Industrialization and ethnic change

Industrialization and Ethnic Change in the Modern World

Elliott Green[1]

Department of International Development, London School of Economics, London, UK

Abstract:

Despite the large recent attention given to ethnicity within the social sciences, the sources of modern ethnic change have remained opaque. Drawing upon social theory from Marx and Gellner, I argue here that industrialization incentivizes ethnic homogenization by lowering the relative value of land. Using carbon emissions per capita as a proxy for industrialization, I show that cross-country changes in ethno-linguistic fractionalization between 1961 and 1985 are negatively correlated with industrialization, and that this result is robust to the use of a variety of control variables, sub-samples and alternative measures of industrialization such as cement production, urbanization and agriculture as a percentage of GDP. In particular I find no evidence for the direct role of the state in promoting ethnic homogenization, which adds to other recent evidence on how economic incentives may trump political ones as regards identity change, at least in the short- to medium term.

Keywords: ethnic change; industrialization; ELF; structural transformation; urbanization; ethnic identity

1. Introduction

The literature on ethnicity has burgeoned across the social sciences in recent decades, with an increasing degree of sophistication. Scholars from political science, sociology, and economics have all contributed to our collective understanding of ethnicity within recent years, including

[1] E.d.green@lse.ac.uk.

notable research into how and why ethnic diversity undermines cooperation (Habyarimana et al. 2009), the interaction between ethnic identity and party politics (Posner 2005), and how ethnicity affects conflict (Cederman and Girardin 2007; Esteban and Ray 2008; Wimmer et al. 2009), among other topics. There has also been an attempt to explain variation in ethnic diversity across the world, with suggestion that ethnic diversity can be explained as a result of state history, genetic diversity, land diversity, modern state strength and colonial rule (Ahlerup and Olsson 2012; Ashraf and Galor 2013; Bleaney and Dimico 2016; Michalopoulos 2012; Wimmer 2015).

Yet despite the advancements of this literature one topic still remains somewhat opaque, namely why ethnic identity changes across time. Despite a general agreement that people can hold multiple ethnic identities and that the salience of these identities is not fixed, there has nonetheless been very few attempts to systematically examine the dynamics of ethnic fractionalization, especially in a cross-national context. As such I show here that industrialization promotes ethnic homogenization. More specifically, I draw upon a long tradition of social theory from Marx and Gellner, among others, to argue that industrialization creates incentives for people to re-identify as part of larger ethnic groups in a bottom-up manner. In particular I suggest that industrialization lowers the relative value of land, thereby disincentiving the use of more narrow rural ethnic identities. I also test for the possibility that ethnic homogenization is driven in a top-down manner by state policies through such proxies such as schooling and the provision of public goods, and fail to find evidence that states are directly responsible for ethnic change.

To test this theory I use cross-national Soviet data on changes in ethnic diversity over time in a long difference model to show that industrialization, as proxied by change in log of carbon emissions per capita, is negatively correlated with change in ethnic diversity, a result which is robust to various control variables and sub-samples. I show that ethnic homogenization is being driven by increasing percentages of people who identify with the largest ethnic group in each country, and that the results do not change when I use alternative measures of industrialization such

as changes in cement production per capita, urbanization and the percentage of GDP derived from agriculture.

The paper is organized as follows. Next in section 2 I discuss the use of ethnicity as both an exogenous and endogenous variable, with attention to social theorists like Marx and Engels, Gellner and others who have argued for a causal effect of industrialization on ethnic homogenization. Section 3 presents my cross-national quantitative analysis, including the use of controls and sub-samples, while section 4 examines change in the percentage identifying with the largest ethnic group and alternative measures of industrialization. In section 5 I conclude.

2. Understanding Ethnicity and Ethnic Change

Ethnicity is a highly contested term within the social sciences (Brubaker 2004). I do not wish to delve into debates on the meaning of ethnicity except to note that I define ethnicity here along the lines suggested by many authors (Bentley 1987; Chandra 2006; Cohen 1978; Fearon 2003), namely that ethnic groups are social groups based on the idea that members share some form of common descent. The question about whether or not members of a given group actually do share common ancestors is largely irrelevant; what instead matters is that a belief in common descent is prevalent enough that it binds people together in a single community.

It is now widely acknowledged that individuals can have a multitude of nested ethnic identities, many of which can be considered as based on a belief in common descent and thus ethnic in nature (Bentley 1987, 35; Cohen 1978, 387). In Uganda and many other parts of Africa, for instance, one can simultaneously be a member of a sub-clan (for instance, the *Busito* sub-clan within the *Nte* clan of Buganda), a clan (*Nte*), a tribe (Buganda), a super-tribe, linguistic or regional group (Bantu-speakers or Southerners, groups which largely overlap) and a race (African). One way to conceptualize these different identities is as differently-sized Russian (*Matryoshka*) dolls or a series of concentric circles, with individuals simultaneously possessing multiple identities based

on different sized groups. A process of ethnic homogenization would thus consist of shifting one's primary identity from a smaller doll or circle to a larger one, thereby emphasizing an identity which is shared by more people than the previous smaller identity. The way this identify shift could be captured would be to examine change across time in a given population's measure of ethno-linguistic fractionalization (ELF). Each country's ELF score is calculated using the Herfindahl concentration formula, namely by summing the squares of the percentages of all ethnic groups larger than one per cent of the population and subtracting this sum from one. Thus a decreasing ELF score would indicate a process of homogenization, while a rising ELF score would indicate increasing levels of diversity.

If individuals have multiple ethnic identities to choose from, one can hypothesize that they will choose to prioritize the ethnic identity that gives them the greatest net benefit. This is not a particularly radical assumption, inasmuch as it is the basis for the current scholarship on identity formation in the social sciences labeled as "instrumentalist." Here ethnicity is an instrument, or a means to an end, rather than an end in itself, in contrast to the "primordialist" school where identities are taken as a given.

The question is thus what major social, economic or political forces create incentives for people to alter their ethnic identity *en masse*. At the macro level most attention has centered on the role of the state in promoting ethnic/national homogenization (Malesevic 2013; Mann 2013; Weber 1976), and thus at the micro-case study level much of the literature has focused on political motivations for choosing ethnic identities in such examples as the former Soviet Union (Laitin 1998), post-colonial Zambia (Posner 2005) and colonial India (Cassan 2015). However, in none of these cases was the focus on explicitly testing competing explanations for what causes ethnic change *per se*, nor was there an attempt to generalize the findings by using cross-national data.

I posit here that the most profound force for promoting ethnic homogenization in the modern world is industrialization, inasmuch as industrialization leads to a decline in the relative value of land. (Here I define industrialization as the shift away from an economy based on agrarian

production.) Thus, as people leave the agrarian sector and enter the industrial sector they lose their attachment to smaller rural ethnic identities which formerly gave them access to land, and instead emphasize the salience of larger group identities which are more useful for gaining access to employment in the modern industrial sector.

To understand the effect of industrialization on ethnic identity it is necessary first to examine the nature of ethnic identity in a broad historical context. Recently many economists, ecologists and anthropologists have suggested that ethnic diversity across the world is correlated with a number of ecological factors like latitude (negatively), elevation (positively) and differential land endowments (positively) (Ahlerup and Olsson 2012; Michalopoulous 2012; Wimmer 2015). There are two possible interrelated reasons for these relationships. First, warm tropical environments with predictable climates (as regards the variability of temperature and rainfall) are ideal for becoming self-sufficient in food and thereby create few incentives for inhabitants to migrate or form social ties with people across a large amount of territory. The same logic applies to differential land endowments, which lead human groups to become specialized in growing certain types of crops and thus discourage migration to other areas that are not suitable for the same crops (Michalopoulous 2012). Second, even if humans wanted to migrate, mountains, different disease environments and dense tropical rainforests can create physical barriers and disincentives to movement and thereby promote cultural isolation.

Thus in the pre-modern or pre-industrial world isolation encouraged greater ethnic diversity. In such isolated societies it would have made no sense to identify with a foreign ethnic group inasmuch as livelihoods and income within the group depended on other members of that group and access to that group's land and ecological skill set. In contrast, however, the incentives for ethnic identification change with industrialization as people leave the rural agrarian economy to move to the industrial urban sector, where access to rural land, ethnic-specific agricultural skills and support from one's rural community are all relatively unimportant in making a living. Instead, what is valuable to the new industrial working class is their ability to earn an income through their labor,

which can be enhanced by shedding their previous narrow rural identity in favor of a larger identity shared by more people. In other words, as both people and profits move from rural to urban areas, access to labor become more important than land in the acquisition of income and wealth.

Theoretically the first and most forceful such statement identifying the homogenizing incentives of industrialization comes from Marx and Engels in their early writings. They clearly state that, in the premodern rural world individuals are “united by some bond: family, tribe, the land itself, etc.” and the economy is guided by an exchange between people and the natural environment (Marx and Engels 1978 [1848], 468-469). The ownership of land is the basis for the economy and thus people remain attached to the land. Individual peasants may be said to be members of an ethnic group, class or nation, but without any "community [or] national bond" they are merely like a "sack of potatoes" in Marx’s memorable phrase.

However, modern industrialization sees a huge shift in the basis of production as the basis for the economy and economic growth shifts from land to labor. Here members from previously isolated and separate rural communities come together to work, thereby making manifest for the first time the division between the “two great classes” of the bourgeoisie and the proletariat. Thus while in the rural sector laborers held location- or ethnic-specific capital, in the industrial sector they lose their distinctiveness and become part of a “floating” army of workers, whereby their exploitation at the hands of the bourgeoisie creates the impetus for class consciousness and subsequent revolution.

In other word, while not claiming explicitly that industrialization promotes ethnic homogenization, Marx and Engels were nonetheless clear about the way in which industrialization created incentives for social homogenization, in two ways. First, the bourgeoisie needed to create a homogenous workforce that it can employ in its factories and shops, a process which I henceforth label the "top-down" mechanism inasmuch as the process of ethnic homogenization is largely a result of the efforts of the ruling elite. However, Marx is explicit that the process of class formation is a byproduct of modern capitalism rather than a deliberate outcome directed by the bourgeoisie,

which has no incentive to create a large united class opposed to its interests. Instead, there is a second process of homogenization at work whereby members of the proletariat have strong incentives to cooperate through unions and political parties in order to fight the exploitation of the bourgeoisie. Indeed, Marx and Engels' famous call at the end of the *Manifesto* for workers of the world to unite is essentially based on their belief that, despite their varied backgrounds and cultural/ethnic differences, the advent of the modern industrial economy means that workers' interests now lie together with each other. This second process I label the "bottom-up" mechanism, since here homogenization is the result of the actions of the non-elite based on what they perceive to be their best interests.

The influence of Marx and Engels on Ernest Gellner's theory of nationalism, where industrialization is instead made explicitly responsible for the rise of modern national identities, is quite clear (despite the fact that Gellner did not consider himself a Marxist). Like Marx and Engels, Gellner also draws a sharp divide between the pre-modern rural world and the modern industrial world, whereby the former consists of "small peasant communities" leading "inward-turned lives, tied to the locality by economic need" (Gellner 2006 [1983], 10). Far from there being incentives to homogenize or assimilate, "no one, or almost no one, has an interest in promoting cultural homogeneity at this social level" (Gellner 2006 [1983], 11).

However, just as with Marx and Engels, Gellner posits that industrialization requires an educated and culturally homogenous workforce and thus employers encourage cultural assimilation of rural-urban migrants into the dominant ethnic group of the newly industrialized areas. Those who assimilate into this dominant group – known as the "Megalomanians" in Gellner's famous story – do so in order to obtain jobs in the modern economy. As with Marx, Gellner implicitly provides both "top-down" and "bottom-up" mechanisms by which industrialization promotes ethnic homogenization. On the one hand the Megalomanian state elite promote a common Megalomanian identity in a conscious "top-down" manner, in particular via mass education in the former high culture of the Megalomanian elite. However, there are also strong incentives for various ethnic

minorities to assimilate as Megalomanian in order to share in the state's "growing prosperity" (Gellner 2006 [1983], 59), with or without state pressure. Thus one of my main goals below will be not only to test the role of industrialization in promoting ethnic homogenization, but to examine which of the two of these mechanisms finds more support with the data.

Before we continue, it is important to note two additional features of the Marxian/Gellnerian framework. First, for both Marx and Gellner industrialization generates incentives for most people not just to identify with a larger ethnic group but specifically with the largest group in society, namely the proletariat and the Megalomanians, respectively. Second, in both cases the process of homogenization is temporary, inasmuch as it takes place while countries transition to a capitalist, industrialized economy. Thus for Marx the development of class consciousness is one stage in history rather than an ongoing process. For Gellner it is the transition from agrarian society to industrial society which gives rise to nationalism, especially via uneven industrialization. In neither case should we expect homogenization to continue or decline once industrialization is complete, at least after controlling for migration. Both of these features then present themselves as additional theories to be tested below.

3. Cross-National Quantitative Evidence

If industrialization leads to ethnic homogenization, then cross-national data on ethnic diversity should exhibit this relationship. As such I could potentially regress a measure of ethnic diversity on measures of industrialization in a cross-sectional database in order to see if there was a correlation between lower levels of diversity and higher levels of industrialization. Such an exercise would also allow me to test the primordial assumption that ethnicity will only change due to the entry and exit of people from the population group in question, namely via international migration, which I measure here as the ratio of immigrants to the total population of a country. (Inasmuch as variance decreases with population size the migration data suffers from

heteroscedasticity, leading me to employ robust standard errors.) Of course, a cross-sectional correlation cannot help identify causality; as such I instead use a long-difference regression, where I regress change over two points in time in levels of ethnic diversity on change in industrialization in a cross-sectional dataset.

For the dependent variable I turn to country-level ELF scores for 1961 and 1985 as listed in two Soviet ethnographic atlases authored or co-authored by the ethnographer Solomon Bruk, which were themselves based on individual country censuses, UN data, Soviet ethnographic journals and other data (Bruk and Apanchenko 1964; Bruk 1986). The data from the *Atlas Narodov Mira* (henceforth ANM) lists in each case the total population of each country before listing individual ethnic groups and their sizes, often in a large amount of detail. For instance, the 1985 dataset lists 148 ethnic groups for Indonesia ranging in population from 74 million Javanese and 20.5 million Sundanese to 4000 Manusela, and 3000 Ambelau, among others. In other cases the Atlas lists the population of very small ethnic groups, such as the Alacalufe of Argentina and Chile, whose populations in 1985 were estimated at 100 and 200 people, respectively. In a small number of very homogenous countries such as Haiti, Japan and Poland, the ANM only lists one ethnic group per country, with the rest of the population listed as "other;" I address this issue in more detail below.

The ANM dataset – which was originally collated by Roeder (2001) but which I independently collected from the original sources – is ideal for several reasons. First, it is the only extant dataset that measures changes in ethnicity over time across the whole world, with specific attention to providing up-to-date population data in both atlases. Indeed, a close comparison of the 1985 ANM data with data on ethnic identity from demographic and health surveys from the late 1980s suggests that it adequately captures the definition of ethnicity discussed above, as discussed in Appendix 5. Second, it covers a large enough amount of time to uncover substantial amounts of ethnic change. Third and finally, the ANM dataset covers a large number of countries.[1]

The ANM dataset is, however, potentially problematic: as it largely relies upon census data, it is possible that it is not really capturing ethnic identity but identities imposed or at least

circumscribed by state officials (Anderson 1991). While this would be a serious problem if one was only examining a cross-sectional relationship, the use of long-difference regressions mean that such census inaccuracies would cancel out as long as they are consistent across the two Atlases. In order to see whether these inaccuracies were consistent or not, I examined whether all groups listed for one country in one Atlas are listed in the other; for cases where the listings are not consistent I either corrected the data or deleted the countries from the analysis, as detailed in Appendix 6. Another way to limit this problem is to examine subsets of countries which exclude potentially problematic cases as identified in the literature, which I employ below. One such group would be former British colonies, which arguably have a legacy of greater fractionalization than other former colonies due to a colonial interest in preventing ethnic homogenization (Robinson 2014); another includes African and European countries, which have lower rates of ethnic enumeration on censuses than other parts of the world (Morning 2008).

As regards measuring change in industrialization, unfortunately there are very few country-year observations for such variables as industry or agriculture as a percentage of GDP that date back to the early 1960s.[2] Instead I use log of carbon emissions in metric tons per capita, or CO_2 emissions, as a proxy for industrialization, as it is available for 133 countries going back to 1961. Carbon emissions are a good measure of industrialization for three reasons. First, thanks to efforts of researchers at the Carbon Dioxide Information Analysis Center in the US, there is high quality annual data on carbon emissions by country dating back to 1960, which derives from various UN country questionnaires and national statistical publications. The quality of data on carbon emissions is thus in sharp contrast to other potential measures of industrialization, as discussed more below. Second, carbon emissions are a good proxy for industrialization inasmuch as they are the result of modern industrial economies, mainly via fossil fuel consumption (for such purposes as electricity, heating and transportation), as well as cement production. Third, carbon emissions correspond to the non-linear, inverted-U relationship between economic development and industrialization, inasmuch as the rate of increase in carbon emissions peaks as countries industrialize and then

declines as they become post-industrial and more efficient at carbon production. Thus for most developed areas of the world the curve representing absolute levels of carbon emissions per capita plotted against time exhibits a sideways S-shape, such that the curve is relatively flat both today and farther back in the past but has a high positive slope as countries industrialize. This non-linear relationship also captures the essence of industrialization as a transition period, as discussed in the theoretical section above. However, as there might be concerns as to what degree carbon emissions are a good proxy for industrialization, I use three alternative measures of industrialization in my robustness tests below.

More formally, the cross-sectional regressions for both 1961 and 1985 can be expressed as

$$ELF_{1961i} = \alpha_{1961} + \beta_1 C_{1961i} + \beta_2 M_{1961i} + \gamma \mathbf{X}_i + \varepsilon_{1961i} \quad (1)$$

$$ELF_{1985i} = \alpha_{1985} + \beta'_1 C_{1985i} + \beta'_2 M_{1985i} + \gamma \mathbf{X}_i + \varepsilon_{1985i} \quad (2)$$

with ELF_{1961i} and ELF_{1985i} as measures of ELF in 1961 and 1985 for country *i*, C_{1961i} and C_{1985i} as levels of carbon emissions per capita in 1960 and 1985 for country *i*, M_{1961i} and M_{1985i} as the percentage of migrants in 1960 and 1985 for country *i*, and $\gamma \mathbf{X}_i$ as a vector of time-invariant controls, including many variables identified above as correlated with levels of ethnic diversity such as mean absolute latitude, mean elevation, mean temperature, and differential land endowments, among others. Subtracting equation 1 from equation 2 therefore eliminates these controls and yields the following basic long-difference model:

$$\Delta ELF_i = \zeta + \gamma_1 \Delta C_i + \gamma_2 \Delta M_i + \Delta \varepsilon_i \quad (3)$$

where for country *i* ΔELF_i is change in ELF between 1961 and 1985, ΔC_i is change in carbon emissions per capita and ΔM_i is change in the percentage of migrants.

Of course, when using a relatively small sample of observations it is important to check for outliers which may be driving any results. Therefore I used the DFbeta outlier detector tool on regression 3, which yielded three clear outliers, namely Kuwait, Qatar and the United Arab Emirates.[3] In all three cases local labor shortages led governments to recruit large numbers of workers from other parts of the Middle East and South Asia to work in their oil fields, which saw immigrants becoming the majority of the population between 1960 and 1975. Moreover, population data in these countries is notoriously questionable, as noted in an earlier attempt to collect cross-national ELF data by Fearon (2003, 219). As a result I exclude all three countries from my analysis here. For a graph of the relationship between change in ELF (on the y-axis) and change in carbon emissions (on the x-axis) between 1961 and 1985 with 95% confidence intervals see Figure 1, and for descriptive statistics and more complete definitions of the variables see Appendices 1 and 2, respectively.[4]

[Insert Figure 1 here]

My initial set of results is presented in Table 1. In column 1 I regress change in ELF on change in carbon emissions while controlling for change in international migration. In column 2 I control for initial levels of both ELF and carbon emissions to account for the possibility that the relationship between change in ELF and carbon emissions is spurious due to the effects of initial levels on subsequent change for both variables.

[Insert Table 1 here]

In columns 3-8 I add six additional variables in six separate specifications due to missing data and potential problems of multicollinearity that result from including too many explanatory variables in one specification.[5] In column 3 I control for change in the log of GDP per capita from

1961 to 1985, inasmuch as it is possible that it is not actually industrialization that is driving ethnic homogenization but instead an increase in wealth, which may come from sources other than industrial production. In column 4 I use the Polity IV database to control for change in democracy over the period, inasmuch as it is plausible that, as countries democratize, ethnic minorities will suffer less discrimination and ethnic conflict will decline (Wimmer 2015, 35).

I next test the argument from both Marx and Gellner, as well as more recently from Malesevic (2013), Mann (2013), Weber (1976) and Wimmer (2015), that ethnic homogenization is driven by the ruling elite via state policies. There are two ways to operationalize the effects of states on ethnic identity change. The first draws explicitly on Gellner's emphasis on schooling as a means to homogenize populations. In particular it is possible that decreases in ELF are a consequence of increased fluency in state languages and/or education policies promoting assimilation among ethnic minorities. Thus in column 5 I control for change in the mean number of years of primary and secondary school attended, which is calculated for the majority of countries in my sample.

Another way to operationalize the effect of state policies on ethnic identity is to consider the role of state capacity, whereby stronger states are able to enforce or incentivize ethnic homogenization while weaker states are unable to do so (Wimmer 2015). In particular this emphasis not just on the existence of a modern state but the potential variation in its power over its citizens is most obvious in Anderson (1991)'s description of "official nationalism." The classic measure of state capacity is taxation, inasmuch as tax collection involves an efficient bureaucracy that is able to extract taxes from a large number of citizens. Thus in column 6 I control for change in government revenue as a percentage of GDP (with similar results if I instead compute change in government expenditure per capita.)

Another measure of state capacity is infant mortality, which has been shown before to be robustly positively correlated with state failure (King and Zeng, 2001). Infant mortality is also an excellent measure of the quality of public goods provision, inasmuch as it is an unambiguously

undesirable phenomenon which all states have the potential to reduce as low as possible, its measurement is not particularly controversial relative to other measures of public goods provision, and reductions in infant mortality generally come about through an increase in the quality and quantity of public goods provision rather than merely through higher levels of income (Abouharb and Kimball 2007). As such in column 7 I control for change in the log of infant mortality rates.[6]

I next control for Anderson (1991)'s argument that the creation of modern nations is the result of print capitalism, by adding a control for change in newspaper circulation per capita in column 8 (with the same results but a much smaller sample size for change in book production per capita.) Finally, in column 9 I control for changes in total fertility rate, crude mortality rate, median age and sex ratio to account for the possibility that demographic changes are causing changes in ethnic identification across time.

As seen in Table 1, the results are quite striking: not only is the coefficient for change in carbon emissions always statistically significant at the 1% level, despite notably reduced sample sizes in columns 3-8, but the coefficient on the migration variable varies widely in magnitude and is statistically significant in only half of the regressions, and the coefficients on the other variables are not statistically significant. Indeed, despite very different sample sizes the carbon emissions variable coefficient does not vary widely around its mean of -0.03. Using this average coefficient, an increase in one standard deviation of log of carbon emissions in metric tons per capita would lead to a 0.02 decline in ELF over the period in question. While these numbers appear small, a 0.02 decrease in ELF is equivalent to a population split evenly between two ethnic groups moving to one split 60% in favour of the larger group and 40% for the smaller one. Moreover, there is considerable heterogeneity in the change in ELF variable within the dataset: for instance, countries like Libya and Oman were able to erase 33% of the difference in ELF between them and Ireland over this short 24-year period. Indeed, it is important to remember that the aforementioned recent theories of the origins of ethnic diversity from Ahlerup and Olson (2012) and Michalopoulos (2012)

propose mechanisms that work over hundreds, thousands and even tens of thousands of years, so any notable amount of change in ELF over a quarter of a century is an impressive feat.

To test for robustness I examined a variety of different-sized samples in Appendix 2, including sub-samples excluding Africa, excluding all but African countries, and excluding the Americas, Asia and Europem as well as former British colonies as well as all countries where the ANM only lists one ethnic group and states with populations less than 500,000 in 1960. In none of these regressions does the carbon emissions coefficient drop below the 1% level of significance.

4. Additional Results

A next step is to examine what process is driving the decline in ELF across countries. The most obvious mechanism is that the largest ethnic group in each country is growing in size as members of ethnic minorities increasingly identify with it over time, a process which corresponds to the Marxian/Gellnerian framework discussed above. I use the ANM dataset to measure the change in the largest ethnic group per country as my new dependent variable, and rerun all of the same specifications from Table 1, as presented in Appendix 3. (I also reran all of the specifications from Appendix 2, with identical results which I have not included here for lack of space.) As seen, the results are very consistent: the coefficient on the carbon emissions variable is now positive and statistically significant at the 5% level in all but one specification.

A further step is to employ alternative measures of industrialization, which is useful not only as a robustness test but also to provide evidence for the operative mechanism, namely the declining value of land. I employ three such alternatives here, namely cement production, urbanization and agriculture as a percentage of GDP. My first alternative measure of industrialization is cement production, which contributes around 5% of total carbon emissions globally and 2% of global energy consumption (Worrell et al. 2001). Cement is the main ingredient in concrete, which is the most widely used product in the modern world by weight after water, and is thus a good measure of

industrialization. Moreover, unlike other industrial products such as steel and electricity, cement production is mostly consumed domestically due to relatively high transport costs and wide availability of raw materials, and is thus a relatively good proxy for cement consumption as well. Finally, like carbon emissions it exhibits a S-shaped curve as countries at high levels of GDP/capita stabilize their production (Mahasenan et al. 2002).

The second alternative measure is urbanization, or the increase in the percentage of people living in urban areas. Urbanization is classically seen as an outcome of rural-urban migration as a product of the creation of an urban industrial sector, as seen in the stylized models by Marx and Gellner presented above, as well as the creation of cheap food for urban workers as a result of industrialized agriculture. Historically there exists a large amount of evidence tying together urbanization and the rise of a non-agricultural industrial or service-based economy (Gollin et al. 2016; Michaels et al. 2012). Moreover, urbanization captures the operative mechanism quite well, inasmuch as it measures the structural shift from a population based in rural areas to one more centered in urban centers. Finally, urbanization levels exhibit an S-shaped curve inasmuch countries stabilize at high levels, often considerably lower than 100%.

Third and finally, I use data on agriculture as a percentage of GDP, which captures the sectoral change that is part and parcel of industrialization. As with urbanization, this variable captures the essence of the operative mechanism as societies become decreasingly reliant upon rural agriculture for their economic output. Here the data clearly exhibits an inverted S-shaped curve (or rather a Z-shaped curve), with many countries stabilizing at a certain level for a long period of time. However, as noted above, the data is only available from the World Bank for 47 countries back to 1961. As such I have backward projected the data from 1970, which yields an additional 19 countries, with the caveat that the data quality is questionable.

As such I reran regression #3 for all three alternative measures of industrialization, with results listed in Table 2. In all three regressions the coefficients have the right sign and are statistically significant at the 5% level.

[Insert Table 2 here]

There is one caveat I should note before concluding. As with any OLS regression there remains concerns about issues of endogeneity whereby some unknown variable could determine change in both ethnicity and industrialization. Yet, as noted above, the causes of ethnic fractionalization suggested in the literature are all time-invariant and thus cancel out by the use of a long-difference regression. Ideally it would be useful to demonstrate that the relationship between industrialization and ethnic homogenization is not just correlative but also causal by instrumenting for carbon emissions or another proxy for industrialization. Yet doing so would require not only finding a variable that is correlated with changes in carbon emissions between 1961 and 1985 but also demonstrating that this variable only affected ethnic identity via changes in carbon emissions. Given the voluminous and disputed literature on why some countries industrialize and others do not, not surprisingly it has proven impossible to find such a variable, and thus I do not attempt to claim causality here.

5. Conclusion

In this paper I have argued that industrialization creates incentives for ethnic homogenization. Based upon a long tradition of social theory and history reaching back to Marx and Gellner, among others, I argued that industrialization lowers ethnic diversity by raising the relative value of labor over land. I then used Soviet data to show that rising levels of carbon emissions per capita was robustly correlated with declining levels of ethnic fractionalization across time. Moreover, consistent with Marx and Gellner's theories, I showed that industrialization is also correlated with an increase in the percentage of people identifying with the largest ethnic group per country. Finally, I found that change in ethnic fractionalization is negatively correlated with

alternative proxies of industrialization such as change in cement production per capita, urbanization and agriculture as a percentage of GDP, which not only served as robustness checks for my analysis but also provided evidence for the relative decline in the value of land as the operative mechanism in a "bottom-up" manner.

There are at least three broader conclusions that I can make from this study. First and most obviously, on an empirical level future research on ethnicity that uses panel data or times series analyses cannot continue to assume that ethnicity is fixed and unchanging, especially in regression models that include measures of industrialization as a covariate. If my analysis is correct then the relationship between ethnic diversity and outcomes such as low economic growth, civil wars and poor public goods provision might merely be correlational rather than causational.

Second, there is little evidence here supporting the role of the state in promoting ethnic identity change. This finding thus fails to support other research which suggests that states have played a major role in the formation of contemporary ethnic and national identities (Gellner 2006 [1983]; Malesevic 2013; Mann 2013; Weber 1976; Wimmer 2015). There are four potential ways to account for these contradictory findings. First, it is possible that, with the accumulation of better data, the collective evidence on modern identity formation is beginning to swing towards the role of economic factors and incentives over political ones, both in the contemporary world and more historically (Robinson 2014; Nix and Qian 2015). Second, it is possible that states were effective at promoting ethnic homogenization in the pre-modern world but have been less effective in the modern world, which would account for Wimmer's findings on the correlation between state history and ethnic fractionalization as well as additional evidence on the inability of modern states to follow pre-modern states in imposing ethnic/linguistic homogeneity on their citizens (Bandyopadhyay and Green 2013; Laitin 2007; Wimmer 2015). Third, it is possible that the mechanism by which modern states have been able to promote ethnic homogenization is actually via industrialization, which is more aligned with aforementioned findings of Gellner and Eugen

Weber. Fourth, it is possible that state-promoted identity change works on a longer time scale than the 24-year period under examination here.

Third and finally, future research may wish to address in more detail the implications of my analysis. For instance, if industrialization contributes towards ethnic change, then the intra-ethnic reciprocity norms discussed by Habyarimana et al. (2009) may be weaker in more industrialized countries where many co-ethnics have recently assimilated into new identities. As regards policy debates, if industrialization contributes to lower levels of ethnic diversity then governments should encourage more industrialization if they want to bypass the negative effects of ethnic diversity on economic and political development. This is a markedly different policy prescription than previous attempts to manage the negative effects of ethnic diversity: for instance, Laitin (2007, 112) argues that governments should instead create sub-national ethnically distinct or homogenous jurisdictions. Moreover, it is one that is at odds with current development discourse, which has moved away from industrialization and modernization in favor of poverty reduction and human development (Chang 2013). As with other conclusions, however, this proposal remains a topic for further investigation.

Acknowledgements

I would like to thank Catherine Boone, Sean Fox, Michael Hechter, Eric Kaufman, Matthias vom Hau, Steffen Hertog, David Laitin and seminar participants at Cornell University, the LSE, Oxford University, Princeton University, the University of Pennsylvania and the Annual Meeting of the American Political Science Association for comments and suggestions. I would also like to thank Cecilia Lanata-Briones and Ulas Karakoc for research assistance and Karina Perez Jvostova, Zhanna Kovaleva, Mélanie Loubet and Alexander Zuev for help with Russian-English translations. All errors remain my own.

Bibliography

Abouharb, M. Rodwan, and Anessa L. Kimball. 2007. "A New Dataset on Infant Mortality Rates, 1816-2002." *Journal of Peace Research*, 44 (6): 743-754.

Ahlerup, Pelle, and Ola Olsson. 2012. "The Roots of Ethnic Diversity." *Journal of Economic Growth*, 17 (2): 71-102.

Anderson, Benedict. 1991. *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. London: Verso.

Ashraf, Quamral, and Oded Galor. 2013. "Genetic Diversity and the Origins of Cultural Fragmentation." Mimeo, Department of Economics, Brown University.

Bandyopadhyay, Sanghamitra, and Elliott D. Green. 2013. "Nation-Building and Conflict in Modern Africa." *World Development*, 45: 108-118.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Bentley, G. Carter. 1987. "Ethnicity and Practice." *Comparative Studies in Society and History*, 29 (1): 24-55.

Bleaney, Michael, and Arcangelo Dimico. 2016. "State History, Historical Legitimacy and Modern Ethnic Diversity." *European Journal of Political Economy*, 43: 159-170.

Brubaker, Rogers. 2004. *Ethnicity Without Groups*. Cambridge, MA: Harvard University Press.

Bruk, Solomon I. 1986. *Naselenie Mira Etnodemograficheskii Spravochnik* Moscow: Izd-vo "Nauka".

Bruk, Solomon I., and V. S. Apenchenko. 1964. *Atlas Narodov Mira*. Moscow: Glavnoe upravlenie geodezii I Kartografii Gosudarstvennogo geologicheskogo komiteta SSSR, Institut etnografii im. N. H. Miklukho-Maklaia Akademii nauk SSSR.

Cassan, Guilhem. 2015. "Identity Based Policies and Identity Manipulation: Evidence from Colonial Punjab." *American Economic Journal: Economic Policy*, 7 (4): 103-131.

Cederman, Lars-Erik, and Luc Girardin. 2007. “Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies.” *American Political Science Review*, 101 (1): 173-186.

Chandra, Kanchan. 2006. “What is Ethnic Identity and Does it Matter?” *Annual Review of Political Science*, 9: 397-424.

Chang, Ha-Joon. 2013. “Hamlet without the Prince of Denmark: How Development has Disappeared from Today’s ‘Development’ Discourse.” In *Global Governance at Risk*, edited by David Held and Charles Roger, 129-148. Cambridge: Polity Press.

Cohen, Ronald. 1978. “Ethnicity: Problem and Focus in Anthropology.” *Annual Review of Anthropology*, 7: 379-403.

Esteban, Joan, and Debraj Ray. 2008. “On the Salience of Ethnic Conflict.” *American Economic Review*, 98 (5): 2185-2202.

Fearon, James D. 2003. “Ethnic and Cultural Diversity by Country.” *Journal of Economic Growth*, 8 (2): 195-222.

Gellner, Ernest. 2006 [1983]. *Nations and Nationalism*. Oxford: Blackwell..

Gollin, Douglas, Remi Jedwab, and Dietrich Vollrath. 2016. “Urbanization with and without Industrialization.” *Journal of Economic Growth* 21 (1): 35-70.

Habyarimana, James, Macartan Humphreys, Daniel Posner, and Jeremy Weinstein. 2009. *Coethnicity: Diversity and the Dilemmas of Collective Action*. New York: Russell Sage Foundation.

King, Gary, and Langche Zeng. 2001. “Improving Forecasts of State Failure.” *World Politics*, 53 (4): 623-658.

Laitin, David. 1998. *Identity in Formation: The Russian-Speaking Populations in the Near Abroad.* Ithaca, NY: Cornell University Press.

Laitin, David. 2007. *Nations, States and Violence*. Oxford: Oxford University Press.

Mahasenan, Natesan, Steve Smith, and Kenneth Humphreys. 2002. *The Cement Industry and Global Climate Change: Current and Potential Future Cement Industry CO2 Emissions*. Richland, WA: Pacific Northwest National Laboratory.

Malešević, Siniša. 2013. *Nation-States and Nationalisms.* Cambridge: Polity Press.

Mann, Michael. 2013. *The Sources of Social Power. Volume IV: Globalizations, 1945-2011*. Cambridge: Cambridge University Press.

Marx, Karl, and Friedrich Engels. 1978 [1848], "Manifesto of the Communist Party." In *The Marx-Engels Reader,* edited by Robert Tucker, 594-618. New York: WW Norton.

Michaels, Guy, Ferdinand Rauch, and Stephen J. Redding. 2012. "Urbanization and Structural Transformation." *Quarterly Journal of Economics*, 127 (2): 535-586.

Michalopoulos, Stelios. 2012. "The Origins of Ethnolinguistic Diversity: Theory and Evidence." *American Economic Review*, 102 (4): 1508-1539.

Morning, Ann. 2008. "Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round." *Population Research and Policy Review*, 27 (2): 239-272.

Nix, Emily, and Nancy Qian. 2015. *The Fluidity of Race: Racial Passing in the United States, 1880-1940*. NBER Working Paper #20828.

Posner, Daniel. 2005. *Institutions and Ethnic Politics in Africa*. New York: Cambridge University Press.

Robinson, Amanda L. 2014. "National Versus Ethnic Identification in Africa: Modernization, Colonial Legacy, and the Origins of Territorial Nationalism." *World Politics*, 66 (4): 709-746.

Roeder, Philip. 2001. Ethnolinguistic Fractionalization (ELF) Indices, 1961 and 1985. Retrieved from http://weber.ucsd.edu/~proeder/elf.htm.

Weber, Eugen J. 1976. *Peasants into Frenchmen: The Modernization of Rural France, 1870-1914*. Stanford, CA: Stanford University Press.

Wimmer, Andreas. 2015. "Nation Building: A Long-Term Perspective and Global Analysis." *European Sociological Review*, 31 (1): 30-47.

Wimmer, Andreas, Cederman, Lars-Erik, and Brian Min. 2009. “Ethnic Politics and Armed Conflict. A Configurational Analysis of a New Global Dataset.” *American Sociological Review*, 74 (2): 316-337.

Worrell, Ernst, Lynn Price, Nathan Martin, Chris Hendriks, and Leticia Ozawa Meida. 2001. “Carbon Dioxide Emissions from the Global Cement Industry.” *Annual Review of Energy and the Environment*, 26: 303-329.

Table 1: Industrialization and Ethnic Change
(Dependent Variable: Change in ELF, 1961-1985)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ CO_2 emissions per capita, 1960-1985	-0.033***	-0.030***	-0.032***	-0.026***	-0.033***	-0.037***	-0.034***	-0.032***	-0.030***
	(0.009)	(0.008)	(0.008)	(0.010)	(0.010)	(0.009)	(0.008)	(0.011)	(0.008)
Δ Immigrants as a % of the Population, 1960-1985	0.517**	0.568***	0.336	0.404	0.481**	0.417	0.308**	0.562*	0.465**
	(0.204)	(0.195)	(0.269)	(0.271)	(0.242)	(0.256)	(0.156)	(0.011)	(0.206)
ELF, 1961		-0.001							
		(0.017)							
CO_2 emissions, 1961		0.005							
		(0.003)							
Immigrants as a % of the Population, 1960		0.142							
		(0.092)							
Δ Log GDP per capita, 1961-1985			0.013						
			(0.009)						
Δ Polity IV democracy score, 1961-1985				0.007					
				(0.007)					
Δ Mean Years of Schooling, 1960-1985					-0.004				
					(0.005)				
Δ Government Revenue as a % of GDP, 1961-1985						0.101			
						(0.122)			
Δ Log Infant Mortality, 1961-1985							-0.006		
							(0.010)		
Δ Newspaper Circulation per capita, 1961-1985								-0.002	
								(0.011)	
Δ Crude Mortality Rate, 1960-1985									0.233
									(0.142)
Δ Total Fertility Rate, 1960-1985									0.003
									(0.005)
Δ Median Age, 1960-1985									0.003
									(0.002)
Δ Sex Ratio, 1960-1985									0.090

									(0.156)
Constant	0.026*** (0.009)	0.018* (0.012)	0.022*** (0.009)	0.019** (0.008)	0.033** (0.013)	0.025* (0.013)	0.015 (0.011)	0.025** (0.010)	0.039*** (0.012)
N	133	133	95	96	110	85	88	92	131
R^2	0.248	0.322	0.236	0.176	0.252	0.280	0.233	0.252	0.286

* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$; robust standard errors in parentheses.

Table 2: Industrialization and Ethnic Change, with Alternative Proxies for Industrialization
(Dependent Variable: Change in % ELF, 1961-1985)

	(1)	(2)	(3)
Δ in Log of Cement Production per capita, 1961-1985	-0.013***		
	(0.005)		
Δ in Urbanization, 1960-1985		-0.176**	
		(0.072)	
Δ in Agriculture as a % of GDP, 1961-1985			0.087**
			(0.042)
Constant	0.011	0.017**	-0.004
	(0.007)	(0.010)	(0.007)
N	109	151	66
R^2	0.209	0.127	0.231

* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$; robust standard errors in parentheses. Change in Migrant Stock is included as a control variable in all regressions but is not shown here.

[1] The dataset does not include Czechoslovakia, the USSR and Yugoslavia inasmuch as data on both carbon emissions and international migration was only available for their various successor states.

[2] Using World Bank data, the number of country observations for these variables dating back to 1961 are 34 and 47, respectively. The same problem of missing country-year observations afflicts a variety of other data as well.

[3] The Dfbeta tool calculates the difference in the regression coefficient for a particular variable with and without each individual observation. The rule of thumb is to exclude outliers that yield a DFbeta value greater than |1|; in this case Kuwait, Qatar and the UAE all had DFbeta values of |1.3| or greater for either change in carbon emissions or change in migrant stock, with no other observation above |0.6|. Cf. Belsley et al. (1980), 28..

[4] It is immediately clear from Figure 1 that Saudi Arabia is far away from the trend line. If it is excluded then the coefficient on the carbon emissions is considerably larger.

[5] The problem of missing data is particularly egregious in the case of African states, which are often missing data on such measures as revenue as a percentage of GDP and infant mortality.

[6] The same results hold when using changes in road density and doctors per capita as additional proxies for state capacity.

Appendix 1: Descriptive Statistics

	Number	*Mean*	*St. Dev*	*Minimum*	*Maximum*
ELF, 1961	133	0.438	0.280	0.003	0.909
Δ ELF, 1961-1985	133	-0.0005	0.063	-0.147	0.269
Log of Carbon Emissions per capita, 1961	133	-0.634	1.829	-4.761	3.595
Δ Carbon Emissions, 1961-1985	133	0.824	0.676	-1.106	2.611
Δ Immigrants (% of the Population, 1960-85)	133	0.002	0.048	-0.190	0.249
Δ Largest Ethnic Group 1960-85	133	-0.004	0.052	-0.179	0.172

Appendix 2: Industrialization and Ethnic Change, Sub-Samples
(Dependent Variable: Change in ELF, 1961-1985)

	Excluding SS Africa	Only SS Africa	Excluding Americas	Excluding Asia	Excluding Europe	Excluding Ex-British Colonies	Excluding Mono-Ethnic Countries	Excluding Small States
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ CO_2 emissions per capita, 1960-1985	-0.038***	-0.021***	-0.031***	-0.045***	-0.025***	-0.038***	-0.036***	-0.022***
	(0.012)	(0.007)	(0.009)	(0.009)	(0.008)	(0.010)	(0.009)	(0.008)
Δ Immigrants as a % of the Population, 1960-1985	0.579**	0.028	0.481**	0.832***	0.353***	0.697***	0.581***	0.365
	(0.228)	(0.217)	(0.219)	(0.210)	(0.213)	(0.267)	(0.223)	(0.242)
Constant	0.033**	0.002	0.021	0.033***	0.016*	0.031***	0.025***	0.016
	(0.012)	(0.006)	(0.010)	(0.008)	(0.008)	(0.010)	(0.009)	(0.006)
N	95	38	105	104	109	95	114	109
R^2	0.280	0.156	0.257	0.394	0.138	0.350	0.295	0.159

* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$; robust standard errors in parentheses. Small states are defined here as states with a population lower than 500,000 in 1960.

Appendix 3: Industrialization and Ethnic Change for the Largest Ethnic Group

(Dependent Variable: Change in % of the largest ethnic group, 1961-1985)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ CO_2 emissions per capita, 1960-1985	0.024***	0.025***	0.020***	0.015**	0.022***	0.024***	0.020***	0.017*	0.020***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.009)	(0.007)	(0.009)	(0.006)
Δ Immigrants as a % of the Population, 1960-1985	-0.295**	-0.371***	-0.202	-0.128	-0.218*	-0.238	-0.324**	-0.201	-0.297**
	(0.119)	(0.122)	(0.181)	(0.112)	(0.126)	(0.232)	(0.124)	(0.155)	(0.140)
Largest Ethnic Group %, 1961		0.002							
		(0.019)							
CO_2 emissions, 1961		-0.0005							
		(0.003)							
Immigrants as a % of the Population, 1960		-0.119							
		(0.067)							
Δ Log GDP per capita, 1961-1985			0.0008						
			(0.010)						
Δ Polity IV democracy score, 1961-1985				-0.001					
				(0.007)					
Δ Mean Years of Schooling, 1960-1985					0.008*				
					(0.005)				
Δ Government Revenue as a % of GDP, 1961-1985						-0.094			
						(0.090)			
Δ Log Infant Mortality, 1961-1985							-0.004		
							(0.010)		
Δ Newspaper Circulation per capita, 1961-1985								0.008	
								(0.008)	
Δ Crude Mortality Rate, 1960-1985									-0.163
									(0.109)
Δ Total Fertility Rate, 1960-1985									-0.004
									(0.004)
Δ Median Age, 1960-1985									-0.0004
									(0.002)
Δ Sex Ratio, 1960-1985									0.041

									(0.137)
Constant	-0.022*** (0.007)	-0.017 (0.016)	-0.021** (0.009)	-0.015** (0.008)	-0.037*** (0.011)	-0.017 (0.011)	-0.017 (0.012)	-0.016* (0.009)	-0.034*** (0.009)
N	133	133	95	96	110	85	88	92	131
R^2	0.149	0.194	0.121	0.051	0.157	0.154	0.134	0.089	0.162

* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$; robust standard errors in parentheses.

Appendix 4: Data Sources

Dependent Variables

Δ in ELF: Change in ELF between 1961 and 1985 as measured by the ANM.

Δ in Largest Ethnic Group: Change in the size of the largest ethnic group between 1961 and 1985 as measured by the ANM.

Independent Variables

Δ CO_2 *emissions 1961-1985*: Change in log of CO_2 emissions in metric tons per capita, 1961-1985. Source: : World Bank World Development Indicators, with data originally taken from the Carbon Dioxide Information Analysis Center at Oak Ridge National Laboratory (run by the US Department of Energy).

CO_2 *emissions, 1961*: Log of CO_2 emissions in metric tons per capita in 1961. Source: World Bank World Development Indicators, with data originally taken from the Carbon Dioxide Information Analysis Center at Oak Ridge National Laboratory (run by the US Department of Energy).

Δ Agriculture as a % of GDP, 1961-1985: Change in agriculture as a percentage of GDP, 1961-1985. Source: World Bank World Development Indicators.

Δ Cement production per capita, 1961-1985: Change in log of cement production in metric tons per capita, 1961-1985. Source: Cross National Times Series Data Archive. In some cases the 1961 data was backward projected from 1962 (Algeria, Burundi, Jamaica, Rwanda, Samoa, Uganda), 1963 (Kenya), 1964 (Malawi, Malta), 1965 (Gambia, Maldives, Singapore, Zambia), and 1966 (Botswana, Guyana, Lesotho)**.** For countries recorded as having no cement production in 1961 but some in 1985 I assigned a value of 0.001 tons per capita in 1961 in order to avoid taking the log of zero.

Δ Crude Mortality Rate, 1960-1985: Change in crude death rates per country, 1960-1985. Source: United Nations Population Division, World Population Prospects.

ELF, 1961: ELF per country as measured in 1961. Source: ANM.

Δ GDP/capita Change 1961-1985: Change in levels of GDP/capita per country, 1961-1985. Source: Alan Heston, Robert Summers, and Bettina Aten. 2009. "Penn World Table, Version 6.3, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Δ Immigrants (% of the Population, 1960-1985): Change in the ratio of total migrants to the total population per country, 1960-1985. Source: United Nations Population Division, World Population Prospects.

Immigrants as a % of the Population, 1960: Ratio of total migrants to the total population per country in 1960. Source: United Nations Population Division, World Population Prospects.

Δ Log Infant Mortality, 1960-1985: Change in the log of infant mortality rates, 1960-1985. Source: United Nations Population Division, World Population Prospects.

Δ Mean Years of Schooling 1960-1985: Change in the mean years of primary and secondary schooling per country, 1960-1985. Source: Barro, R., & Lee, J.-W. (2010). A New Data Set on Educational Attainment in the World, 1950-2010. NBER Working Paper #15902, National Bureau of Economic Research.

Δ Median Age, 1960-1985: Change in median age per country, 1960-1985. Source: United Nations Population Division, World Population Prospects.

Δ Newspaper Circulation per capita, 1961-1985: Change in daily newspaper circulation per capita per country, 1961-1985. Source: Cross National Times Series Data Archive.

Δ Polity IV, 1961-1985: Change in a country's democracy rating, 1961-1985. Source: Polity IV.

Δ Sex Ratio, 1960-1985: Change in sex ratio (total number of women divided by total number of men) per country, 1960-1985. Source: United Nations Population Division, World Population Prospects.

Δ Government Revenue as a % of GDP, 1960-1985: Change in government revenue as a percentage of GDP per country, 1960-1985. Source: Cross National Times Series Data Archive.

Δ Total Fertility Rate, 1960-1985: Change in total fertility rates per country, 1960-1985. Source: United Nations Population Division, World Population Prospects.

Δ Urbanization, 1960-1985: Change in the percentage of a country's population living in urban areas, 1960-1985. Source: United Nations, "World Urbanization Prospects: The 2012 Revision," Department of Economic and Social Affairs/Population Division (2013).

Appendix 5: Assessing the Accuracy of the Soviet ELF Data

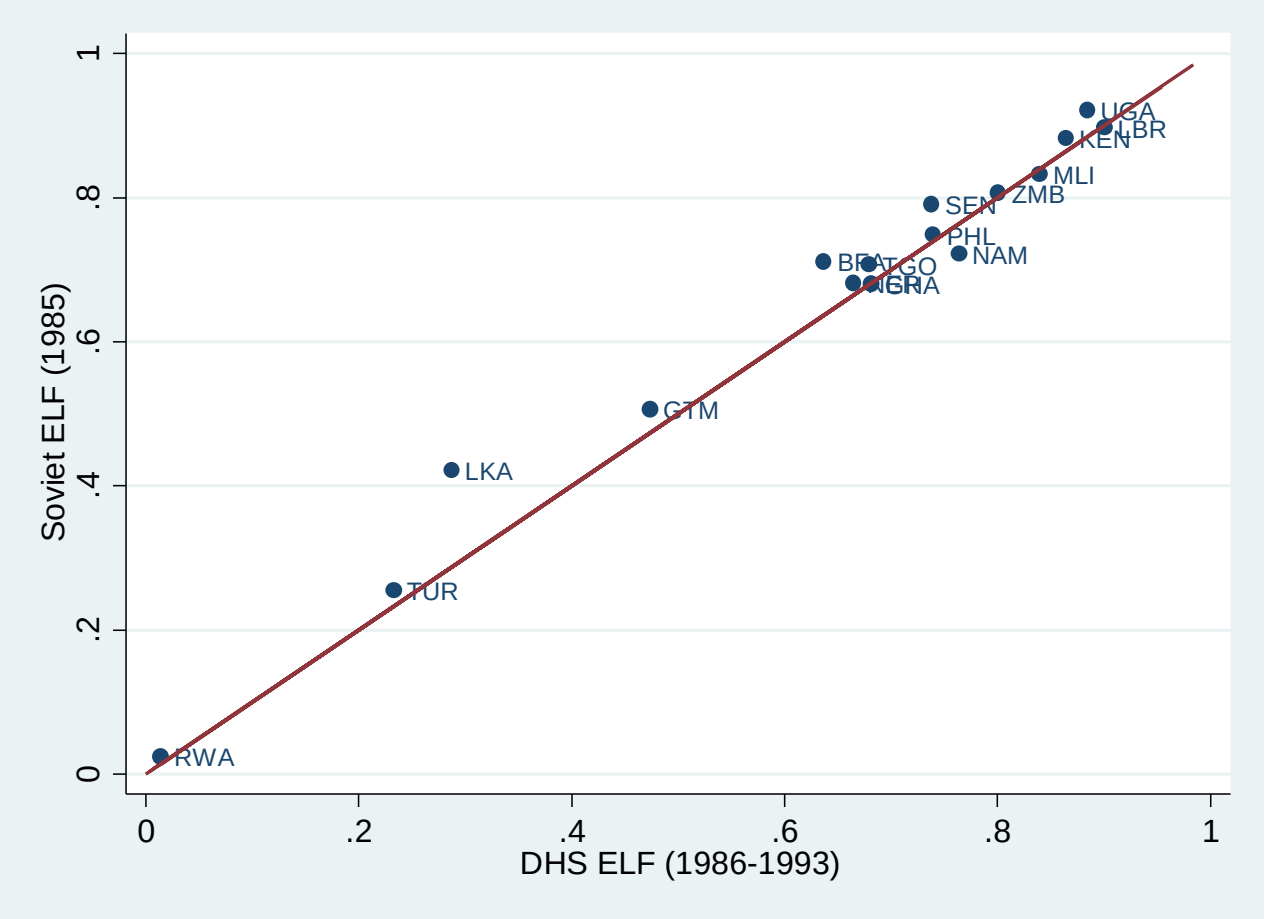
The Soviet data on ELF is given in two volumes. The first volume is an atlas, with ethnographic maps in the beginning followed by data on the ethnic composition at the end. The second volume does not have any maps and does not list ethnic groups per country in tabular form for the same number of countries that were listed in the first volume, but instead has textual descriptions for 72 out of 187 countries. With the help of Russian-speaking research assistants I coded ethnic data for all countries in the dataset across the two years in question.

As a means of checking the accuracy of the Soviet data I compared its ELF scores to ELF scores generated from the first eight years of data from demographic and health surveys (DHS) conducted across a number of developing countries between 1986 and 1993. The DHS surveys provide a good independent data source, both because they were conducted just after the Soviet data was published (and were thus not a source for the Soviet data) and because they provide primary evidence of ethnic identity in a non-census format, thereby bypassing at least some of the problems associated with census data noted above.

In most cases data on ethnic identity was not collected, but in the sixteen countries listed below in Table A2.1 we can compile an ELF score based on the DHS survey data collected between 1986 and 1993 (which varied in number from 3199 people in Mali to 7150 in Kenya). The DHS researchers mostly collected data on the same lines as the Soviet data, with three exceptions. In Guatemala the DHS data only listed two ethnic groups, namely indigenous (Indian) and Ladino, while the Soviet data was more detailed, which led me to collapse the Soviet data into two comparable groups. In the Philippines the Soviet data listed the Bisaya or Visayan group, while the DHS listed the three major sub-groups of the Bisaya (Cebuano, Ilonggo and Waray), which led me to collapse the DHS data. Finally, in Ghana data was tabulated at the level of the sub-ethnic group for the Akan peoples in the Soviet data but was only tabulated at the level of the ethnic group in the DHS data, therefore leading me to count the Akan as one ethnic group in both cases. In four cases, namely Ghana, Guatemala, Kenya and Senegal, DHS data on ethnic identity was collected for both rounds; as such I take the mean values across the two rounds.

As can be seen in Figure A1, the correlation between the Soviet and DHS data is very high: when regressing the Soviet data on the DHS data the DHS variable has a coefficient of 0.946 and the R2 is 0.978. Sri Lanka is the only outlier in the relationship, which is exactly as we would expect since the DHS did not survey Northern and Eastern Provinces due to the ongoing civil war at the time, which meant that it oversampled Sinhalese and undersampled Tamils compared to the Soviet data.

Figure A5.1: Comparing Soviet and DHS I ELF data



We can observe the same close correlation by just examining the largest ethnic group in each country as well in Table A3.2, where we again observe a very high correlation (regression coefficient of 0.946 and an R2 of 0.958).

Figure A5.2: Largest Ethnic Group, Soviet and DHS I data

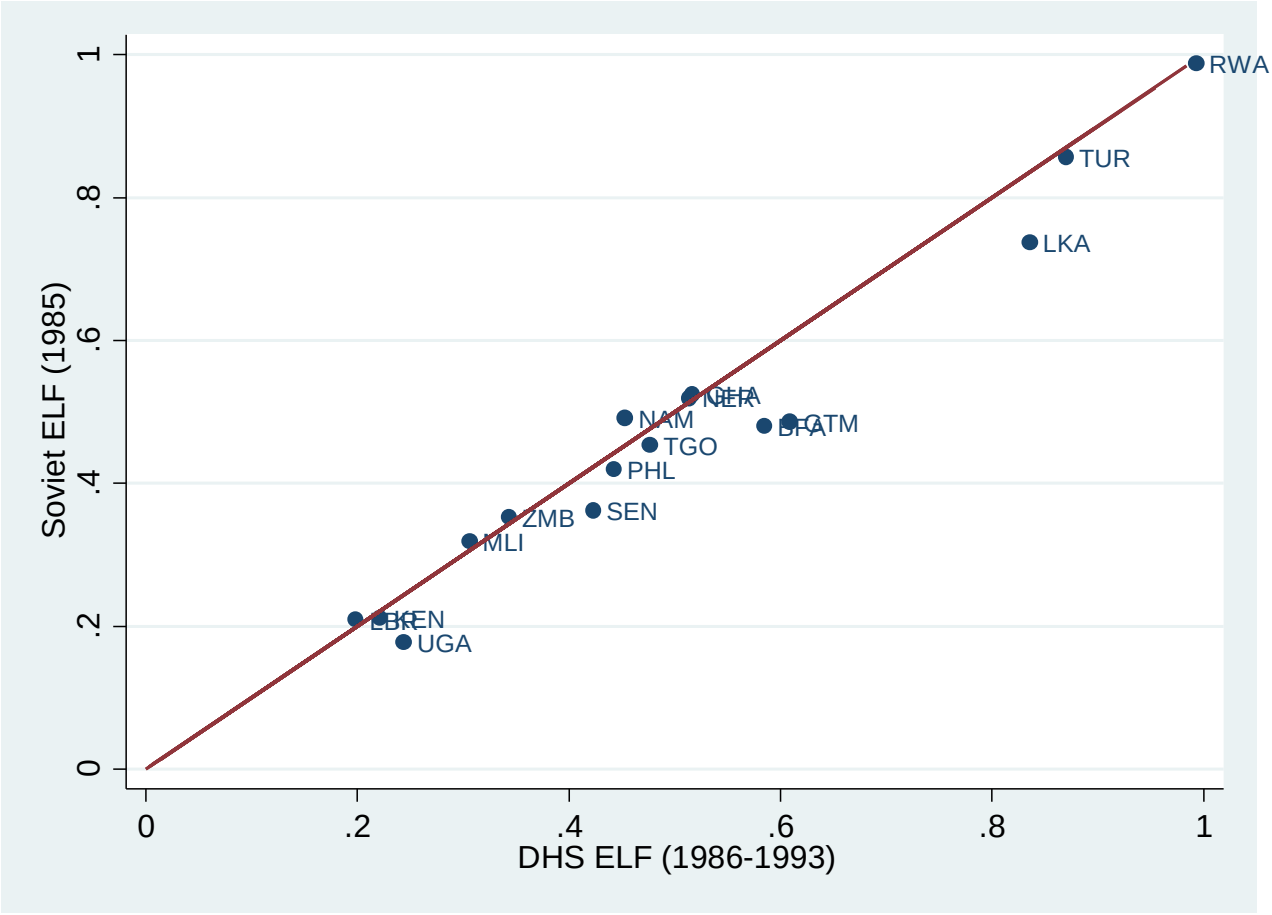


Table A5.1: Countries Included in DHS/ANM Comparison

Country	Year(s) collected
Burkina Faso	1992
Ghana	1988, 1993
Guatemala	1987, 1993
Kenya	1989, 1993
Liberia	1986
Mali	1987
Namibia	1992
Niger	1992
Philippines	1993
Rwanda	1992
Senegal	1986, 1992
Sri Lanka	1987
Togo	1988
Turkey	1993
Uganda	1995
Zambia	1988

Appendix 6: Data Alterations

The following countries were left out of the analysis for the following reasons:

- **Anguilla**, **Antigua and Barbuda**, **Barbados**, **Dominica**, **Grenada**, **Montserrat**, **Saint Kitts and Nevis**, and **Saint Lucia**, **Saint Vincent and the Grenadines**. Listed together as one unit in the 1961 but separately in the 1985 Atlas.
- **Bangladesh** and **Pakistan.** Listed together under Pakistan in the 1961 Atlas but separately in the 1985 Atlas due to the partition of Pakistan in 1971.
- **Cape Verde**. The 1961 Atlas lists groups according to language (where all Portuguese-speakers are listed as one group) while the 1985 Atlas lists group according to nationality (with native Cape Verdeans and Portuguese listed separately).
- **Central African Republic**. The 1985 Atlas lists several indigenous ethnic groups not included in the 1961 Atlas, including the Ngbandi (10.6% of the population), Sara (6.9%) and Mbum (4.0%), among others.
- **Gabon**. The 1985 Atlas lists several indigenous ethnic groups not included in the 1961 Atlas, including the Mbete (14.2% of the population) and the Ngundi (5.3%).
- **Kiribati** and **Tuvalu**. Listed together under the Gilbert and Ellice Islands in the 1961 Atlas but separately in the 1985 Atlas due to their separate independence in 1976.
- **Papua New Guinea**. Listed separately as New Guinea and Papua in the first volume as the two constituent units of what was then the Territory of Papua and New Guinea.
- **Réunion**. The 1961 Atlas lists groups according to language (where all French-speakers are listed as one group) while the 1985 Atlas lists group according to race (with Creoles and French listed separately).
- **Surinam**. The 1961 Atlas lists groups according to language (where all Dutch-speakers are listed as one group) while the 1985 Atlas lists group according to race (with Creoles, "Bush Negros" and Dutch listed separately).

In the following cases I coded ELF differently from Roeder (2001):

- **Algeria, Iraq, Lebanon, Libya, Morocco, Sudan**, **Syria** and **Tunisia**. I coded Arabs as one ethnic group rather than coding them by nationality.
- **Brunei**. I coded the Malays and Kedayan as one group in 1985 inasmuch as they were coded as one group in 1961 (and in every census in Brunei since 1960).
- **Liberia**. I coded the Grebo, Kran and Kru as one group in 1985 inasmuch as they were recorded as one group in 1961.
- **Mali**. I coded the Bambara, Duala and Mandinka as one group in 1985 inasmuch as they were recorded as one group in 1961.
- **Netherlands**. I coded the Flemish as Dutch in 1985 to match the fact that both were recorded as one group in 1961.
- **Switzerland**. I coded the French speakers, German speakers and Italian speakers as single ethnic groups in 1985 to match the fact that they were recorded as such in 1961.
- **Vietnam**. I added data for 1961 to the dataset inasmuch as it was listed as a single unit in 1961 despite the fact that it was partitioned between 1954 and 1975.

Finally, I added the following countries which were listed in both Atlases but were not included in (Roeder, 2001)'s original dataset: Belize, Bermuda, French Polynesia, Guam, Hong Kong, Macao, Solomon Islands, São Tomé and Príncipe, Seychelles, Timor-Leste, Tonga and the US Virgin Islands.