Ethnic Structure and Cultural Diversity around the World: A Cross-National Data Set on Ethnic Groups*

James D. Fearon
Department of Political Science
Stanford University
Stanford, CA 94305-6044

January 3, 2003

Abstract

Development economists have recently focused on ethnic diversity, or "fractionalization," as a possible cause of corruption, political instability, and poor economic performance. Political scientists have argued for years over possible links between ethnic diversity (or structure) and civil violence, democratic stability, and party systems. For its empirical evaluation in a cross-national setting, all such research requires data on ethnic groups across countries. This paper tries to do a better job of conceptually grounding, operationalizing, and constructing a list of ethnic groups across countries than is currently available. After addressing conceptual and practical problems involved in enumerating "ethnic groups," I present a list of 820 ethnic groups in 160 countries that made up at least 1% of country population in the early 1990s. I compare a measure of ethnic fractionalization based on this list with the commonly used fractionalization measure based on the *Atlas Narodov Mira* (1964). Finally, I construct an index of *cultural* fractionalization that uses the structural distance between languages as a proxy for the cultural distance between groups in a country. This latter measure may be more appropriate when testing hypotheses that assume that ethnic groups "matter" because they have diverse preferences and cultural differences that pose obstacles to cooperation.

1 Introduction

Does ethnic diversity lower a country's economic growth rate or level of public good provision, as Easterly and Levine (1997) and Alesina, Baqir and Easterly (1997) claim? Are more ethnically divided states more civil war prone? Less (or more) likely to experience democratic transitions or stable democracy? More likely to have highly fractionalized party systems if

*To be presented at the 2002 Annual Meeting of the American Political Science Association, August 29-Sept. 1, Boston. Some of the research for this paper was supported by grants from the National Science Foundation (SES-9876477 and SES-9876530) and the Carnegie Corporation. I am deeply indebted to excellent research assistance by Christina Maimone, Alexander Rosas, Atsuko Suga, and Nikolay Marinov. I also thank David Laitin for many helpful conversations about the project, which is linked to our larger joint project on civil conflict.

they are democratic? What factors differentiate ethnic groups that rebel or have secessionist movements from those that do not? What differentiates the ethnic groups whose members mobilize in the political sphere from those that remain "latent"?[1]

Empirical efforts to answer any of these questions require that we collect data on ethnic groups in different countries. And before any such data is collected, we need a list or sample of "ethnic groups" for some sample of countries.

This paper discusses conceptual and practical problems involved in constructing a cross-national list of ethnic groups (sections 2 and 3) and then presents the results of an effort to carry out the task (sections 4-8). Restricting attention to groups that had at least 1% of country population in the 1990s, I identify 820 ethnic and "ethnoreligious" groups in 160 countries. Hypothetically, my objective is to include those groups that would be listed most often if randomly chosen individuals in the country in question were asked "what are the main ethnic (or racial or ascriptive) groups in this country?" I lack the resources to carry out such a survey and have not done so. Faute de mieux, I rely on the secondary sources and existing lists discussed below. The list offered here should be viewed as a continual work in progress, to be improved as more country-specific expertise, or actual survey data, is brought to bear case by case.

Sections 5 presents descriptive statistics. Section 6 proposes a simple way to use the data to represent how *ethnic structures* differ across countries, as opposed to an aggregate measure of ethnic diversity. Section 7 compares the standard diversity measure constructed from my data with the commonly used measure based on the *Atlas Narodov Mira*, published

[1]On the relationship (or lack thereof) between ethnic divisions and civil conflict, see among others Hibbs (1973), Horowitz (1985), Powell (1982), VanHanen (1999), Fearon and Laitin (2002), Collier and Hoeffler (2001), and Huntington (1996). Przeworski, Alvarez, Cheibub and Limongi (2001) consider ethnic diversity as a possible predictor of the likelihood of democratic transitions. Cox (1997) considers the effect of ethnic diversity on party systems in democracies. Dudley and Miller (1998), Fearon and Laitin (1999), Gurr (1993), Gurr and Moore (1997), and Lindstrom and Moore (1995) examine the determinants of ethnic group rebellion and protest.

by Soviet ethnographers in 1964. In section 8, I use the data to construct an index of *cultural* fractionalization that uses a measure of the structural relationship between languages to take into account the cultural distance between groups in a country. For instance, Belarus and Cyprus have somewhat similar ethnic structures, but the groups in Belarus are culturally much closer than those in Cyprus. Using the structural distance between languages as a proxy for extent of cultural difference, the cultural fractionalization measure attempts to take such cultural proximity into account.

Relation to existing work. As noted, the best known and most widely used similar effort was carried out by a team of Soviet ethnographers in the early 1960s, and published as *Atlas Narodov Mira*. Their list of "ethnolinguistic" groups and population figures has been employed by several generations of political scientists, sociologists, and, more recently, economists to produce cross-national estimates of "ethnic fractionalization."[2] Most empirical studies concerning the implications of ethnic diversity, such as Easterly and Levine's work on economic growth, have employed this measure. The Soviet team mainly used language to define groups, but sometimes included groups that seem to be distinguished by some notion of race rather than language, and quite often use national origin (e.g., "Anglo-Canadians" are listed in the United States).

More recently, Ted R. Gurr and his collaborators have developed a list of "minorities at risk" in 115 countries, along with a remarkable array of variables coding group characteristics, situations, and experiences (Gurr 1996). This data set has allowed for the first time large-N research on the correlates of group oppression, protest, and rebellion. Unfortunately, the selection criteria for the sample – the groups must be judged "at risk" in one or more of four

[2]Hibbs's (1973) cross-national study of causes of political violence is an early example in political science. Taylor and Jodice's (1983) handbook included the version of the measure that is widely cited. See Easterly and Levine (1997) and Alesina, Devleeschauwer, Easterly, Kurlat and Wazciarg (2002) for references in economics.

ways – render problematic efforts to use the data to draw inferences about these phenomena. The difficulty is the same as that of trying to learn the effect of SAT scores on academic performance by looking only at elite colleges. If we consider only oppressed or disadvantaged groups, we are truncating variation on the independent variable, and thus making it harder to detect a relationship between (say) discrimination and rebellion. This sample selection problem in Minorities at Risk (MAR) is one of the motivations for the present study.

Alesina et al. (2002) attempt to distinguish between ethnic, linguistic, and religious groups in a sample of about 190 countries, and then use their lists to construct measures of ethnic, linguistic, and religious fractionalization. Though it is not clear how "ethnic" and "linguistic" groups are distinguished (as they themselves allow), the descriptive statistics for their "ethnic" measure look broadly similar to those for the measure constructed here. Roeder (2002) has made available a series of fractionalization measures for 1961 and 1985 based almost entirely on Soviet ethnographic sources; his measures appear to be closer to those of the *Atlas Narodov Mira* than mine, though I have not seen the group list that underlies his estimates.

2 Coding ethnic groups

"Primordialists" are said to believe that ethnic groups are either fixed, biologically given entities, or, if they are social conventions, that they are deeply rooted, clearly drawn, and historically rigid conventions.

Anyone with primordialist leanings should be quickly disabused of them by undertaking to code "ethnic groups" in many different countries.[3] It rapidly becomes clear that one must

[3]I have been told, in effect, that the very idea of listing ethnic groups is "primordialist," because it somehow presumes or implies that these groups exist in the wrong sort of way. I see no contradiction in seeing ethnic groups as purely social facts, and trying to enumerate them. Indeed, I don't see how anyone could maintain that social facts can't be identified.

make all manner of borderline-arbitrary decisions, and that in many cases there simply does not seem to be a single right answer to the question "what are the ethnic groups in this country?" Constructivist or instrumentalist arguments about the contingent, fuzzy, and situational character of ethnicity seem amply supported.

Take, for example, the United States. What are its ethnic (or racial) groups? Let's make things *much* easier by restricting attention to groups with at least 1% of country population.[4] If we consult official census categories, we get three "races" – white, African American, and Asian – and an additional group, Hispanic, which is emphatically said "not [to be] a race." Is this the right list for the U.S.? Why not disaggregate Hispanic into Puerto Rican Americans, Cuban Americans, Mexican Americans, and so on, or likewise for Asian?[5] Why not distinguish between Arab Americans, Irish Americans, Italian Americans, German Americans, and so on? And why should we use the current census categories, when earlier censuses formulated the categories quite differently (Nobles 2000)?

Looking farther afield, does Somalia have a single ethnic group (Somalis), or several, corresponding to the major clans? If the latter, at what level in the extremely detailed hierarchical system of clan and subclan do we locate the "ethnic groups" for our list? What about castes in India? What about Berbers in several North African countries, where a large majority of the population could if they wished claim Berber descent, but attitudes vary on whether to characterize oneself as "Arab" or "Berber"? What about many Latin American countries, where the lines between "indigenous" and "mestizo," and between "mestizo" and "white," are often vague to the point of being imperceptible or situation-dependent. What about Sudan, where one might code oneself as a (black, not Arab) Southerner in one context, but as a Dinka in another?

[4]I suspect that without this restriction or some other low threshold it will be impossible to enumerate all "ethnic groups" in all countries.

[5]With the 1% threshold, this could mean dropping Asian entirely from the list.

An explicit definition of "ethnic group" could help with some of these questions, but it is important to see that it could not plausibly solve all of them. For example, any definition of ethnic group that said unambiguously that "Hispanic" is an ethnic group in the U.S. but "Cuban American" is not is prima facie implausible. The nature of the concept of "ethnic group" is such that there can be multiple ways to specify the set of ethnic groups in a country, all of which include more-or-less equally valid "ethnic groups."

This observation has an important implication for social science research that uses measures of ethnic diversity to explain outcomes such as economic growth or political violence (for example). If there are multiple plausible ways of listing a country's "ethnic groups," we must be careful that we do not, in effect, choose the coding that best supports our theory, after the fact. Somalia was viewed by the Soviet ethnographers in 1960 as highly homogenous, a nation of ethnic Somalis sharing religion, language, and customs. This was a perfectly plausible coding then and it remains so today. Since the civil war of the 1990s, however, analysts seeking to explain poor prior economic growth or the war itself would be drawn to argue that Somalia is highly ethnically fractionalized along clan lines, and thus a good example of the proposition that ethnic heterogeneity causes poor economic performance and civil strife. Designating Somalia as highly fractionalized is not implausible either, whether for 1960 or 1990. Or consider Botswana, a case often used to support the argument that "Africa's economic growth tragedy" is explained by ethnic heterogeneity. With its large Tswana ethnic group, Botswana can be plausibly coded as homogeneous by African standards, and its economy has performed very well. Yet Botswana's ethnic structure is fundamentaly similar to Somalia's – the Tswana are divided into eight subtribes that are socially and politically consequential. If for some reason Botswana's economy had done poorly over the last 30 years, and if it had seen significant internal fighting along tribal lines, it would have been viewed ex post as confirmation of the "regularity" that ethnic diversity

makes for low growth and a greater risk of civil conflict!

So what can be done? Many of the problematic cases noted above have a common origin: Where to locate the "ethnic group" when there are two groups, and group B is a subset of group A? One approach is to avoid a decision, instead incorporating these set/subset relations in the structure of the data. That is, we might code multiple "levels" of ethnic groups, where, in the U.S. for example, the census categories form level 1, a disaggregation by country of origin forms level 2 (Mexican-American, Vietnamese-American, etc.), and so on. Following Scarritt and Mozaffar (1999), I partially build such structure into the group lists for sub-Saharan Africa, where this issue is particularly common and difficult.

But for many purposes, such as producing a cross-national measure of ethnic diversity, we will want a single list of groups for a country. It is not evident that the "levels" would correspond across countries, making it sensible to compute "level 1 fractionalization," "level 2 fractionalization," etc. Moreover, sets and subsets are not the only problem we encounter. Should Mexico be divided between "indigenous" and "mestizo/white," or should "white" be broken out? Or if we are listing hyphenated Americans for the U.S., do we include, say, "German Americans," even if this is at best a vague category rather than a group in the sense of a set of people who recognize and feel motivated to act on the basis of this membership?

Implicit in the idea of an ethnic group is the idea that members and non-members recognize the distinction and anticipate that significant actions are or could be conditioned on it. So it is natural and perhaps necessary that the "right list" of ethnic groups for a country depend on what people in the country identify as the most socially relevant ethnic groupings. I adopt this approach for the list discussed below, in principle if not literally in practice. Ideally, the standard for "the right list" that I am seeking would be defined by a procedure like the following:

1. Randomly sample a large number of people in the country.

2. Ask each of them to list the major or main ethnic groups in the country.

3. Show them or read a list of many possible formulations of the ethnic groups in the country, and ask them to say of which they consider themselves members.

4. Repeat (3), asking them to say of which groups on the list most other people in the country would consider them to be members.

5. Ask them to try to rank the groups they identified in (3) according to how strongly they identify with the group (e.g., which is "most important to you," or some such language).

Such a survey could be useful for many interesting purposes besides that of constructing a list of ethnic groups by country (for which I would expect to analyze responses to question 2). It could be used to assess the degree of social consensus on what are the country's "ethnic groups," which might not be particularly high in many cases. If taken at multiple points in time, it could be used to study the political or economic determinants of "situational ethnicity," factors that lead people to see this-or-that element of their "identity repertoire" as more or less important at different times. Rankings by importance in question 5 could allow a more subtle and nuanced mapping of levels of ethnic identity and possibilities for reformulation and coalitions. The differences between answers to questions 3 and 4 could allow an inquiry into gaps between subjective understandings and objective assessments of ethnicity (for example, many white Americans might identify "Asian" as a race, but how often would "Asians" self-identify this way?).

Without survey data of this sort, we are forced to review existing lists and secondary sources to apply this standard as best we can. The main sources employed are dicussed in section 4 below.

Before proceeding, I stress two points that follow from the observation that what the

ethnic groups in a country are depends on what the people in the country *think* they are at a given time. First, it cannot be assumed, without argument, that ethnic distinctions are wholly exogenous to other political, economic, and social variables of interest. For example, poor economic performance could exacerbate distributional struggles, causing people to see and act along lines of ethnic division that were formerly considered unimportant. By contrast, robust economic growth might lead to the downplaying of ethnic divisions and a greater emphasis on national identity. If Botswana seems more ethnically homogeneous than Somalia does at this point, it may be that this is in part a result rather than a cause of economic growth. Likewise, many examples, such as Somalia, show that political violence can lead or force people to identify more strongly along ethnic lines that formerly were less salient. This may be an argument for using a list of ethnic groups constructed in 1960, such as the *Atlas Narodov Mira*, to study subsequent economic growth or political conflict.

Second, we can't use the list to ask empirically why some possible ethnic groups become actual ethnic groups at a given time, or why ethnic as opposed to other political cleavages develop. We might want to know why possible ethnic groups such as German-American or Scots-Irish do not have the same social and political salience that white and black do in the United States at present, or why Kenyan national electoral coalitions are structured by divisions among Kikuyu, Luo, Kalenjins, Kamba, etc., rather than between men and women, or rich and poor. Obviously, if a criterion for inclusion in the list is that people in the country see the category in question as an ethnic group, then we do not have a sample of all hypothetically possible ethnic (or other) groups. Nonetheless, a list of "actual" or existing ethnic groups would be a prerequisite for such a study. The trick would be constructing the list of "potential" ethnic (or other) groups. Since it is not clear that the population of "all possible ethnic groups in a country" is well-defined, even in theory, some sort of case-control

approach would be necessary.[6]

3 Ethnicity

I argued in the last section that no plausible definition of "ethnic group" will by itself imply a unique list of groups for a country. Still, a definition would be useful to bound the phenomenon we are trying to capture, and to address questions like the following. Are Protestants and Catholics in Northern Ireland, or Bosnian Serbs and Muslims, to be included if the only significant cultural difference is *religion*?

Standard definitions of "ethnic group" in terms of a shared belief of common ancestry and/or shared cultural features are problematic (Fearon and Laitin 2000). It is almost always possible to give examples of groups that fit the definition taken literally, but that are not intuitively "ethnic," or of groups that do not fit the definition but that are often described as "ethnic."

Fearon and Laitin (2000) attempt to deal with this problem by examining the implicit rules that people (or at least English speakers) use to decide which groups are "ethnic" in everyday talk. They argue that in common speech a group may be designated as "ethnic" if the group is larger than a family and membership in the group is reckoned primarily by a descent rule. These are the core criteria, although the concept may be further restricted to rule out cases such castes or nobility, and groups that are "legislated" into existence (i.e., have no "naturalized history" as a group). It is worth noting that shared cultural features seem to play no necessary role in whether a group can be described as "ethnic" in everyday talk. For example, "Jews" are often described as an ethnic group despite lacking a common language, universally shared customs, or even common religious practice (since

[6]That is, "randomly" select a set of potential but not actual groups (or non-ethnic groups), making no pretense of getting the whole population (or even defining it, conceptually). Then use techniques such as those discussed in King and Zeng (2001) to analyze the resulting sample.

non-believers are typically included in the group, and it is contested whether conversion can make one ethnically Jewish). Somali clans are frequently referred to as "ethnic" formations, even though their members do not see the clans as culturally distinct in any significant respect.

The results of the ordinary language analysis also help explain when groups sharing a common religion will be considered "ethnic" – namely, when membership in the group is reckoned primarily by descent rather than by public confession of faith. In Bosnia, thugs distinguished between Serbs and Muslims on the basis of local knowledge and records concerning descent, not by tests of religious faith or practice. In the United States, one can make oneself Protestant or Catholic by adopting the appropriate religious practices and beliefs, something that is hard or impossible to do in Northern Ireland.

Another approach to definition – in several ways more useful for the purpose of constructing a list by countries – is to employ the idea of "radial categories" advanced by linguists and cognitive scientists (Lakoff 1987; see Collier and Mahon 1993 for a discussion of with respect to political science). In practice, people may understand the meaning of a concept X by reference to *prototypical cases*. Less prototypical cases may not share all the features of a prototype, and yet still be validly classed as Xs, at least in some circumstances.

For example, the prototypical ethnic group has the following features:

1. Membership in the group is reckoned primarily by descent by both members and non-members.

2. Members are conscious of group membership and view it as normatively and psychologically important to them.

3. Members share some distinguishing cultural features, such as common language, religion, and customs.

4. These cultural features are held to be valuable by a large majority of members of the group.

5. The group has a homeland, or at least "remembers" one.

6. The group has a shared and collectively represented history as a group. Further, this history is not wholly manufactured, but has some basis in fact.

7. The group is potentially "stand alone" in a conceptual sense – that is, it is not a caste or caste-like group (e.g., European nobility or commoners).

The term "radial" comes from the observation that by taking away one or more of these features, one may get types of "ethnic groups" that are not prototypical but nonetheless are often seen as ethnic groups. For example:

- Take away 2, 4, and 6, (and possibly others except 1), and you get an *ethnic category* rather than an ethnic group. The extent or degree to which these conditions apply might be said to determine the "groupness" of a group (Brubaker 200x).

- Take away 5 (and possibly others except 1) and you get some *nomadic* ethnic groups, such as the Roma.

- Take away 7 and you get castes in South Asia, or noble/commoner distinctions in Europe.

In assembling the list discussed below, I am looking for groups that meet the "prototype" conditions as much as possible. This implies that I allow groups distinguished from others in the same country primarily by religion provided that they meet condition 1 (membership has a strong descent basis) and condition 2 (self-consciousness as group). It also implies that I do not count castes in South Asia as ethnic groups, even though I would readily admit that they share an important "family resemblance" to ethnic groups through the

descent criterion, and could be validly considered as ethnic groups in some research designs (Horowitz 1985; Chandra 2000).

I believe that the vast majority of the groups in the list discussed below meet the conditions for a "prototypical" group fairly well, although for a number of cases, especially in Asia and Africa, the extent to which 2, 4, and 6 are met is unclear. These continents have many "groups" that are identified by some language commonality, which in most cases does mark some cultural similarity. But the extent of their "groupness," or sense of common identity (conditions 2,4, and 6) is not clear from the sources I have been able to consult so far.

4 Sources

Working with Alex Rosas, Christina Maimone, and Atsuko Suga, I used the CIA's *World Factbook* online for a "first pass." The *Factbook*'s numbers and designations were then compared with those in *Encyclopedia Brittanica* (*EB*) and, when possible, the relevant Library of Congress Country Study (*LCCS*). Significant discrepancies between these sources prompted an investigation using country-specific sources. For a number of countries and particularly for Latin America, *LCCS* provides a nuanced discussion of the nature of ethnic identity. These were often used to modify the *Factbook*'s listing. For example, for choices about whether to code "whites" separate from "mestizos" in Latin America I followed *LCCS* when possible.

I also compared the *Factbook*, *EB*, and *LCCS* groups and numbers with the minority groups listed in the Minorities at Risk data set. Though MAR does not purport to cover all ethnic minorities in a country, it has the advantage of including groups that are almost all "mobilized" or have a nontrivial level of "groupness." In a few cases I included a group they identified but which does not appear in the *Factbook*.

The *Factbook* generally does not list the large non-citizen populations that inhabit

many Western European countries and many of the Gulf states. Excluding these seems hard to defend if we want a list of ethnic groups in a country at a given time – would a country with 50% white citizens and 50% black noncitizens be properly regarding as ethnically homogenous? For information on noncitizens, I consulted recent census figures for OECD countries, and a variety of web sources both for these and for the Gulf states.[7]

The subSaharan African countries pose special problems. In general they are remarkably ethnically diverse, and Africans often manifest their multiple ascriptive affiliations in highly complex, situation-dependent ways. At the time of access, at least, the *Factbook* was unusable for much of the continent, providing either uninformative or superficial breakdowns (e.g., Bantu/Nilotic, or a statement about the total number of ethnic groups in the country). Fortunately, Scarritt and Mozaffar (1999) have carefully constructed a list of over 300 "ethnopolitical" groups in 48 African countries. Working from Morrison et al. (1989) and a large number of country-specific accounts, Scarrit and Mozaffar sought to list ethnic groups with "contemporary or past political relevance" at the national level. For my purposes, an important advantage of their data is that they required country-specific evidence on the shared awareness and political significance of a ethnic category in order to include it, so that these are more likely to be "real" groups.

A disadvantage is that for my purposes *political* significance is too restrictive. For example, for Burkina Faso Scarritt and Mozaffar list only the Mossi, at 50% of the population, because they found no evidence that the other ethnic groups had any "political relevance" at the national level (the other groups are excluded by the Mossi). So we returned to Scarritt and Mozzafar's main source, Donald Morrison et al.'s *Black Africa: A Comparative Handbook* and tried to identify those groups greater than 1% of country pop-

[7]Note that estimates for Saudi Arabia and some other Gulf states are fairly uncertain, as it appears that these kingdoms are overstating their citizen populations so as to look more like "real" countries.

ulation that were excluded under the political-relevance rule. We reconsidered all countries for which the sum of the group percentages in Scarritt and Mozaffar's list was less than 95.[8] Parallel to the process for the rest of the world's countries, we took Morrison (1989) as our base, and then compared Morrison's list with those provided by the Summer Institute of Language's *Ethnologue*, and Levinson (1998). Significant discrepancies were resolved by resort to country-specific sources.[9] For a number of these countries – for example, Chad, Congo-Brazzaville, the Democratic Republic of Congo, and Liberia – I do not have great confidence that all of the groups listed accurately reflect how people in the country mentally divide the social terrain in ethnic terms. The sources used overwhelmingly identify groups by shared languages, but there are often so many closely related languages/dialects that it is difficult to know where perceptions of groupness attach most strongly.

An innovative feature of Scarritt and Mozaffar's (1999) data is that they code groups at three levels of aggregation which they term "national dichotomy," "middle-level of aggregation," and "lower level of aggregation." The first refers to situations where "Virtually the entire population ... is at least fairly intensely politicized as part of one side or the other of a long-standing national ethnopolitical dichotomy" (89). Hutu/Tutsi in Rwanda and Burundi, and Mainlanders/Zanzibaris in Tanzania, are examples of this coding. The "middle level" lists ethnopolitical groups and in some cases coalitions of groups that act together politically, but which do not necessarily partition the whole population. "Lower level" groups are bro-

[8]Using Scarritt and Mozaffar's second level of aggregation, on which see below.

[9]Although it is difficult to do because the language groups listed in *Ethnologue* are highly disaggregated, we often constructed population estimates for African groups identified in Morrison and other sources by searching the *Ethnologue* list for languages closely related to the group name (or some variant of it). In general, we found that the group population proportions based on *Ethnologue*'s language-speaker estimates – which are typically dated in the early 1990s and presumably come from linguists and missionaries – were remarkably close to the population proportion estimates from Morrison, who references sources going back to the 50s and 60s (often the last colonial census). When there were significant differences and some other source (especially the *Factbook* tended to corroborate the more recent estimate based on *Ethnologue*, we adjusted the figures accordingly.

ken out under middle-level groups in some cases, where there is a "significant ethnopolitical cleavage within middle-level groups" (90).

Above, I noted that a major obstacle to listing a country's "ethnic groups" is that people commonly have multiple ascriptive attachments organized in set/subset relationships – Hispanic/Mexican-American, for example, or (black) Southerner/Nuer, in Sudan. One way to deal with this issue is simply to incorporate it in the structure of the data, coding groups at different levels. Although Scarritt and Mozaffar's three levels are not motivated by this same observation about the structure of ethnic attachments, in practice the codings for their three levels tend to reflect the set/subset phenomenon noted here. For instance, Scarritt and Mozaffar code Kalenjins and Luhyas in Kenya at the middle level of aggregation, but also list a number of Kalenjin and Luhya tribes at the lower level.

In the raw data used to generate the ethnic group list examined below I have preserved and in some cases added to Scarritt and Mozaffar's scheme of three levels of aggregation. In future work I would like to rationalize and extend this approach to the rest of the world's countries. Such data would provide a richer and more accurate rendering of the organization of ethnicity across countries. For present purposes, however, I have gone through the subSaharan countries and selected out the level of aggregation that produces a list of groups that, in the mid-90s, are judged to be collectively closest to the "prototypical case," as assessed by additional country-specific research. This task is made less subjective than it may sound by the facts that (1) Scarritt and Mozaffar code only 12 cases of "national dichotomies" and these are mainly obvious cases, and (2) in many cases there are virtually no "lower" level groups listed. But certainly there are some difficult countries here, such as Somalia (should ethnic groups be measured at the subclan level or just Hawiye, Issaq, Darod, etc.?).

5 Descriptive statistics

The list of ethnic groups resulting from the procedures described above has 820 groups in the 160 countries that had over half a million in population in 1990.[10] Table 1 provides descriptive statistics for the sample as a whole and for six cultural regions.

Considering the sample as a whole, we find that the "average country" has about five ethnic groups that are larger than 1% of the population, with half of the world's countries having between 3 and 6 such groups (this is the interquartile range). Tanzania, with 23 groups, tops the list, while Papua New Guinea, with zero, is the somewhat anomalous minimum. How can a country have zero ethnic groups? Recall that I am coding only ethnic groups that make up over 1% of country population. The sources I have consulted are consistent in characterizing the primary ethnic units of Papua New Guinea (PNG) as extremely small. The U.S. State Department's *Background Notes* for PNG are indicative of what the anthropologists say as well.

> The indigenous population of PNG is one of the most heterogeneous in the world. PNG has several thousand separate communities, most with only a few hundred people. ... The isolation created by the mountainous terrain is so great that some groups, until recently, were unaware of the existence of neighboring groups only a few kilometers away.[11]

While broad classifications, such as Papuans/Melanesians, or Highlanders/ Sepak Valley/etc., are sometimes mentioned, there is general agreement that PNG citizens' primary ethnic attachments are to these very small groups, which are almost always differentiated

[10] Because of a large changes in their ethnic compositions following their break ups, the Soviet Union and Russia are entered as two different countries, as are Yugoslavia and Yugoslavia/Serbia.

[11] "Background Note: Papua New Guinea," U.S. Department of State, Bureau of East Asian and Pacific Affairs, October 2001. See Reilly (2000/01) for a similar assessment and references to the anthropological literature.

by language. By the ethnic fractionalization measure discussed in the next section, PNG approximates a perfectly fractionalized state.

Returning to Table 1, we see that about 70% of the countries in the world have an ethnic group that forms an absolute majority of the population, although the average population share of such groups is only 65% and only 18% of countries are "homogenous" in the weak sense of having a group that claims 9 out of 10 residents. The average size of the *second* largest group, or largest ethnic minority, is surprisingly large, at 17%. This is not due to the influence of a single highly diverse region, such as SubSaharan Africa. Seventeen percent is close to the average size of the largest minority in every region except the West, where the largest minorities tend to be smaller (and the majority ethnic group larger).

Turning to regional variation, what is most striking is how much more ethnically divided are the subSaharan African countries. With 350 groups coded, Africa accounts for about quarter of all countries but 43% of the world's ethnic groups (larger than 1% of population). While the rest of the world's regions average between 3.2 and 4.7 groups per country, the African countries' average is greater than eight. The average population share of the largest ethnic group in these countries is 42%, less than a majority, in sharp contrast to all other regions. SubSaharan Africa has no "homogenous" countries and less than a third have an ethnic majority.

A second interesting feature of the regional statistics is how *small* are the aggregate differences between the countries of North Africa/Middle East, Latin America/Caribbean, Asia, and Eastern Europe/Former Soviet Union. The Western countries are somewhat more homogeneous, and, as noted, the subSaharan countries are considerably more diverse on average. But the rest of the world's regions show broadly similar ethnic demographies. The average number of groups per country and the average size of the top two groups are all quite similar. The percentage of countries that are "homogeneous" or that have an ethnic

majority are also fairly similar in this set (although Eastern Europe has a somewhat higher proportion of ethnic majorities and small number "homogeneous" countries).

Of course, similarity in broad ethnic demography does *not* imply that these regions have similar ethnic politics, interethnic relations, or economic or political outcomes. To the contrary, we know that they do not. Though hardly definitive, these data suggest that scholars who want to explain differences in political or economic outcomes by reference to cross-national differences in ethnic demography may face an uphill task.

6 Ethnic structures

In cross-national studies of political violence, economic growth, and other outcomes in political economy, analysts most often use ethnic fractionalization as a measure of ethnic diversity. This is defined as the probability that two randomly selected individuals in a country will be from different ethnic groups. But many hypotheses and arguments in the literature refer not just to measures of ethnic diversity like this one, but to more fine-grained conceptualizations of ethnic structure. For example, Horowitz (1985) and others say that ethnic conflict is more likely in countries with an ethnic majority and a large ethnic minority, as opposed to homogenous or highly heterogeneous countries. Reilly (2000/01) observes that fractionalization is ill-suited to capture different structures of ethnic cleavages – for instance, highly fragmented (Papua New Guinea), bipolar (Cyprus), multipolar and balanced (Bosnia), dominant majority (Sri Lanka), or dominant minority (Burundi).

A simple way to use these data to get a sense of how ethnic structures vary around the world is to graph the population share of the second largest group (the largest minority) against the share of largest group (the plurality group). This is undertaken in Figure 1 for each region, using an abbreviation of the country name as the plotting symbol. Because the second largest group is by definition no larger than the largest group, the points in

these graphs necessarily fall within a triangle with vertices at (0, 0), (.5, .5), and (1, 0). Countries located near the (1, 0) vertex have a large ethnic majority and are thus relatively homogeneous (e.g., Tunisia). Countries located nearer to (0, 0) are highly fragmented (e.g., Tanzania and Uganda; PNG would be approximately at (0,0)). A country near to (.5, .5) is roughly "bipolar," with two large ethnic groups dividing most of the population (e.g., Fiji). Finally, a country located near to the x-axis has a single plurality group and a highly fragmented set of ethnic minorities (e.g. India).[12]

Figure 1 illustrates more dramatically how different are "ethnic structures" in subSaharan Africa from those in other regions. Whereas countries with no ethnic majority are fairly rare in the rest of world, this is the norm in Africa. Most African countries cluster on the left side of the triangle, around a point that implies a plurality group of about 22%, with the second largest slightly less than this. The figure also shows considerable variation in ethnic structures within Africa. Rwanda, Burundi, Lesotho, Swaziland, and Zimbabwe have a large majority group and a minority that makes up almost all of the rest of the population. Botswana is coded as having a large majority (the Tswana) and a set of smaller minorities.[13] Mauritania and Djibouti are fairly evenly divided between two large groups, while there is a set of countries (e.g., Mali, Burkina Faso, and Namibia) that has a relatively large plurality group with the rest of the population divided among quite small groups.[14]

Outside of Africa, the figure shows the West and Eastern Europe/FSU as the regions with the largest clusters of relatively homogeneous states. Among the less homogeneous

[12]I coded India using language groups, of which Hindi is the largest. Certainly there are other plausible renderings of India's "ethnic groups," but most likely all of them would imply a high level of diversity, which is true for language groups.

[13]The sources I consulted stressed that identity as a Tswana is generally more important than identity as a member of one of the subtribes of the Tswana (though no doubt this is highly context specific). As noted above, this could well be a result of Botswana's strong economic and political performance rather than a cause.

[14]White and Black Moors in Mauritania, Afars and Issas in Djibouti.

countries in the West, the largest ethnic minority tends to make up about half of the population outside of the majority group. This is true for EE/FSU as well, but these countries show much more variation around this pattern. The U.S.S.R. had and Kyrgyzstan has a bare majority group and a large number of small ethnic minorities; the Baltic states are approximately bipolar with a moderate sized majority (i.e. the titular) group; the former Yugoslavia had a structure typical of subSaharan countries, and Kazakhstan is not too far from an evenly balanced bipolarity.

Latin America and the Caribbean are notable for the high proportion of the countries that are approximately partitioned between a majority group and a single minority group, usually "mestizos" (or "whites") and "indigenous peoples." "Indigenous peoples" is of course a catch-all, often combining groups that were historically divided among many smaller tribes speaking diverse languages. A long history of assimilation and the numerical and political dominance of the settler populations has blurred these distinctions and made the commonsense ethnic categories in many of these countries "indigenous" versus "white/mestizo." Exceptions are Guatemala and the Andean countries Bolivia, Peru, and Ecuador, which are coded as having large indigenous populations, along with noteworthy distinctions between whites and mestizos (in the Andean countries). For Bolivia, the sources suggested a distinction between Quechua and Aymara speaking indigenous peoples. Along with Trinidad and Tobago, Guatemala, Ecuador and Peru look approximately bipolar by this rendering. Structurally similar to Bosnia, Bolivia is divided between three (if whites and mestizos are combined) or four fairly equal sized groups. The coding of Colombia towards the middle of the triangle depends on distinguishing between white and mestizo. In the *cultural* diversity measure discussed below, Colombia will look much more homogeneous.

Finally, Asia and North Africa/Middle East show similar patterns of ethnic structure. Both regions' countries mostly have ethnic majorities, but in both there are a number with

a sometimes slim ethnic majority that faces a large number of small ethnic groups. For Asia this often reflects a configuration of a large lowland majority that is ringed or edged by more fragmented mountain peoples (Burma, Laos, Thailand, Vietnam, Pakistan (slightly)). For the Middle East, it reflects the political economy of oil production in the Persian Gulf. Saudi Arabia, Bahrain, United Arab Emirates, Oman, and Kuwait have ethnically homogeneous groups of citizens who are either a bare majority or a mere plurality; the rest of the population is typically made up of ethnically diverse noncitizen workers. Iran comes by this structure more honestly, as it were, with a bare majority of Persians, 24% Azeris, and seven other quite small groups. North Africa/Middle East is also notable for the number of countries that, by my codings, are almost strictly divided by two ethnic or ethnoreligious groups[15]: Arabs and Berbers in Morocco, Algeria, Libya, and Tunisia; Muslims and Copts in Egypt; Turks and Kurds in Turkey; Greeks and Turks in Cyprus; and Palestinians and TransJordan Arabs in Jordan.

7 Ethnic fractionalization

The most commonly employed measure of aggregate ethnic diversity is *fractionalization*, defined as the probability that two individuals selected at random from a country will be from different ethnic groups. If the population shares of the ethnic groups in a country are denoted $p_1, p_2, p_3, \ldots, p_n$, then fractionalization is $F = 1 - \sum_{i=1}^{n} p_i^2$. Table 2 gives a few examples of how the measure works.

[15]That is, right on the downward-sloping line of the triangle.

Table 2: Fractionalization Examples

Country	Structure	F
A	Perfectly homogeneous	0
B	2 groups, (.95, .05)	.10
C	2 groups, (.8 ,.2)	.32
D	2 groups, (.5, .5)	.50
E	3 groups, (.33, .33, .33)	.67
F	3 groups, (.55, .30, .15)	.59
G	3 groups, (.75, .20, .05)	.40
H	$(.48, .01, .01, \ldots)$	.76
I	(.25, .25, .25, .25)	.75
J	n groups, $(1/n, 1/n, \ldots)$	$1-(1/n)$

In line with the discussion about ethnic structures above, notice that the fractionalization scores for countries E and F are not that different, even though one might expect their ethnic politics to differ markedly given that there is an absolute majority in F but not in E. As a continuous measure, F is not sensitive to discontinuities related to the idea of majority rule. The comparison between countries H and I makes a different point. As a one-dimensional measure, F cannot fully capture differences in ethnic structures that may seem intuitively significant.

Still, as an index of overall ethnic diversity F has much to recommend it. It has a natural intuitive interpretation. It is far superior to the number of ethnic groups because it takes account of population shares. It encodes more information than would using the population share of the largest group (though these measures are quite close). And its empirical distribution – summarized numerically in Table 3 – is not highly skewed.[16] The average value of .47 for all countries implies that if one were to select a country at random, then randomly select two people from it, there is about a 50-50 chance that they would come

[16]Cox (1997) and others sometimes prefer to use the "effective number of ethnic groups" (or political parties' vote or seat shares), which is $1/(1-F)$. Thus, a country with n equal-sized groups has an "effective number" of n groups, with departures from equal shares shrinking the effective number continuously. Although the interpretation is "nice," this measure is *highly* skewed, at least for ethnic fractionalization, so that it tends to exaggerate the influence of very diverse countries like Tanzania when used as an explanatory variable.

from different ethnic groups.

Table 3: Ethnic Fractionalization

Region	N	10th pctile	Median	Mean	90th pctile
World	160	.11	.50	.47	.81
West	21	.04	.15	.24	.57
EE/FSU	31	.13	.39	.41	.68
LA/Ca	23	.13	.48	.41	.65
Asia	23	.15	.43	.44	.77
NA/ME	19	.08	.51	.45	.74
SSA	43	.35	.76	.71	.89

For each region, Figure 2 plots F as measured using the *Atlas Narodov Mira* against F computed using the data discussed above. The agreement between the two codings is quite high, except in North Africa/Middle East and Latin America/Caribbean, where my constructions show considerably more diversity for a number of countries. The bivariate correlation for the whole sample – which consists of only 135 states because of new countries (mainly in the FSU) not coded by the Soviet ethnographers – is .75.

Different conceptions of ethnicity explain some differences between the two measures. The Soviet geographers code ethno*linguistic* groups, adopting the common Eastern European assumption that native language marks ethnicity. As discussed above, I allow for other cultural criteria distinguishing groups, provided that the groups are locally understood as (primarily) descent groups and are locally viewed as socially or politically most consequential. In Eastern Europe/F.S.U. and to a slightly lesser extent in the West, language does indeed tend to mark ethnicity in "my" sense, so the correlation between the two measures is nearly perfect. In Latin America, however, the Soviet ethnographers code all Spanish speakers as one ethnolinguistic group, and tend to break out the "indigenous peoples" by tribal language. On net, this makes for considerably greater homogeneity by their measure for this region. This consideration also explains a number of prominent outliers in other regions. The Soviets

code Burundi as ethnically homogenous, since both Hutus and Tutsis speak Kirundi![17] The common languages of Somali and Malagasy make Somalia and Madagascar appear nearly perfectly homogenous in the Soviet coding (which could be argued as plausible in each case). They draw no distinction between White and Black Moors in Mauritania because both speak Arabic. In an exception to their normal practice, they code Papua New Guinea by racial categories (Papuans and Melanesians) rather than by language groups, which leads to a much less fractionalized estimate for PNG in their data.[18]

Several countries in the Middle East are coded quite differently by the two measures for this same reason. I distinguish between Palestinians and TransJordan Arabs in Jordan whereas the Soviet team sees them all as Arabs because they speak Arabic. Likewise, I code the ethnoreligious groups in Lebanon whereas the Soviet team sees this country as almost all "Arab"; and similarly for Alawi and Christians in Syria, and Sunni and Shia Arabs in Iraq. But there is another reason for the low correlation (.22) between the two measures in this region. I code the large noncitizen populations in the Gulf states, who comprised much smaller groups in these countries in the early 1960s (it appears that the Soviet team did try to include them). Ironically, the states that show by far the greatest increase in ethnic diversity due to "globalization" over the last 40 years are the Gulf monarchies.[19]

[17]Interestingly, we differ very little for Rwanda even though we identify different ethnic groups – the Soviets code a moderately large number of Kirundi speakers in Rwanda as a distinct group, next to a large majority of (Hutu and Tutsi) Kinyarwanda speakers.

[18]The Soviets coded the Philippines by a combination of language and islands (e.g., "Visayans"), which makes for a much larger fractionalization estimate than I have (I code "lowland Christian Malays" as the main group, in line with the *Factbook* and the discussion in *LCCS*.

[19]Which they have managed, of course, in the same way as "the West" – by keeping the newcomers largely as noncitizens who come and go.

8 A measure of cultural diversity

With a fractionalization score of .37, Belarus falls at about the 40th percentile on ethnic diversity from a cross-national perspective. This reflects a division between the majority Byelorussian group (78%), Russians (13%), Poles (4%), and Ukrainians (3%). Cyprus, coded as 78% Greek and 18% Turkish, is assessed as practically the same as Belarus in terms of ethnic fractionalization, at .36.

If one has a theory that says that ethnic diversity matters because ethnic differences make it harder for people to cooperate and coordinate, then one might be interested in some notion of the *cultural distance* between ethnic groups rather than just fractionalization. Intuitively, even though their F scores are about the same, Belarus is much less culturally divided than Cyprus. Byelorussians, Ukrainians, and Russians are quite similar in terms of religion, language, and customs, and Poles speak a Slavic language and share many of the same customs. By contrast, Greeks and Turks speak languages that come from completely different families (Indo-European and Altaic), subscribe to two different world religions (Orthodox Christianity and Islam), and have very different customs. In this section I construct a measure of cultural distance that modifies fractionalization so as to take some account of cultural distances between groups. To continue the above example, by this "cultural fractionalization" measure, Belarus moves down to .22 – about the 40th percentile on cultural fractionalization – while Cyprus stays at .36 – which is now at the 60th percentile of the new measure.

Linguists classify and represent the structural relationships between languages with the help of tree diagrams. Fearon and Laitin (1999, 2000) and Laitin (2000) propose using the distance between the "tree branches" of two languages as a measure, albeit a noisy one, of the cultural distance between groups that speak them as a first language. For example, Greek and Turkish diverge at the first branch or level, since they come from structurally

unrelated language families. By contrast, Byelorussian, Russian, and Ukrainian share their first three classifications as Indo-European, Slavic, East Branch languages. Polish shares only the first two levels with these, since it is Indo-European, Slavic, West Branch. The idea is that the number of common classifications in the language tree can be used as a measure of cultural proximity.[20]

For two ethnic groups i and j, consider defining a *resemblance factor* r_{ij} (Greenberg 1956) that works as follows. r_{ij} is zero when the two groups' languages come from completely different families (like Indo-European and Altaic). r_{ij} is 1 when the two groups speak exactly the same language. In between, we let r_{ij} be some increasing function of the number of shared classifications between i's and j's languages. Since early divergence in a language tree probably signifies much more cultural difference on average than later divergence, the function should be concave as well. (For example, coming from structurally unrelated families such as Bantu and Indo-European, denotes more cultural difference on average than does the difference between, say, Slavic East Branch and Slavic West Branch.)[21]

To construct a measure of "cultural fractionalization" analogous to the ethnic fractionalization measure F discussed above, consider drawing two people at random from a country and then computing their *expected cultural resemblance*, using r_{ij} as defined above. In a country with one language group or a set of ethnic groups that all speak highly similar languages, the expected resemblance will be close to 1. In a country with a large number of groups that speak structurally unrelated languages, the expected resemblance will be closer

[20]For some cases, there is a question about whether to use the "historical language" of the group – e.g., Gaelic for Catholics in Northern Ireland or for Scots in Britain – or the language currently spoken as a first language by most members of the group. Ideally, I would like to take into account how many generations have been speaking the "new" language. For the data discussed below, this issue is handled in a somewhat ad hoc fashion at present. For more discussion on this point see Fearon and Laitin (2000).

[21]A function that fits the bill is $r_{ij} = ((l-1)/(m-1))^{\alpha}$, where l is the "level" or branch at which i's and j's languages diverge, m is the highest number of common classifications in the data set, and α is a positive parameter less than 1. For the measure constructed below, m is 14 and I use $\alpha = 1/2$.

to zero. To get a fractionalization measure analogous to ethnic fractionalization, simply subtract expected cultural resemblance from 1.[22] If the groups in the country speak structurally unrelated languages, their cultural fractionalization index will be the same as the ethnic fractionalization index F. The more similar are the languages spoken by the different ethnic groups, the more will the cultural measure be reduced below the value of F for the country.[23]

Using the linguistic classications given by Grimes and Grimes (1996), I calculated cultural fractionalization as defined above – call it C. As shown in Table 4, its avergage value of .29 is much smaller than the average value of ethnic fractionalization (.49 when computed using my data). This indicates that taking linguistic similarities into account has a large effect for a significant number of countries. Even so, C is correlated fairly strongly with ethnic fractionalization, at .76 with my measure F, and .78 with fractionalization based on the Soviet Atlas (ELF). So by these measures, ethnic fractionalization is reasonable if hardly perfect proxy for cultural fractionalization in a broad cross-section.

[22]Formally, cultural fractionalization is $1 - \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j r_{ij}$, where p_i is the proportion of group i and n is the number of groups.

[23]This measure was first proposed by linguist Joseph Greenberg (1956) in a paper on ways of guaging linguistic diversity in a region, though he had a different proposal for assessing the resemblance r_{ij} between two languages. He termed it the "B index" (the "A index" in his 1956 paper was just F, where his groups referred to groups of first language speakers). Laitin (2000) discusses Greenberg's three measures and uses the language-tree approach to measuring resemblance to compute the B index for six Soviet republics.

Table 4: Cultural vs. Ethnic Fractionalization

	N	C	F	ELF	N_{ELF}
World	159	.29	.48	.40	135
West	21	.18	.24	.22	21
LA/Ca	23	.19	.41	.24	23
NA/ME	19	.28	.45	.23	19
EE/FSU	31	.29	.41	.29	9
Asia	22	.31	.44	.50	21
SSA	43	.40	.71	.64	42

Notes: C = Avg. cultural fractionalization. F = Avg. ethnic fractionalization using my data. ELF = Avg. ethnolinguistic fractionalization using the Soviet Atlas data. N_{ELF} = the size of the sample available from the Soviet data.

When cultural/linguistic similarity is taken into account, Latin America looks much more homogeneous than it does by the ethnic fractionalization measure. This is due mainly to the use of Spanish across "white" and "mestizo" groups, which indeed reflects considerable (some might say near total) cultural similarity. This is one of example of an attractive feature of the measure C. In many cases where there is a question about where to "draw the line" between ethnic groups, C in effect makes a principled decision. Another example is Somalia, which will have a low C regardless of where one thinks the "ethnic groups" should be located.

After Latin America, the subSaharan countries show the greatest average change when we take account of cultural/linguistic similarities. The great ethnic and linguistic diversity of Africa is represented by a fairly small number of highly articulated language trees. For example, most of Tanzania's many small groups share eight common levels (Niger-Congo, Atlantic Congo, Volta Congo, Benue Congo, Bantoid, Southern, Narrow Bantu, Central). As a result – and plausibly if arguably – the measure judges some African countries significantly less culturally diverse than they are ethnically diverse.

Not surprisingly, the cultural diversity measure C tends to be closer on average to ethnolinguistic fractionalization computed using the Soviet Atlas data (ELF). As noted,

the Soviet ethnographers defined ethnicity in terms of language and national origin, so that Latin America and North Africa/Middle East come out more homogeneous by ELF than F, while subSaharan Africa is judged highly heterogeneous by both measures. One implication is that the ELF measure may be particularly favorable to the thesis that low economic growth in Africa is due to ethnic diversity (Easterly and Levine 1997), since it represents Africa as more ethnically diverse compared to the rest of the world than C or F does.

Figure 3 plots cultural fractionalization against the ethnic fractionalization measure F by region. Under C, the Latin American countries are essentially partitioned into two sets, those with substantial indigenous populations and those without. The cultural measure shows much greater variation within subSaharan Africa than F does, as a number of countries that appear highly ethnically diverse appear much less so when we take into account language proximities. Angola, Somalia, Zambia, and Madagascar are most affected in this respect.

9 Conclusion

Several active research programs in economics and political science require, for empirical evaluation, data on ethnic groups across countries. The research reported here tries to do a better job of conceptually grounding, operationalizing, and constructing a list of ethnic groups across countries than is available in the literature. As shown, the list of ethnic groups presented here can be used to produce cross-national measures of ethnic diversity, ethnic structures, and cultural diversity. Another use, not illustrated, would be "cross-group" research on the factors that distinguish the groups with members involved in secessionist struggles (Fearon and Laitin 1999).

The concept of an "ethnic group" is inherently slippery. There are often multiple plausible ways of partitioning the "ethnic groups" of a country. For example, the 11 largest groups listed for the United States by the *Atlas Narodov Mira* are "Americans (including

blacks), Jews, Germans, Italians, Mexicans, Poles, Irish, Swedes, Austrians, Puerto Ricans, Anglo-Canadians." I don't know that I would say that this is a highly plausible way of rendering the United States' ethnic groups, but at any rate it is quite different from White, Black, Hispanic, and Asian, the groups that appear in my list.

It is interesting to learn, then, that despite sharply different formulations of "ethnic group," the aggregate measure of ethnic fractionalization based on the *Atlas Narodov Mira* data and the data presented here are moderately well correlated, at .75. Very similar correlations obtain between the Soviet ELF and the "ethnic" and "linguistic" fractionalization measures produced by Alesina et al. (2002). Roeder's (2002) several measures correlate at around .81 with my measure F and at about .88 with the Soviet ELF. So as a measure of aggregate ethnic diversity across countries, fractionalization appears to be fairly robust to the looseness of the concept of "ethnic group."[24]

Still, a correlation of .75 means that only a little more than half of the variation in the two measures is "shared." The analysis above showed that there are some systematic regional differences in how my measure and the Soviet ELF assess ethnic diversity, so that certainly not all of the unshared variation is pure noise. In addition, there is some reason to be concerned that perceptions of what the ethnic groups are in a country can be caused by the dependent variables that ethnic fractionalization is supposed to predict (like growth and conflict). Researchers should therefore check to see whether their results concerning the effect of ethnic fractionalization on economic growth, political conflict, political party structure,

[24]Another way that cross-national fractionalization measures could be misleading is if the estimates of group proportions are systematically wrong. For many countries, especially in the developing world, it seems likely that the group proportion estimates found in the CIA Factbook and other such sources ultimately derive from the last colonial era census, since very few post-colonial censuses ask questions about ethnicity. My experience trying to match the recent population estimates based on *Ethnologue* with the much older, usually census based estimates in Morrison, Mitchell and Paden (1989) showed a remarkable degree of consistency, which is reassuring. Also, by construction fractionalization is not very sensitive to small changes in group proportions, and its value is determined mainly by the share of the largest group. So I doubt that there are major problems being caused by errors in the estimated population shares.

etc., depend on the specific measure used, and if they do, why. Relatedly, if a researcher's theory is that ethnic fractionalization matters because it makes for diverse preferences and consequent difficulties cooperating, then the measure of *cultural* fractionalization introduced in section 8 may be more appropriate.

Finally, the partial robustness of ethnic fractionalization measures is of no help if one's research project is at the level of ethnic groups (e.g., a study of determinants of group oppression or rebellion). In this case, it matters that the groups listed be the "right" groups in some defensible sense. I have argued that in principle the right list must depend on contemporary views of the people in the country in question, so that in the end survey data is required. In lieu of such data, the best we can do is to consult country experts who have a sense of how citizens "map" ethnicity in the country. Thus, the list discussed here is offered as provisional and to be amended and corrected, not as a definitive statement of an objective, unchanging reality.

Many possibilities for further, related research could be noted. In concluding I will mention just one. The cultural diversity measure constructed here used structural relationships between languages as a proxy for cultural similarity. This obviously leaves out other dimensions of cultural resemblance, most notably shared religion. While a variety of measures of religious fractionalization have been constructed (Alesina et al. 2002, Barro and McCleary 2002, Fearon and Laitin 2002), so far no cross-national data examines whether cross-cutting or overlapping cleavages between language/ethnicity and religion matter for dependent variables of interest. It should be relatively straightforward to use the group list discussed here to categorize countries by the interaction of religious and linguistic cleavage structures.

References

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wazciarg. 2002. "Fractionalization." Unpublished manuscript, Harvard University.

Alesina, Alberto, R. Baqir and William Easterly. 1997. "Public Goods and Ethnic Divisions." *Quarterly Journal of Economics* 114(4):1243–84.

Atlas Narodov Mir. 1964. Moscow: Glavnoe upravlenie geodezii i kartografii.

Barro, Robert and Rachel McCleary. 2002. "Religion and Political Economy in an International Panel." Unpublished manuscript, Harvard University, August 6.

Chandra, Kanchan. 2000. "Counting Heads: A Theory of Voting in Patronage Democracies." Unpublished paper, Harvard University.

Collier, David and James E. Mahon. 1993. "Conceptual 'Stretching' Revisited." *American Political Science Review* 87(4):845–855.

Collier, Paul and Anke Hoeffler. 2001. "Greed and Grievance in Civil War." World Bank, DECRG.

Cox, Gary W. 1997. *Making Votes Count*. New York: Cambridge University Press.

Dudley, Ryan and Ross A. Miller. 1998. "Group Rebellion in the 1980s." *Journal of Conflict Resolution* 42(1):77–96.

Easterly, William and Ross Levine. 1997. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *Quarterly Journal of Economics* 112(4):1203–50.

Fearon, James D. and David D. Laitin. 1999. "Weak States, Rough Terrain, and Large-Scale Ethnic Violence since 1945." Presented at the Annual Meetings of the American Political Science Association, Atlanta, GA, 2-5 September 1999.

Fearon, James D. and David D. Laitin. 2000. "Ordinary Language and External Validity." Paper presented at the Annual Meetings of the American Political Science Association, Washington, D.C., September 2000.

Fearon, James D. and David D. Laitin. 2002. "Ethnicity, Insurgency, and Civil War." *American Political Science Review*. forthcoming.

Greenberg, Joseph H. 1956. "The Measurement of Linguistic Diversity." *Language* 32:109–105.

Grimes, Joseph E. and Barbara F. Grimes. 1996. *Ethnologue: Languages of the World*. Thirteenth ed. Dallas, TX: Summer Institute of Linguistics.

Gurr, Ted Robert. 1993. "Why Minorities Rebel: A Global Analysis of Communal Mobilization and Conflict since 1945." *International Political Science Review* 14(2):161–201.

Gurr, Ted Robert. 1996. "Minorities at Risk III Dataset: User's Manual." CIDCM, University of Maryland.

Gurr, Ted Robert and Will H. Moore. 1997. "Ethnopolitical Rebellion: A Cross-Sectional Analysis of the 1980s with Risk Assessments for the 1990s." *American Journal of Political Science* 41(4):1079–1103.

Hibbs, Douglas A. 1973. *Mass Political Violence*. New York: Wiley.

Horowitz, Donald L. 1985. *Ethnic Groups in Conflict*. Berkeley, CA: University of California Press.

Huntington, Samuel. 1996. *The Clash of Civilizations and the Remaking of World Order*. New York: Simon and Shuster.

King, Gary and Langche Zeng. 2001. "Explaining Rare Events in International Relations." *International Organization* 55(3):693–715.

Laitin, David D. 2000. "What is a Language Community?" *American Journal of Political Science* 44(1):142–55.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Levinson, David. 1998. *Ethnic Groups Worldwide*. Phoenix: Oryx Press.

Lindstrom, Ronny and Will H. Moore. 1995. "Deprived, Rational, or Both? 'Why Minorities Rebel' Revisited." *Journal of Political and Military Sociology* 23:167–190.

Morrison, Donald, Robert Mitchell and John Paden. 1989. *Black Africa: A Comparative Handbook*. 2nd ed. New York: Paragon House.

Nobles, Melissa. 2000. *Shades of Citizenship*. Stanford: Stanford University Press.

Powell, G. Bingham. 1982. *Contemporary Democracies: Participation, Stability, and Violence*. Cambridge: Harvard University Press.

Przeworski, Adam, Michael E. Alvarez, Jose A. Cheibub and Fernando Limongi. 2001. *Democracy and Development : Political Institutions and Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.

Reilly, Benjamin. 2000/01. "Democracy, Ethnic Fragmentation, and Internal Conflict." *International Security* 25(3):162–85.

Roeder, Phillip G. 2002. "Ethnolinguistic Fractionalization (ELF) Indices, 1961 and 1985." http://weber.ucsd.edu/ proeder/data.htm [accessed 5/9/02].

Scarritt, James R. and Shaheen Mozaffar. 1999. "The Specification of Ethnic Cleavages and Ethnopolitical Groups for the Analysis of Democratic Competition in Africa." *Nationalism and Ethnic Politics* 5(1):82–117.

Taylor, Charles Lewis and David A. Jodice. 1983. *World Handbook of Political and Social Indicators*. 3rd ed. New Haven: Yale University Press.

VanHanen, Tatu. 1999. "Domestic ethnic conflict and ethnic nepotism: A comparative analysis." *Journal of Peace Research* 36(1):55–73.

Table 1
Descriptive Statistics on Ethnic Groups larger than
1% of country population, by Region

	World	West[a]	NA/ME	LA/Ca	Asia	EE/FSU	SSA[b]
# countries	160	21	19	23	23	31	43
% total		.13	.12	.14	.14	.19	.27
# groups	819	68	70	81	109	141	350
% total		.08	.09	.1	.13	.17	.43
Groups/country	5.11	3.24	3.68	3.52	4.74	4.55	8.14
Std. Dev.	3.54	2.41	2.14	.93	3.62	2.5	5.04
Max. # groups	23[c]	9	9	6	13	12	23
Min. # groups	0[d]	1	1	2	0	1	2
Avg. pop. share of largest group	.65	.85	.68	.69	.72	.73	.42
Avg. pop. share of 2nd largest	.17	.09	.19	.21	.16	.15	.2
% countries with a group $\geq$ 50%	.72	1.00	.84	.78	.82	.90	.30
% countries with a group $\geq$ 90%	.18	.57	.21	.17	.18	.13	0

Notes: [a]Includes Australia, New Zealand, and Japan. [b]Includes Sudan. [c]Tanzania. [d]Papua New Guinea is coded as having no ethnic groups that meet the 1% threshold.