0.820
El Salvador	0.004	0.004	0.004	0.004	0.004	0.009	0.009	0.009	0.009	0.009
Equatorial Guinea	0.103	0.103	0.103	0.448	0.453	0.198	0.197	0.197	0.651	0.636
Eritrea	0.118	0.518	0.749	0.749	0.749	0.236	0.691	0.656	0.656	0.656
Estonia	0.455	0.463	0.476	0.476	0.476	0.903	0.896	0.870	0.870	0.870
Ethiopia	0.020	0.579	0.830	0.843	0.843	0.040	0.919	0.494	0.465	0.465
Falkland Islands	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Fiji	0.499	0.505	0.505	0.589	0.607	0.998	0.991	0.991	0.911	0.883
Finland	0.119	0.140	0.140	0.140	0.140	0.238	0.268	0.268	0.268	0.268
France	0.082	0.154	0.195	0.272	0.272	0.161	0.279	0.337	0.438	0.438
French Guiana	0.279	0.480	0.480	0.480	0.480	0.467	0.648	0.648	0.648	0.648
French Polynesia	0.384	0.384	0.384	0.384	0.596	0.645	0.645	0.645	0.645	0.706
Gabon	0.115	0.115	0.125	0.819	0.919	0.231	0.231	0.247	0.529	0.281
Gambia	0.018	0.500	0.735	0.739	0.748	0.036	0.998	0.690	0.683	0.669
Georgia	0.435	0.574	0.576	0.576	0.576	0.715	0.667	0.662	0.662	0.662
Germany	0.055	0.121	0.189	0.189	0.189	0.110	0.225	0.330	0.330	0.330
Ghana	0.000	0.019	0.599	0.796	0.805	0.000	0.037	0.782	0.523	0.492
Gibraltar	0.498	0.498	0.498	0.498	0.498	0.996	0.996	0.996	0.996	0.996
Greece	0.024	0.142	0.175	0.175	0.175	0.047	0.259	0.310	0.310	0.310
Greenland	0.242	0.242	0.242	0.242	0.242	0.484	0.484	0.484	0.484	0.484
Grenada	0.016	0.064	0.064	0.064	0.064	0.032	0.126	0.126	0.126	0.126
Guadeloupe	0.033	0.084	0.084	0.084	0.084	0.066	0.163	0.163	0.163	0.163
Guam	0.438	0.640	0.640	0.640	0.640	0.761	0.814	0.814	0.814	0.814
Guatemala	0.502	0.583	0.691	0.691	0.691	0.997	0.898	0.665	0.665	0.665
Guinea	0.000	0.505	0.720	0.735	0.748	0.000	0.978	0.743	0.708	0.679
Guinea-Bissau	0.229	0.401	0.820	0.853	0.853	0.458	0.666	0.567	0.478	0.478
Guyana	0.077	0.078	0.078	0.078	0.078	0.150	0.151	0.149	0.149	0.149
Haiti	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Honduras	0.045	0.056	0.056	0.056	0.056	0.088	0.109	0.109	0.109	0.109
Hungary	0.153	0.157	0.158	0.158	0.158	0.305	0.289	0.286	0.286	0.286
Iceland	0.000	0.000	0.019	0.019	0.019	0.000	0.000	0.038	0.038	0.038
India	0.386	0.412	0.910	0.930	0.930	0.723	0.654	0.298	0.244	0.244
Indonesia	0.065	0.783	0.846	0.846	0.846	0.126	0.614	0.438	0.438	0.438
Iran	0.493	0.517	0.772	0.797	0.797	0.931	0.921	0.640	0.572	0.572
Iraq	0.309	0.310	0.661	0.666	0.666	0.579	0.576	0.763	0.747	0.747
Ireland	0.004	0.164	0.223	0.223	0.223	0.008	0.325	0.412	0.412	0.412
Israel	0.387	0.434	0.664	0.665	0.665	0.731	0.627	0.635	0.634	0.634

country	ELF (1)	ELF (3)	ELF (6)	ELF (10)	ELF (15)	POL (1)	POL (3)	POL (6)	POL (10)	POL (15)
Italy	0.002	0.019	0.559	0.593	0.593	0.004	0.037	0.766	0.667	0.667
Jamaica	0.011	0.011	0.011	0.011	0.011	0.022	0.022	0.022	0.022	0.022
Japan	0.012	0.028	0.028	0.028	0.028	0.024	0.056	0.056	0.056	0.056
Jordan	0.027	0.027	0.484	0.484	0.484	0.053	0.053	0.698	0.698	0.698
Kazakhstan	0.501	0.622	0.701	0.701	0.701	0.968	0.857	0.749	0.749	0.749
Kenya	0.458	0.458	0.350	0.800	0.901	0.829	0.828	0.558	0.565	0.337
Kiribati	0.016	0.016	0.016	0.033	0.033	0.033	0.033	0.033	0.065	0.065
Korea, North	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Korea, South	0.003	0.003	0.003	0.003	0.003	0.006	0.006	0.006	0.006	0.006
Kuwait	0.000	0.034	0.556	0.556	0.556	0.000	0.069	0.803	0.803	0.803
Kyrgyzstan	0.461	0.633	0.670	0.670	0.670	0.887	0.828	0.771	0.771	0.771
Laos	0.454	0.491	0.513	0.678	0.678	0.743	0.674	0.629	0.595	0.595
Latvia	0.006	0.530	0.595	0.595	0.595	0.012	0.941	0.845	0.845	0.845
Lebanon	0.143	0.145	0.161	0.161	0.161	0.285	0.276	0.300	0.300	0.300
Lesotho	0.000	0.000	0.000	0.253	0.260	0.000	0.000	0.000	0.506	0.483
Liberia	0.054	0.693	0.883	0.911	0.912	0.109	0.796	0.403	0.307	0.305
Libya	0.015	0.075	0.362	0.362	0.362	0.030	0.146	0.653	0.653	0.653
Liechtenstein	0.000	0.050	0.128	0.128	0.128	0.000	0.100	0.244	0.244	0.244
Lithuania	0.003	0.331	0.339	0.339	0.339	0.006	0.576	0.559	0.559	0.559
Luxembourg	0.013	0.429	0.452	0.498	0.498	0.027	0.840	0.789	0.762	0.762
Macedonia	0.160	0.566	0.566	0.566	0.566	0.321	0.811	0.810	0.810	0.810
Madagascar	0.016	0.016	0.656	0.656	0.656	0.032	0.032	0.657	0.657	0.657
Malawi	0.053	0.053	0.053	0.519	0.519	0.106	0.106	0.106	0.667	0.667
Malaysia	0.504	0.650	0.758	0.758	0.758	0.835	0.663	0.573	0.572	0.572
Maldives	0.000	0.000	0.010	0.010	0.010	0.000	0.000	0.020	0.020	0.020
Mali	0.239	0.653	0.788	0.867	0.876	0.432	0.752	0.569	0.405	0.381
Malta	0.016	0.016	0.016	0.016	0.016	0.031	0.031	0.031	0.031	0.031
Marshall Islands	0.027	0.027	0.027	0.027	0.027	0.053	0.053	0.053	0.053	0.053
Martinique	0.043	0.043	0.043	0.043	0.043	0.085	0.085	0.085	0.085	0.085
Mauritania	0.170	0.171	0.172	0.172	0.172	0.323	0.319	0.317	0.317	0.317
Mauritius	0.557	0.584	0.641	0.641	0.641	0.930	0.888	0.781	0.781	0.781
Mayotte	0.432	0.432	0.432	0.432	0.459	0.822	0.822	0.822	0.822	0.827
Mexico	0.126	0.135	0.135	0.135	0.135	0.234	0.244	0.243	0.243	0.243
Micronesia	0.105	0.105	0.211	0.384	0.792	0.207	0.207	0.383	0.586	0.585
Moldova	0.062	0.552	0.589	0.589	0.589	0.123	0.823	0.733	0.733	0.733
Monaco	0.000	0.000	0.000	0.521	0.521	0.000	0.000	0.000	0.799	0.799
Mongolia	0.027	0.162	0.331	0.331	0.331	0.053	0.305	0.534	0.534	0.534
Montserrat	0.026	0.026	0.026	0.026	0.026	0.051	0.051	0.051	0.051	0.051
Morocco	0.008	0.412	0.466	0.466	0.466	0.015	0.814	0.688	0.688	0.688
Mozambique	0.004	0.004	0.004	0.730	0.929	0.007	0.007	0.007	0.671	0.255
Myanmar	0.230	0.427	0.520	0.521	0.521	0.418	0.619	0.626	0.621	0.621
Namibia	0.478	0.479	0.491	0.784	0.808	0.758	0.754	0.719	0.583	0.536
Nauru	0.241	0.241	0.241	0.596	0.596	0.435	0.435	0.435	0.729	0.729
Nepal	0.300	0.300	0.738	0.742	0.742	0.595	0.595	0.591	0.577	0.577
Netherlands	0.118	0.123	0.389	0.389	0.389	0.219	0.227	0.567	0.567	0.567
Netherlands Antilles	0.072	0.260	0.266	0.266	0.266	0.145	0.475	0.458	0.458	0.458
New Caledonia	0.528	0.578	0.578	0.806	0.834	0.907	0.882	0.882	0.547	0.462
New Zealand	0.094	0.100	0.100	0.101	0.102	0.185	0.196	0.196	0.195	0.191
Nicaragua	0.081	0.081	0.081	0.081	0.081	0.160	0.159	0.159	0.159	0.159
Niger	0.525	0.638	0.642	0.646	0.646	0.842	0.761	0.745	0.730	0.730

country	ELF (1)	ELF (3)	ELF (6)	ELF (10)	ELF (15)	POL (1)	POL (3)	POL (6)	POL (10)	POL (15)
Nigeria	0.434	0.496	0.855	0.870	0.870	0.788	0.772	0.462	0.415	0.415
Niue	0.071	0.071	0.071	0.071	0.071	0.143	0.143	0.143	0.143	0.143
Norfolk Island	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Northern Mariana Islands	0.313	0.605	0.605	0.605	0.642	0.625	0.870	0.870	0.870	0.784
Norway	0.254	0.255	0.257	0.257	0.257	0.474	0.471	0.466	0.466	0.466
Oman	0.241	0.326	0.693	0.693	0.693	0.464	0.543	0.734	0.734	0.734
Pakistan	0.034	0.333	0.750	0.762	0.762	0.068	0.634	0.630	0.592	0.592
Palau	0.000	0.077	0.077	0.077	0.077	0.000	0.155	0.155	0.155	0.154
Palestine	0.002	0.002	0.208	0.208	0.208	0.004	0.004	0.408	0.408	0.408
Panama	0.321	0.324	0.324	0.324	0.324	0.553	0.543	0.543	0.543	0.543
Papua New Guinea	0.564	0.699	0.941	0.989	0.990	0.748	0.662	0.203	0.042	0.038
Paraguay	0.319	0.334	0.340	0.347	0.347	0.620	0.587	0.585	0.563	0.563
Peru	0.347	0.366	0.376	0.376	0.376	0.644	0.585	0.552	0.552	0.552
Philippines	0.027	0.396	0.836	0.849	0.849	0.053	0.671	0.510	0.466	0.466
Pitcairn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Poland	0.000	0.032	0.060	0.060	0.060	0.000	0.063	0.116	0.116	0.116
Portugal	0.011	0.011	0.011	0.022	0.022	0.022	0.022	0.022	0.043	0.043
Puerto Rico	0.000	0.047	0.047	0.049	0.049	0.001	0.093	0.095	0.097	0.097
Qatar	0.538	0.607	0.608	0.608	0.608	0.960	0.872	0.869	0.869	0.869
Reunion	0.066	0.066	0.066	0.066	0.066	0.128	0.128	0.128	0.128	0.128
Romania	0.129	0.166	0.168	0.168	0.168	0.256	0.313	0.317	0.317	0.317
Russia	0.172	0.214	0.283	0.283	0.283	0.316	0.363	0.448	0.448	0.448
Rwanda	0.001	0.001	0.001	0.004	0.004	0.002	0.002	0.002	0.008	0.008
St Helena	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
St Kitts and Nevis	0.010	0.010	0.010	0.010	0.010	0.020	0.020	0.020	0.020	0.020
St Lucia	0.020	0.020	0.020	0.020	0.020	0.040	0.040	0.040	0.040	0.040
St Pierre & Miquelon	0.070	0.134	0.134	0.134	0.134	0.140	0.253	0.253	0.253	0.253
St Vincent & Grenadines	0.009	0.009	0.009	0.009	0.009	0.017	0.017	0.017	0.017	0.017
Samoa	0.002	0.002	0.002	0.002	0.002	0.004	0.004	0.004	0.004	0.004
San Marino	0.000	0.000	0.494	0.494	0.494	0.000	0.000	0.988	0.988	0.988
Sao Tome e Principe	0.302	0.389	0.389	0.389	0.389	0.562	0.642	0.642	0.642	0.642
Saudi Arabia	0.167	0.172	0.609	0.609	0.609	0.309	0.313	0.860	0.860	0.860
Senegal	0.017	0.248	0.757	0.767	0.772	0.034	0.484	0.670	0.643	0.625
Serbia and Montenegro	0.076	0.359	0.359	0.359	0.359	0.150	0.600	0.600	0.600	0.600
Seychelles	0.066	0.067	0.067	0.067	0.067	0.132	0.130	0.130	0.130	0.130
Sierra Leone	0.180	0.502	0.792	0.817	0.817	0.359	0.996	0.609	0.542	0.542
Singapore	0.491	0.747	0.748	0.748	0.748	0.708	0.629	0.628	0.628	0.628
Slovakia	0.182	0.289	0.307	0.307	0.307	0.364	0.503	0.521	0.521	0.521
Slovenia	0.010	0.023	0.174	0.174	0.174	0.019	0.045	0.336	0.336	0.336
Solomon Islands	0.269	0.273	0.524	0.877	0.965	0.480	0.464	0.733	0.376	0.131
Somalia	0.014	0.014	0.179	0.179	0.179	0.028	0.028	0.334	0.334	0.334
South Africa	0.380	0.388	0.412	0.724	0.869	0.742	0.723	0.659	0.718	0.438
Spain	0.032	0.033	0.034	0.438	0.438	0.064	0.066	0.067	0.696	0.696
Sri Lanka	0.305	0.313	0.313	0.313	0.313	0.605	0.611	0.611	0.611	0.611
Sudan	0.440	0.564	0.584	0.587	0.587	0.797	0.689	0.623	0.611	0.611
Suriname	0.571	0.744	0.788	0.788	0.788	0.877	0.734	0.631	0.631	0.631
Swaziland	0.000	0.000	0.000	0.050	0.228	0.000	0.000	0.000	0.099	0.428
Sweden	0.079	0.136	0.150	0.167	0.167	0.155	0.249	0.271	0.297	0.297
Switzerland	0.016	0.477	0.510	0.547	0.547	0.033	0.856	0.819	0.727	0.727
Syria	0.186	0.189	0.503	0.503	0.503	0.367	0.354	0.689	0.689	0.689

country	ELF (1)	ELF (3)	ELF (6)	ELF (10)	ELF (15)	POL (1)	POL (3)	POL (6)	POL (10)	POL (15)
Taiwan	0.033	0.488	0.488	0.488	0.488	0.065	0.757	0.757	0.757	0.757
Tajikistan	0.346	0.445	0.467	0.482	0.482	0.685	0.706	0.706	0.709	0.709
Tanzania	0.123	0.123	0.104	0.925	0.965	0.232	0.232	0.195	0.266	0.131
Thailand	0.232	0.233	0.234	0.753	0.753	0.408	0.405	0.404	0.674	0.674
Togo	0.002	0.027	0.616	0.897	0.897	0.003	0.054	0.871	0.344	0.342
Tokelau	0.054	0.054	0.054	0.054	0.054	0.108	0.108	0.108	0.108	0.108
Tonga	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.000	0.030
Trinidad and Tobago	0.467	0.571	0.696	0.696	0.696	0.850	0.752	0.713	0.713	0.713
Tunisia	0.005	0.011	0.012	0.012	0.012	0.009	0.021	0.023	0.023	0.023
Turkey	0.256	0.257	0.289	0.289	0.289	0.468	0.462	0.475	0.474	0.474
Turkmenistan	0.204	0.372	0.386	0.386	0.386	0.402	0.584	0.584	0.584	0.584
Turks and Caicos Islands	0.145	0.145	0.145	0.145	0.145	0.291	0.291	0.291	0.291	0.291
Tuvalu	0.000	0.000	0.000	0.139	0.139	0.000	0.000	0.000	0.279	0.279
U.S. Virgin Islands	0.316	0.339	0.339	0.339	0.339	0.632	0.587	0.587	0.587	0.587
Uganda	0.445	0.479	0.312	0.688	0.928	0.865	0.793	0.508	0.608	0.255
Ukraine	0.027	0.126	0.492	0.492	0.492	0.053	0.236	0.777	0.777	0.777
United Arab Emirates	0.611	0.662	0.775	0.777	0.777	0.810	0.738	0.579	0.573	0.573
United Kingdom	0.031	0.135	0.139	0.139	0.139	0.060	0.250	0.251	0.251	0.251
Uruguay	0.000	0.027	0.075	0.092	0.092	0.000	0.054	0.146	0.176	0.176
USA	0.058	0.273	0.351	0.353	0.353	0.111	0.479	0.552	0.543	0.543
Uzbekistan	0.246	0.419	0.428	0.428	0.428	0.479	0.624	0.601	0.600	0.600
Vanuatu	0.213	0.215	0.215	0.838	0.972	0.388	0.382	0.382	0.459	0.108
Vatican State	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Venezuela	0.024	0.026	0.026	0.026	0.026	0.048	0.051	0.051	0.051	0.051
Viet Nam	0.154	0.204	0.234	0.234	0.234	0.288	0.362	0.393	0.389	0.389
Wallis and Futuna	0.018	0.018	0.018	0.018	0.407	0.036	0.036	0.036	0.036	0.793
Yemen	0.029	0.126	0.579	0.579	0.579	0.058	0.239	0.914	0.914	0.914
Zambia	0.010	0.010	0.010	0.834	0.855	0.020	0.020	0.020	0.497	0.439
Zimbabwe	0.049	0.049	0.049	0.300	0.526	0.098	0.098	0.098	0.507	0.660

Table 1 – Growth, Conflict, Redistribution and Linguistic Diversity in Chad and Zambia

	Chad	Zambia
Per capita growth 1960-1990 (Easterly-Levine), %	-1%	-1%
Per capita growth 1965-2000* (PWT 6.2), %	-1%	-1%
Years of civil war 1965-1999*	35	0
Redistribution as % of GDP, 1985-1995	0.9%	3.8%
ELF (most disaggregated level)	0.95	0.85
ELF (at the aggregated level of language families)	0.55	0.01
Polarization (most disaggregated level)	0.18	0.43
Polarization (at the aggregated level of language families)	0.89	0.02

* We choose 1965 as the start date for the data on growth and conflict as this is the date of independence for Zambia.

Table 2 – Summary Statistics for Ethnolinguistic Diversity Measures

Panel A. Means and Standard Deviations

Variable	Mean	Std. Dev.	Min	Max
ELF(1)	0.156	0.180	0.000	0.647
ELF(3)	0.241	0.221	0.000	0.818
ELF(6)	0.328	0.272	0.000	0.941
ELF(10)	0.394	0.301	0.000	0.989
ELF(15)	0.412	0.308	0.000	0.990
POL(1)	0.283	0.314	0.000	0.999
POL(3)	0.384	0.316	0.000	0.998
POL(6)	0.423	0.297	0.000	0.996
POL(10)	0.435	0.279	0.000	0.996
POL(15)	0.432	0.278	0.000	0.996

(226 observations)

Panel B. Correlations

	ELF(1)	ELF(3)	ELF(6)	ELF(10)	ELF(15)	POL(1)	POL(3)	POL(6)	POL(10)
ELF(3)	0.770	1							
ELF(6)	0.579	0.826	1						
ELF(10)	0.544	0.708	0.848	1					
ELF(15)	0.526	0.672	0.798	0.977	1				
POL(1)	0.988	0.754	0.565	0.530	0.514	1			
POL(3)	0.720	0.939	0.788	0.683	0.651	0.737	1		
POL(6)	0.545	0.691	0.821	0.697	0.654	0.563	0.763	1	
POL(10)	0.444	0.568	0.643	0.664	0.638	0.466	0.637	0.838	1
POL(15)	0.391	0.513	0.595	0.542	0.555	0.408	0.572	0.777	0.925

(226 observations)

Table 3: Civil Conflict and Linguistic Fractionalization (1945-1999)
Dependent variable: Onset of Civil War, logit estimator

	(1)	(2)	(3)	(4)	(5)
	ELF(1)	**ELF(3)**	**ELF(6)**	**ELF(10)**	**ELF(15)**
ELF (at different levels of aggregation)	**1.060 [2.02]****	**0.777 [1.57]**	**0.085 [0.22]**	**-0.138 [0.37]**	**-0.051 [0.14]**
Lagged civil war	-0.908 [3.45]***	-0.906 [3.58]***	-0.861 [3.37]***	-0.852 [3.33]***	-0.855 [3.33]***
Log lagged GDP/cap	-0.700 [5.45]***	-0.651 [4.87]***	-0.693 [5.05]***	-0.717 [5.12]***	-0.706 [5.10]***
Log lagged population	0.292 [4.80]***	0.270 [4.51]***	0.282 [4.65]***	0.288 [4.92]***	0.286 [4.90]***
% mountainous	0.008 [1.69]*	0.008 [1.60]	0.008 [1.75]*	0.008 [1.81]*	0.008 [1.81]*
Noncontiguous state dummy	0.555 [1.90]*	0.474 [1.67]*	0.464 [1.64]	0.467 [1.66]*	0.465 [1.65]*
Oil exporter dummy	0.775 [2.81]***	0.758 [2.81]***	0.865 [3.32]***	0.903 [3.42]***	0.888 [3.37]***
New State dummy (1st or 2nd year from independence)	1.751 [5.07]***	1.741 [5.04]***	1.756 [5.08]***	1.762 [5.12]***	1.760 [5.11]***
Instability dummy (3 years prior)	0.649 [3.08]***	0.675 [3.16]***	0.649 [3.05]***	0.644 [3.03]***	0.646 [3.04]***
Democracy lagged (Polity 2)	0.015 [0.81]	0.017 [0.96]	0.019 [1.01]	0.019 [1.05]	0.019 [1.04]
Religious fractionalization	-0.122 [0.22]	-0.061 [0.11]	0.037 [0.07]	0.093 [0.17]	0.065 [0.12]
Constant	-2.307 [1.94]*	-2.493 [2.03]**	-2.175 [1.77]*	-1.986 [1.55]	-2.070 [1.62]

(t-statistics based on robust standard errors, clustered at the level of countries, in parentheses)
* significant at 10%; ** significant at 5%; *** significant at 1%
All columns involve 6,059 observations from 149 countries from 1945 to 1999.
The table reports logit coefficients, not marginal effects.
The specification is based on Fearon and Laitin (2003), Table 1, column 1, page 84. Results are robust to controlling for the growth of GDP per capita, the growth of GDP per capita lagged, a lagged dichotomous indicator of democracy (instead of the Polity2 index), a Subsaharan Africa dummy, an Asian dummy, a North Africa / Middle East dummy, and an anocracy dummy.
The data is from Fearon and Laitin (2003), except for ELF (authors' calculations from Ethnologue database).

Table 4: Civil Conflict and Linguistic Polarization (1945-1999)
Dependent variable: Onset of Civil War, logit estimator

	(1)	(2)	(3)	(4)	(5)
	POL(1)	**POL(3)**	**POL(6)**	**POL(10)**	**POL(15)**
POL (at different levels of aggregation)	**0.623 [2.06]****	**0.350 [0.95]**	**-0.317 [0.79]**	**-0.600 [1.41]**	**-0.632 [1.39]**
Lagged civil war	-0.912 [3.45]***	-0.872 [3.40]***	-0.873 [3.31]***	-0.869 [3.38]***	-0.874 [3.39]***
Log lagged GDP per capita	-0.702 [5.46]***	-0.681 [5.24]***	-0.702 [5.37]***	-0.701 [5.37]***	-0.699 [5.35]***
Log lagged populations	0.290 [4.76]***	0.275 [4.65]***	0.293 [4.85]***	0.295 [4.85]***	0.296 [4.80]***
% Mountainous	0.008 [1.65]*	0.008 [1.53]	0.008 [1.83]*	0.009 [1.95]*	0.009 [1.99]**
Noncontiguous state	0.568 [1.93]*	0.503 [1.75]*	0.422 [1.48]	0.417 [1.46]	0.422 [1.50]
Oil exporter dummy	0.765 [2.75]***	0.828 [3.17]***	0.904 [3.52]***	0.940 [3.60]***	0.940 [3.63]***
New State dummy	1.749 [5.06]***	1.747 [5.08]***	1.779 [5.12]***	1.808 [5.16]***	1.807 [5.12]***
Instability dummy	0.651 [3.09]***	0.657 [3.09]***	0.649 [3.06]***	0.654 [3.09]***	0.655 [3.09]***
Democracy lagged (Polity 2)	0.015 [0.81]	0.018 [0.99]	0.019 [1.04]	0.018 [1.00]	0.018 [1.00]
Religious fractionalization	-0.131 [0.24]	-0.024 [0.04]	0.048 [0.09]	0.022 [0.04]	-0.038 [0.07]
Constant	-2.282 [1.92]*	-2.282 [1.91]*	-2.039 [1.71]*	-1.946 [1.63]	-1.953 [1.64]

(t-statistics based on robust standard errors, clustered at the level of countries, in parentheses)
* significant at 10%; ** significant at 5%; *** significant at 1%
All columns involve 6,059 observations from 149 countries from 1945 to 1999.
The table reports logit coefficients, not marginal effects.
The specification is based on Fearon and Laitin (2003), Table 1, column 1, page 84. Results are robust to controlling for the growth of GDP per capita, the growth of GDP per capita lagged, a lagged dichotomous indicator of democracy (instead of the Polity2 index), a Subsaharan Africa dummy, an Asian dummy, a North Africa / Middle East dummy, and an anocracy dummy.
The data is from Fearon and Laitin (2003), except for POL (authors' calculations from Ethnologue database).

Table 5: Redistribution and Linguistic Fractionalization (1985-1995)
Dependent variable: Transfers and Subsidies as Share of GDP, least squares estimator

	(1)	(2)	(3)	(4)	(5)
	ELF(1)	**ELF(3)**	**ELF(6)**	**ELF(10)**	**ELF(15)**
ELF (at different levels of aggregation)	**-4.013 [2.26]****	**-3.797 [2.48]****	**-1.314 [1.00]**	**-1.731 [1.23]**	**-1.614 [1.15]**
Log GDP per capita 1985-95	1.036 [2.03]**	0.980 [1.98]*	0.944 [1.81]*	0.953 [1.84]*	0.969 [1.86]*
Log population 1985-95	0.033 [0.12]	0.036 [0.14]	0.096 [0.36]	0.103 [0.38]	0.108 [0.40]
Population above 65	0.684 [4.17]***	0.717 [4.27]***	0.704 [4.16]***	0.703 [4.13]***	0.699 [4.10]***
UK legal origin	5.169 [2.45]**	5.187 [2.56]**	4.643 [2.17]**	4.804 [2.20]**	4.810 [2.19]**
French legal origin	4.497 [1.86]*	4.657 [1.99]*	4.371 [1.76]*	4.378 [1.75]*	4.388 [1.75]*
Socialist legal origin	9.224 [2.97]***	9.191 [3.04]***	8.615 [2.77]***	8.634 [2.77]***	8.66 [2.78]***
Scandinavian legal origin	7.908 [1.91]*	7.490 [1.84]*	7.349 [1.77]*	7.317 [1.76]*	7.343 [1.76]*
% Catholic 1980	0.034 [1.62]	0.035 [1.68]*	0.035 [1.69]*	0.036 [1.73]*	0.036 [1.72]*
% Muslim 1980	-0.024 [1.03]	-0.019 [0.81]	-0.024 [1.02]	-0.023 [1.00]	-0.024 [1.03]
% Protestant 1980	-0.024 [0.59]	-0.023 [0.54]	-0.026 [0.60]	-0.024 [0.57]	-0.025 [0.58]
Small island dummy	-6.456 [3.62]***	-6.442 [3.57]***	-5.980 [3.50]***	-6.075 [3.55]***	-6.056 [3.54]***
Latitude	8.423 [1.84]*	8.750 [1.97]*	9.245 [2.02]**	8.847 [2.01]**	8.924 [2.02]**
Latin America and Caribbean	-4.464 [2.78]***	-4.618 [2.90]***	-4.861 [3.05]***	-5.077 [3.22]***	-5.050 [3.21]***
Sub Saharan Africa	-1.464 [0.98]	-1.446 [0.96]	-1.480 [0.99]	-1.171 [0.74]	-1.103 [0.69]
East Asia & Pacific	-3.218 [1.68]*	-2.799 [1.51]	-3.254 [1.72]*	-3.286 [1.77]*	-3.299 [1.78]*
Constant	-10.058 [1.44]	-9.938 [1.46]	-10.550 [1.52]	-10.491 [1.49]	-10.712 [1.52]
R-squared	0.84	0.85	0.84	0.84	0.84

(t-statistics based on robust standard errors, in parentheses)
* significant at 10%; ** significant at 5%; *** significant at 1%
All columns involve 101 country observations.
The specification is based on Desmet, Ortuño-Ortin and Weber (2009), Table 2, column 8, i.e. the specification with the most controls (considering instead the other seven specifications in Desmet, Ortuño-Ortin and Weber, 2009, with fewer regressors did not change the pattern of coefficients on ELF, although the degree of statistical significance remains high at higher levels of aggregation).
The data is from Desmet, Ortuño-Ortin and Weber (2009), except for ELF (authors' calculations from Ethnologue database).

Table 6: Redistribution and Linguistic Polarization (1985-1995)
Dependent variable: Transfers and Subsidies as Share of GDP, least squares estimator

	(1)	(2)	(3)	(4)	(5)
	POL(1)	**POL(3)**	**POL(6)**	**POL(10)**	**POL(15)**
POL (at different levels of aggregation)	**-2.232 [2.09]****	**-3.066 [2.63]****	**-0.994 [0.69]**	**-1.726 [0.98]**	**-1.459 [0.82]**
Log GDP per capita 1985-95	1.026 [2.00]**	0.994 [2.02]**	0.983 [1.82]*	1.027 [1.89]*	0.989 [1.84]*
Log population 1985-95	0.060 [0.22]	0.048 [0.18]	0.074 [0.27]	0.052 [0.19]	0.052 [0.19]
Population above 65	0.685 [4.13]***	0.689 [4.10]***	0.700 [4.10]***	0.685 [3.91]***	0.691 [3.96]***
UK legal origin	5.120 [2.41]**	5.175 [2.61]**	4.490 [2.13]**	4.646 [2.07]**	4.618 [2.09]**
French legal origin	4.550 [1.87]*	4.669 [2.05]**	4.266 [1.72]*	4.309 [1.67]*	4.284 [1.69]*
Socialist legal origin	9.218 [2.95]***	9.065 [3.09]***	8.533 [2.78]***	8.487 [2.71]***	8.447 [2.73]***
Scandinavian legal origin	7.813 [1.90]*	6.864 [1.71]*	7.082 [1.73]*	6.883 [1.69]*	7.004 [1.72]*
% Catholic 1980	0.033 [1.57]	0.034 [1.66]	0.035 [1.69]*	0.034 [1.66]	0.034 [1.67]*
% Muslim 1980	-0.025 [1.09]	-0.023 [0.98]	-0.026 [1.13]	-0.027 [1.18]	-0.027 [1.18]
% Protestant 1980	-0.024 [0.59]	-0.019 [0.48]	-0.026 [0.62]	-0.026 [0.65]	-0.027 [0.66]
Small island dummy	-6.316 [3.52]***	-6.538 [3.54]***	-6.054 [3.42]***	-6.303 [3.45]***	-6.224 [3.40]***
Latitude	8.547 [1.87]*	9.466 [2.13]**	9.599 [2.11]**	9.839 [2.17]**	9.967 [2.16]**
Latin America and Caribbean	-4.463 [2.80]***	-4.82 [3.07]***	-4.776 [3.00]***	-4.936 [3.18]***	-4.856 [3.13]***
Sub Saharan Africa	-1.464 [0.99]	-1.707 [1.17]	-1.445 [0.98]	-1.474 [1.01]	-1.532 [1.05]
East Asia & Pacific	-3.292 [1.71]*	-3.279 [1.84]*	-3.438 [1.83]*	-3.607 [1.97]*	-3.524 [1.93]*
Constant	-10.400 [1.48]	-9.777 [1.45]	-10.455 [1.52]	-10.030 [1.45]	-9.938 [1.42]
R-squared	0.84	0.85	0.84	0.84	0.84

(t-statistics based on robust standard errors, in parentheses)
* significant at 10%; ** significant at 5%; *** significant at 1%
All columns involve 101 country observations.
The specification is based on Desmet, Ortuño-Ortin and Weber (2009), Table 2, column 8, i.e. the specification with the most controls (considering instead the other seven specifications in Desmet, Ortuño-Ortin and Weber, 2009, with fewer regressors did not change the pattern of coefficients on POL, although the degree of statistical significance remains high at higher levels of aggregation).
The data is from Desmet, Ortuño-Ortin and Weber (2009), except for POL (authors' calculations from Ethnologue database).

Table 7 – Public Goods and ELF at various levels of aggregation, OLS estimates
(dependent variable listed in the leftmost column)

	ELF(1)	ELF(3)	ELF(6)	ELF(10)	ELF(15)	# of obs.	Adj-R^2 min	Adj-R^2 max
Output of Public Goods (from LLSV, 1999)[a]								
Log infant mortality	0.189	0.125	0.183	0.352	0.367	172	0.82	0.84
	[1.08]	[0.88]	[1.67]*	[3.28]***	[3.48]***			
Log of school attainment	-0.084	-0.139	-0.180	-0.212	-0.224	101	0.77	0.78
	[0.42]	[0.89]	[1.30]	[1.88]*	[2.11]**			
Illiteracy rate	-0.261	5.115	9.529	12.643	12.209	119	0.65	0.67
	[0.04]	[0.87]	[1.87]*	[2.97]***	[2.89]***			
Infrastructure quality index	-0.390	0.633	0.294	0.164	0.174	59	0.82	0.83
	[0.63]	[1.12]	[0.63]	[0.37]	[0.41]			
Additional Measures of Public Goods								
Hospital beds	0.254	1.206	0.849	0.854	0.992	169	0.47	0.47
(per 1,000 people)	[0.24]	[1.25]	[1.02]	[1.10]	[1.27]			
Measles immunization rates	-24.179	-17.987	-18.889	-14.782	-13.690	168	0.51	0.54
(% of children 12-23 months)	[4.00]***	[3.79]***	[4.75]***	[3.76]***	[3.46]***			
Improved sanitation facilities	-11.400	-13.431	-19.116	-22.281	-24.098	147	0.72	0.76
(% of population with access)	[1.32]	[1.94]*	[3.27]***	[3.93]***	[4.52]***			
Improved water source	-12.874	-10.881	-6.638	-13.264	-12.612	157	0.61	0.63
(% of population with access)	[1.74]*	[1.74]*	[1.29]	[2.94]***	[2.95]***			
Road network density	-4.203	-0.349	-0.918	-0.849	-1.031	150	0.33	0.33
(km per 1,000 inhabitants)	[1.13]	[0.13]	[0.44]	[0.39]	[0.48]			
Rail network density	0.206	0.183	0.022	0.044	0.049	88	0.45	0.46
(km per 1,000 inhabitants)	[0.91]	[1.10]	[0.19]	[0.37]	[0.42]			

Robust t statistics in brackets; * significant at 10%; ** significant at 5%; *** significant at 1%.
For all regressions, the specification is that of Table 6 of LLSV (1999), p. 256-260, including the broadest set of controls: Socialist legal origin dummy, French legal origin dummy, German legal origin dummy, Scandinavian legal origin dummy, Catholic share, Muslim share, other religions' share, latitude, log GNP per capita.
The table reports coefficient estimates on linguistic fractionalization (ELF) at various levels of aggregation, in regressions where the dependent variable is the one listed in the leftmost column.
The data is from LLSV (1999), World Bank (2008), except for ELF (authors' calculations from Ethnologue database).

Table 8 – Public Goods and POL at various levels of aggregation, OLS estimates (dependent variable listed in the leftmost column)

	POL(1)	POL(3)	POL(6)	POL(10)	POL(15)	# of Obs.	Adj-R^2 min	Adj-R^2 max
Output of Public Goods (from LLSV, 1999)[a]								
Log infant mortality	0.088	0.128	0.128	0.314	0.260	172	0.82	0.83
	[0.91]	[1.40]	[1.26]	[2.87]***	[2.40]**			
Log of school attainment	-0.025	-0.026	-0.034	0.042	0.034	101	0.77	0.77
	[0.25]	[0.29]	[0.34]	[0.38]	[0.28]			
Illiteracy rate	-0.060	5.908	4.395	4.836	-1.336	119	0.65	0.66
	[0.02]	[1.46]	[0.94]	[0.93]	[0.24]			
Infrastructure quality index	-0.274	0.348	0.378	0.429	0.424	59	0.82	0.83
	[0.76]	[0.86]	[0.80]	[0.88]	[0.82]			
Additional Measures of Public Goods								
Hospital beds	0.322	1.052	1.096	0.916	0.966	169	0.47	0.47
(per 1,000 people)	[0.49]	[1.37]	[1.14]	[0.88]	[0.85]			
Measles immunization rates	-12.868	-9.928	-13.710	-8.554	-6.936	168	0.48	0.51
(% of children 12-23 months)	[3.66]***	[2.80]***	[3.68]***	[2.04]**	[1.52]			
Improved sanitation facilities	-4.816	-10.852	-17.658	-15.255	-15.305	147	0.72	0.74
(% of population with access)	[0.98]	[2.24]**	[3.23]***	[2.54]**	[2.59]**			
Improved water source	-7.409	-7.450	-7.160	-10.169	-3.948	157	0.61	0.62
(% of population with access)	[1.73]*	[1.68]*	[1.59]	[2.08]**	[0.79]			
Road network density	-2.670	0.270	0.533	-0.332	-0.285	150	0.33	0.34
(km per 1,000 inhabitant)	[1.35]	[0.14]	[0.22]	[0.13]	[0.11]			
Rail network density	0.085	0.101	0.074	-0.005	-0.040	88	0.45	0.45
(km per 1,000 inhabitant)	[0.77]	[0.87]	[0.57]	[0.03]	[0.29]			

Robust t statistics in brackets; * significant at 10%; ** significant at 5%; *** significant at 1%.

For all regressions, the specification is that of Table 6 of LLSV (1999), p. 256-260, including the broadest set of controls: Socialist legal origin dummy, French legal origin dummy, German legal origin dummy, Scandinavian legal origin dummy, Catholic share, Muslim share, other religions' share, latitude, log GNP per capita.

The table reports coefficient estimates on linguistic polarization (POL) at various levels of aggregation, in regressions where the dependent variable is the one listed in the leftmost column.

The data is from LLSV (1999) and the World Bank (2008), except for POL (authors' calculations from Ethnologue database).

Table 9 – Growth and Linguistic Fractionalization
(Easterly and Levine specification, random effects estimator, 1960-1989 panel)

	(1)	(2)	(3)	(4)	(5)
	ELF(1)	**ELF(3)**	**ELF(6)**	**ELF(10)**	**ELF(15)**
ELF (at different levels of aggregation)	**-0.575 [0.58]**	**-0.560 [0.68]**	**-0.851 [1.25]**	**-1.367 [1.97]****	**-1.423 [2.10]****
Log of initial income	4.925 [1.85]*	4.860 [1.84]*	4.926 [1.87]*	4.575 [1.78]*	4.564 [1.78]*
Log of initial income squared	-0.348 [1.99]**	-0.342 [1.98]**	-0.345 [2.01]**	-0.328 [1.95]*	-0.327 [1.95]*
Dummy for Sub-Saharan Africa	-1.584 [2.81]***	-1.591 [2.83]***	-1.658 [3.03]***	-1.424 [2.48]**	-1.334 [2.28]**
Dummy for Latin America and Caribbean	-2.045 [4.97]***	-2.092 [4.96]***	-2.226 [4.98]***	-2.382 [5.24]***	-2.396 [5.30]***
Log of schooling	1.177 [2.34]**	1.147 [2.28]**	1.025 [2.03]**	1.050 [2.15]**	1.066 [2.19]**
Dummy for the 1960s	2.282 [8.54]***	2.280 [8.57]***	2.249 [8.45]***	2.215 [8.37]***	2.218 [8.34]***
Dummy for the 1970s	1.969 [7.22]***	1.967 [7.23]***	1.947 [7.24]***	1.936 [7.22]***	1.938 [7.21]***
Constant	-16.895 [1.64]	-16.614 [1.63]	-16.542 [1.63]	-14.649 [1.48]	-14.581 [1.48]
R-squared overall	0.36	0.36	0.37	0.38	0.38

(t-statistics based on robust standard errors in brackets)
* significant at 10%; ** significant at 5%; *** significant at 1%
Random effects estimates from an unbalanced panel of up to 94 countries, 259 observations over 3 decades (1960s, 1970s, 1980s).
The specification is that of column 1 of Table 4 in Easterly and Levine (1997), pp. 1225-1226 (also column 1 of Table 8 in Alesina et al. (2003), p. 168), which allows for the largest number of observations.
The pattern of significance and magnitude of the coefficients on ELF at various levels of aggregation are robust (and in some cases strengthened) when including additional controls included in Easterly and Levine (1997), namely 1) controlling for political assassinations 2) controlling for political assassinations plus financial depth, the black market premium, and the fiscal surplus to GDP ratio. 3) controlling for all of the above plus the number of coups d'etat, the number of revolutions, a dummy for civil wars and a measure of political rights (Gastil's index of democracy).
The data is from Easterly and Levine (1997), except for ELF (authors' calculations from Ethnologue database).

Table 10 – Growth and Linguistic Polarization
(Easterly and Levine specification, random effects estimator, 1960-1989 panel)

	(1)	(3)	(6)	(10)	(15)
	POL(1)	**POL(3)**	**POL(6)**	**POL(10)**	**POL(15)**
POL (at different levels of aggregation)	**-0.367 [0.65]**	**-0.699 [1.23]**	**-0.682 [1.04]**	**-0.467 [0.66]**	**-0.028 [0.04]**
Log of initial income	4.919 [1.85]*	4.830 [1.84]*	4.845 [1.83]*	4.915 [1.85]*	4.883 [1.83]*
Log of initial income squared	-0.347 [1.99]**	-0.340 [1.99]**	-0.339 [1.97]**	-0.346 [2.00]**	-0.344 [1.97]**
Dummy for Sub-Saharan Africa	-1.579 [2.80]***	-1.559 [2.76]***	-1.593 [2.83]***	-1.570 [2.78]***	-1.595 [2.85]***
Dummy for Latin America and Caribbean	-2.044 [4.97]***	-2.127 [5.07]***	-2.170 [4.90]***	-2.137 [4.79]***	-2.052 [4.57]***
Log of schooling	1.183 [2.35]**	1.157 [2.31]**	1.108 [2.22]**	1.183 [2.32]**	1.181 [2.34]**
Dummy for the 1960s	2.281 [8.53]***	2.284 [8.54]***	2.286 [8.50]***	2.293 [8.45]***	2.288 [8.45]***
Dummy for the 1970s	1.969 [7.22]***	1.969 [7.23]***	1.965 [7.22]***	1.973 [7.18]***	1.971 [7.20]***
Constant	-16.853 [1.64]	-16.394 [1.62]	-16.482 [1.62]	-16.831 [1.64]	-16.866 [1.64]
R-squared overall	0.36	0.37	0.37	0.36	0.36

(t-statistics based on robust standard errors in brackets)
* significant at 10%; ** significant at 5%; *** significant at 1%.
Random effects estimates from an unbalanced panel of up to 94 countries, 259 observations over 3 decades (1960s, 1970s, 1980s).
The specification is that of column 1 of Table 4 in Easterly and Levine (1997), pp. 1225-1226 (also column 1 of Table 8 in Alesina et al. (2003), p. 168), which allows for the largest number of observations.
The pattern of significance and magnitude of the coefficients on ELF at various levels of aggregation are similar when including additional controls included in Easterly and Levine (1997), namely 1) controlling for political assassinations 2) controlling for political assassinations plus financial depth, the black market premium, and the fiscal surplus to GDP ratio. 3) controlling for all of the above plus the number of coups d'etat, the number of revolutions, a dummy for civil wars and a measure of political rights (Gastil's index of democracy).
The data is from Easterly and Levine (1997), except for POL (authors' calculations from Ethnologue database).

Table 11 – Growth and Linguistic Fractionalization
(Augmented Solow specification, random effects estimator, 1960-2004 panel)

	(1)	(2)	(3)	(4)	(5)
	ELF(1)	**ELF(3)**	**ELF(6)**	**ELF(10)**	**ELF(15)**
ELF (at various levels of aggregation)	**-1.074 [1.11]**	**-1.280 [1.64]**	**-1.354 [2.13]****	**-1.849 [2.98]*****	**-1.640 [2.65]*****
Log initial real per capita GDP	-1.848 [7.52]***	-1.855 [7.60]***	-1.821 [7.69]***	-1.845 [7.83]***	-1.832 [7.74]***
Investment share of GDP	0.095 [4.51]***	0.093 [4.46]***	0.089 [4.18]***	0.088 [4.30]***	0.088 [4.25]***
Avg. schooling years in the total population aged 25+	0.260 [3.13]***	0.259 [3.15]***	0.232 [2.89]***	0.222 [2.76]***	0.229 [2.82]***
Growth of population	-0.507 [2.92]***	-0.505 [2.90]***	-0.488 [2.73]***	-0.474 [2.69]***	-0.478 [2.71]***
Log population	0.290 [1.66]*	0.316 [1.81]*	0.322 [1.86]*	0.338 [1.97]**	0.341 [1.97]**
Interaction between openness and log population	-0.006 [2.29]**	-0.006 [2.25]**	-0.005 [2.26]**	-0.005 [2.13]**	-0.005 [2.15]**
Openness (imports + exports over GDP)	0.057 [2.69]***	0.057 [2.70]***	0.056 [2.72]***	0.054 [2.63]***	0.055 [2.64]***
Latin America and Carribean dummy	-0.739 [1.86]*	-0.833 [2.08]**	-1.062 [2.49]**	-1.208 [2.85]***	-1.151 [2.71]***
Sub-Saharan Africa dummy	-2.193 [3.69]***	-2.189 [3.71]***	-2.322 [3.90]***	-1.950 [3.40]***	-1.869 [3.21]***
Constant	13.895 [5.40]***	13.863 [5.50]***	13.933 [5.62]***	14.237 [5.76]***	13.992 [5.64]***
R-squared	0.38	0.39	0.39	0.40	0.40

Absolute value of t statistics, based on robust standard errors, in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%.
Random effects estimates based on an unbalanced panel of 101 countries over 5 periods (1960-69, 1970-79, 1980-89, 1990-99, 2000-2004), 428 observations.
The specification includes period dummies for 1970-1979, 1980-1989, 1990-1999 and 2000-2004 (estimates not reported).
Investment, schooling, population growth, and openness are entered as period averages; log initial per capita income and log population are for the first year of each period.
The data on income per capita, income growth, population, population growth, openness and investment are from the Penn World Tables, version 6.2 (Heston, Summers and Aten, 2006). The data on human capital is from Barro-Lee (2000). The geographic controls are from the CIA World Factbook (2009). The ELF data is from the authors' calculations using the Ethnologue database.

Table 12 – Growth and Linguistic Polarization
(Augmented Solow specification, random effects estimator, 1960-2004 panel)

	(1)	(2)	(3)	(4)	(5)
	POL(1)	**POL(3)**	**POL(6)**	**POL(10)**	**POL(15)**
POL (at various levels of aggregation)	**-0.659 [1.17]**	**-0.980 [1.87]***	**-0.874 [1.57]**	**-0.207 [0.34]**	**-0.034 [0.05]**
Log initial real per capita GDP	-1.851 [7.51]***	-1.851 [7.57]***	-1.814 [7.57]***	-1.833 [7.59]***	-1.833 [7.57]***
Investment share of GDP	0.095 [4.51]***	0.092 [4.40]***	0.092 [4.33]***	0.095 [4.44]***	0.095 [4.46]***
Avg. schooling years in the total population aged 25+	0.261 [3.15]***	0.268 [3.23]***	0.263 [3.18]***	0.268 [3.21]***	0.267 [3.20]***
Growth of population	-0.506 [2.92]***	-0.499 [2.88]***	-0.493 [2.79]***	-0.513 [2.91]***	-0.518 [2.94]***
Log population	0.289 [1.65]*	0.303 [1.74]*	0.281 [1.60]	0.286 [1.62]	0.285 [1.62]
Interaction between openness and log population	-0.006 [2.29]**	-0.005 [2.24]**	-0.005 [2.16]**	-0.006 [2.24]**	-0.006 [2.26]**
Openness (imports + exports over GDP)	0.057 [2.68]***	0.056 [2.66]***	0.054 [2.56]**	0.055 [2.60]***	0.055 [2.61]***
Latin America and Carribean dummy	-0.741 [1.87]*	-0.861 [2.17]**	-0.933 [2.24]**	-0.783 [1.84]*	-0.748 [1.76]*
Sub-Saharan Africa dummy	-2.190 [3.69]***	-2.173 [3.71]***	-2.217 [3.77]***	-2.180 [3.69]***	-2.181 [3.68]***
Constant	13.953 [5.41]***	14.026 [5.52]***	13.986 [5.55]***	13.772 [5.42]***	13.701 [5.40]***
R-squared	0.38	0.39	0.39	0.38	0.38

Absolute value of t statistics, based on robust standard errors, in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%.
Random effects estimates based on an unbalanced panel of 101 countries over 5 periods (1960-69, 1970-79, 1980-89, 1990-99, 2000-2004), 428 observations.
The specification includes period dummies for 1970-1979, 1980-1989, 1990-1999 and 2000-2004 (estimates not reported).
Investment, schooling, population growth, and openness are entered as period averages; log initial per capita income and log population are for the first year of each period.
The data on income per capita, income growth, population, population growth, openness and investment are from the Penn World Tables, version 6.2 (Heston, Summers and Aten, 2006). The data on human capital is from Barro-Lee (2000). The geographic controls are from the CIA World Factbook (2009). The POL data is from the authors' calculations using the Ethnologue database.

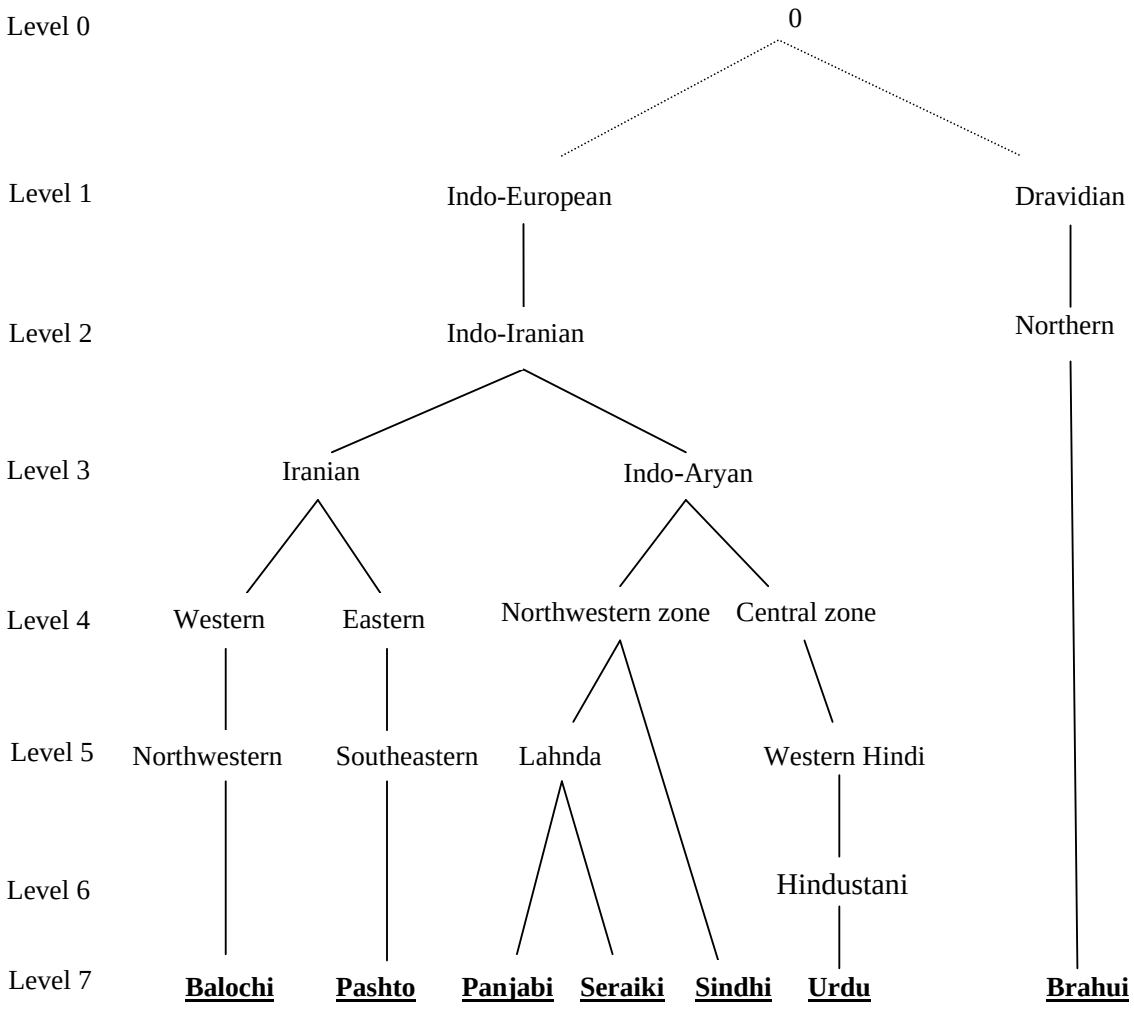


Figure 1. Phylogenetic Tree of Major Languages in Pakistan

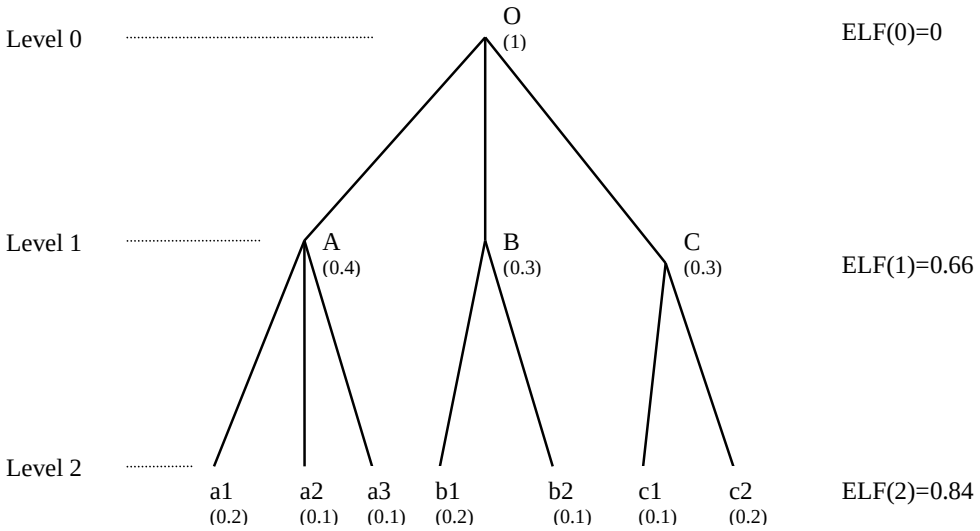


Figure 2 - Hypothetical Language Tree.

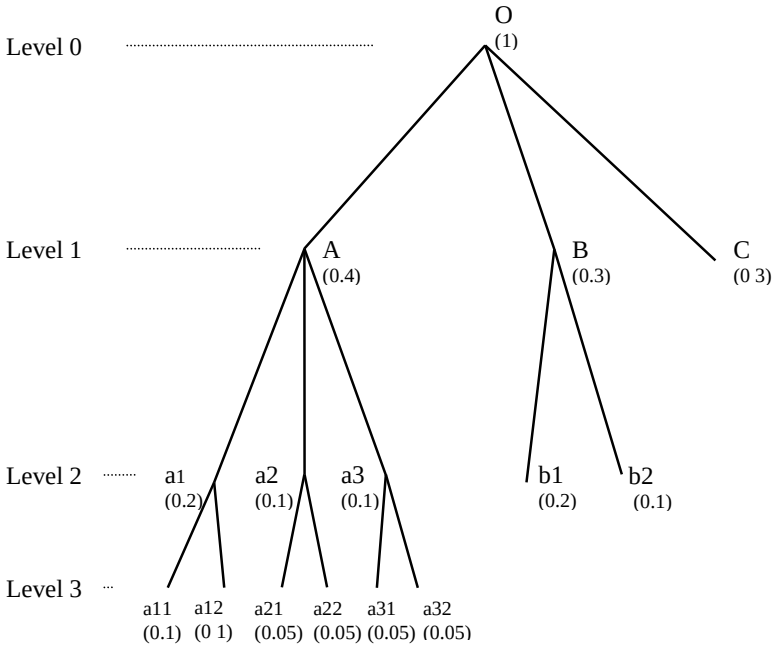
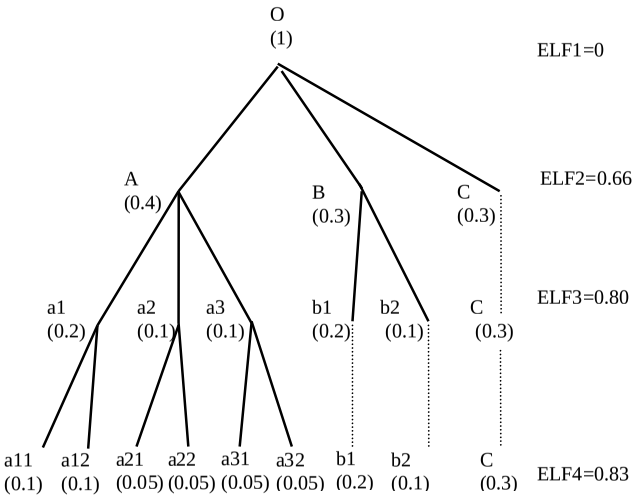
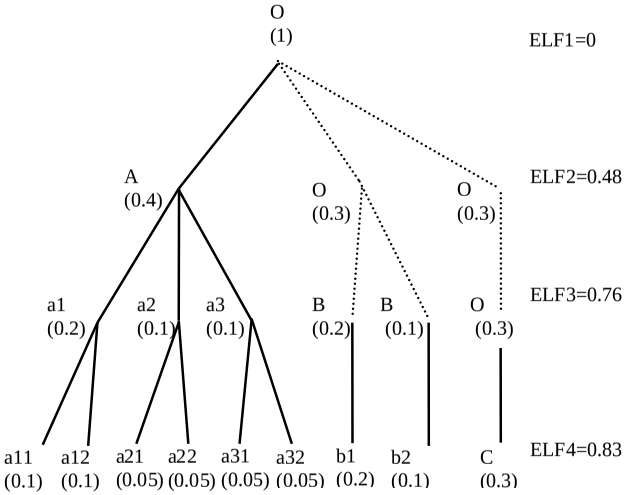


Figure 3 - Language Tree from Ethnologue.



Panel I - Approach 1



Panel II - Approach 2

Figure 4 - Two Different Approaches.

Figure 5 - Distributions of ELF(1) and ELF(15)

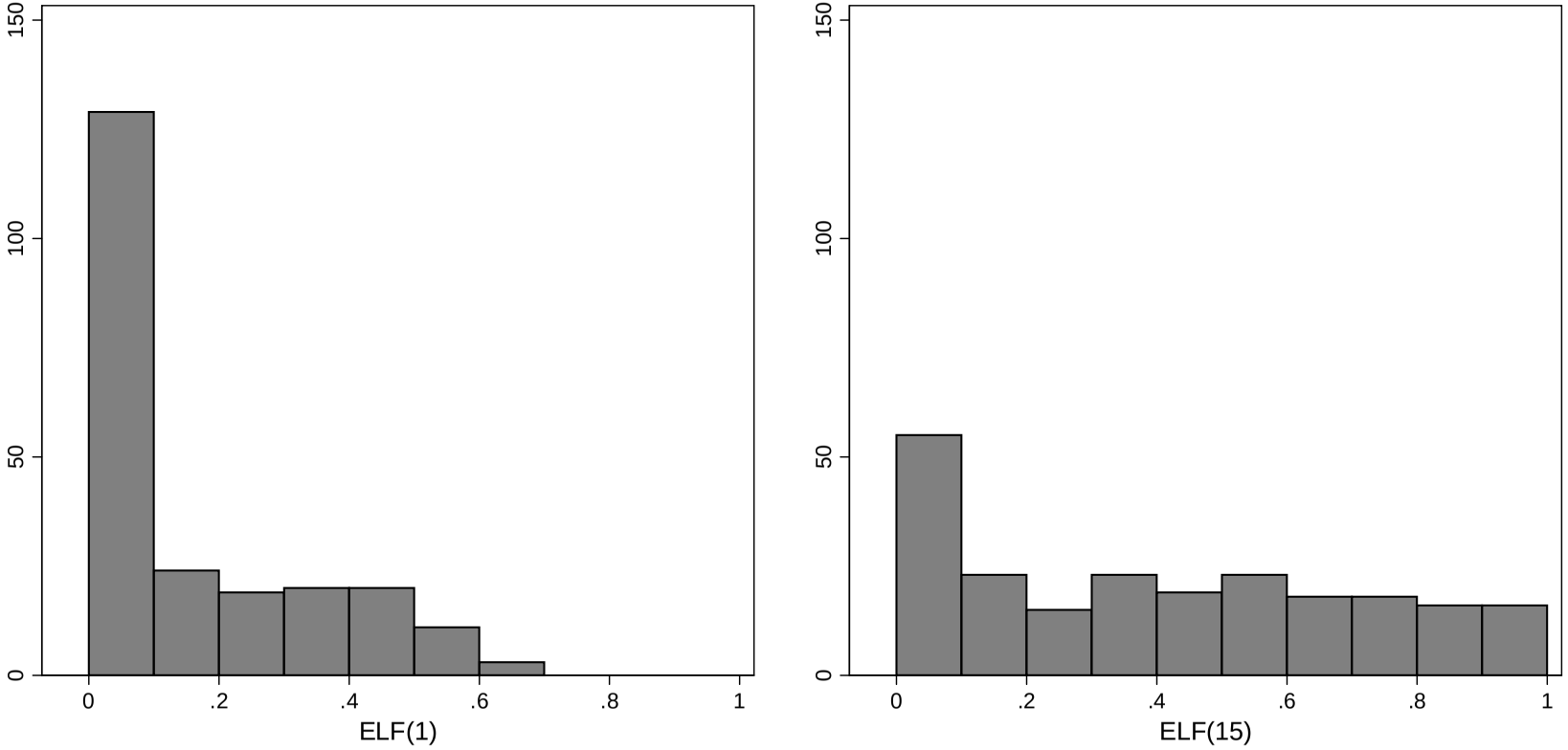


Figure 6 - Distributions of POL(1) and POL(15)

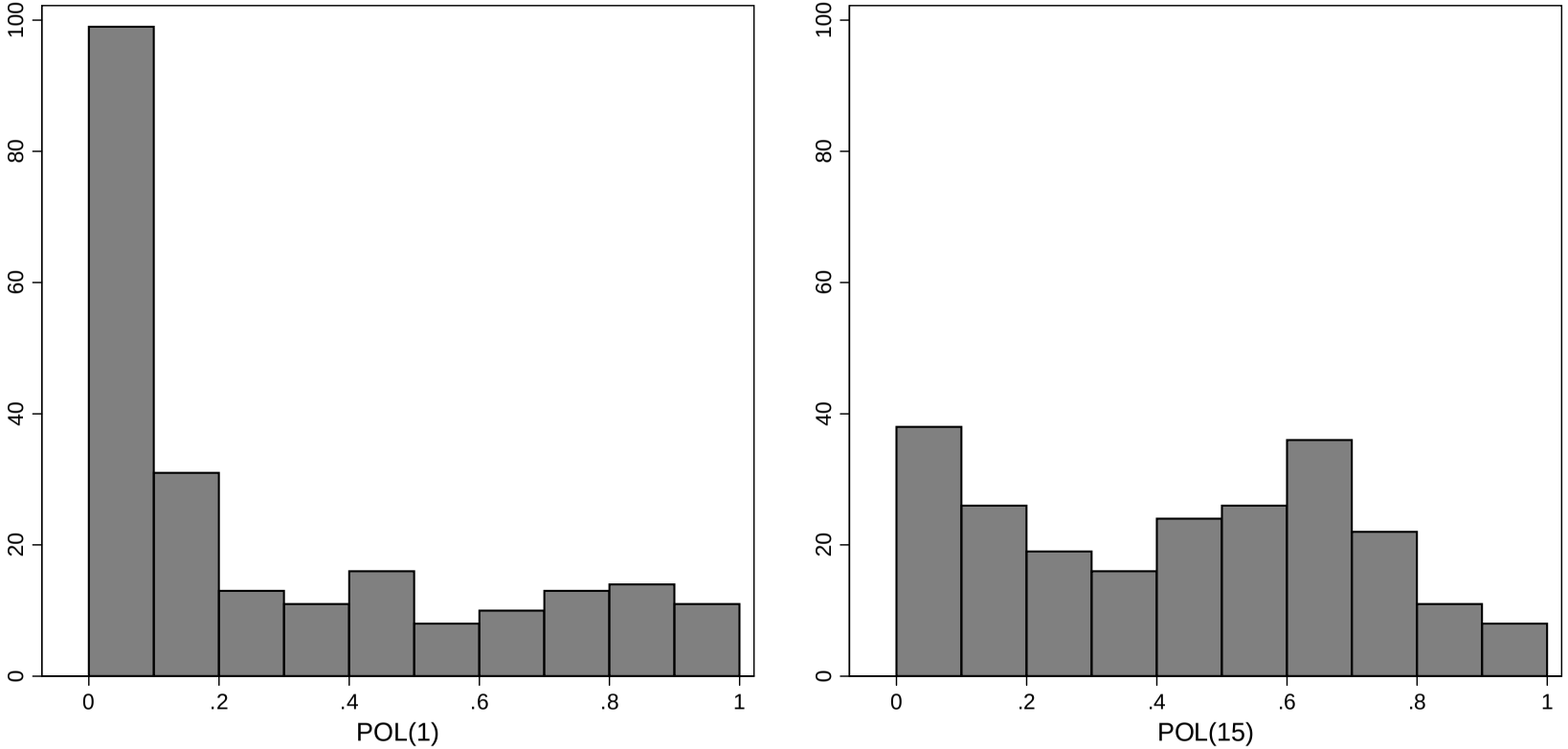


Figure 7 - Marginal effect of a one standard deviation increase in ELF as a % of the mean probability of civil conflict onset

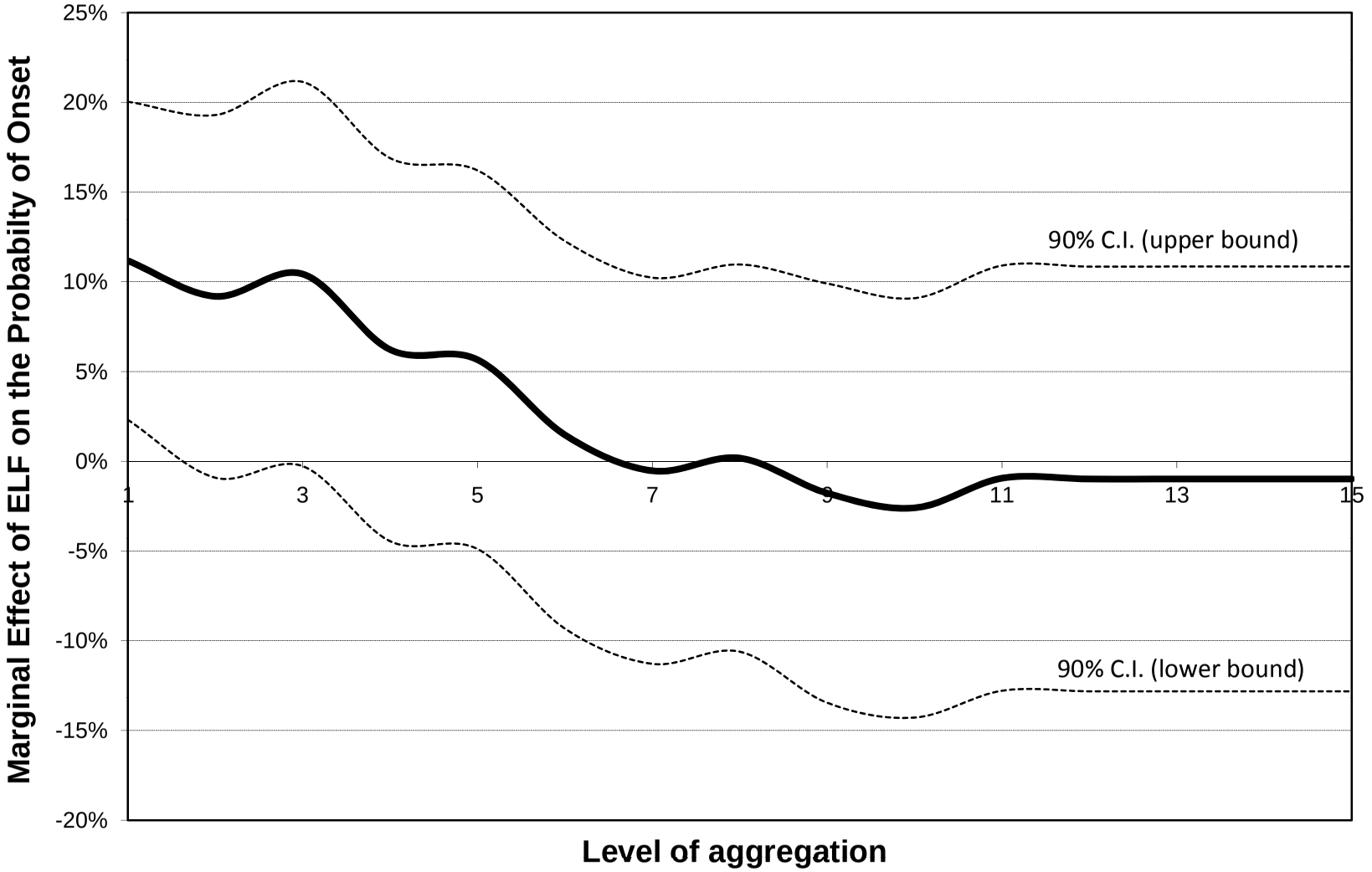


Figure 8 - Effect of a one standard deviation increase in ELF on redistribution (as a % of standard deviation of redistribution)

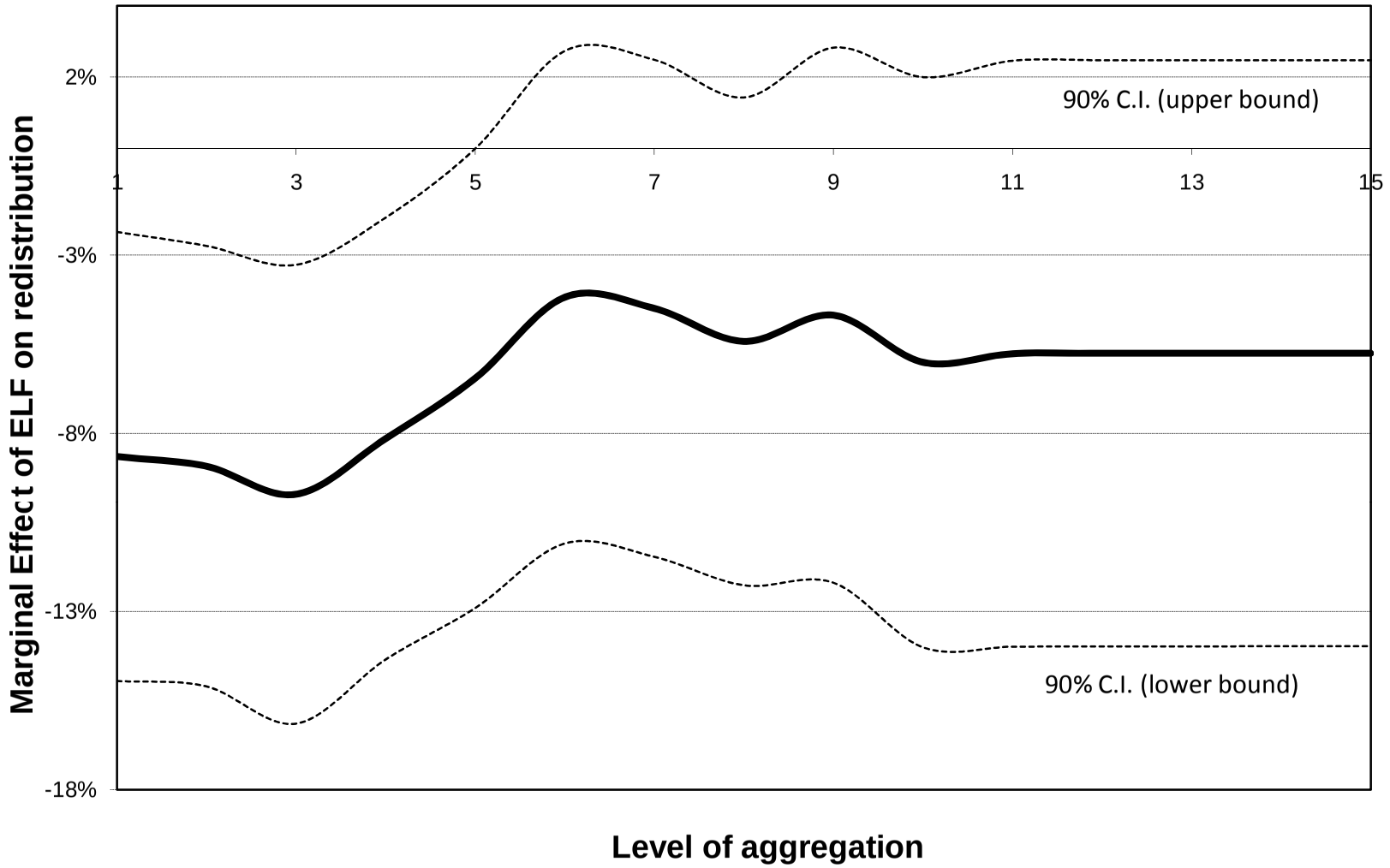


Figure 9 - Marginal effect of a one standard deviation increase in ELF as a % of a standard deviation of growth - Easterly-Levine regressions

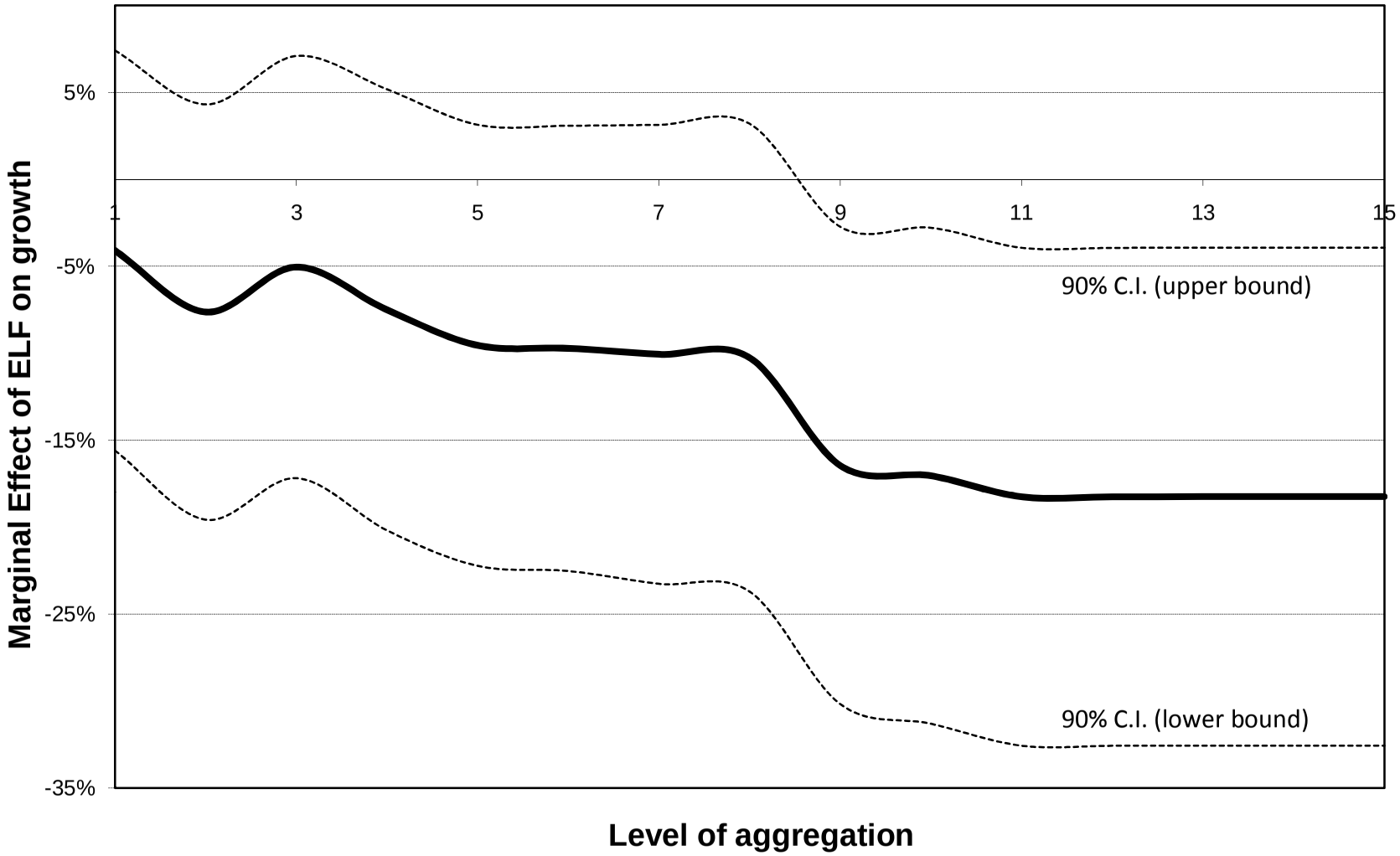


Figure 10 - Marginal effect of a standard deviation increase in ELF as a % of a standard deviation of growth - Augmented Solow regressions