Ethno-Linguistic Fractionalization: Dataset Review

Michael D. Driessen
Department of Political Science
University of Notre Dame
Notre Dame, IN 46556
mdriesse@nd.edu

Published in ***APSA Comparative Politics Newsletter*** (Winter, 2008)

Ethno-Linguistic Fractionalization
Dataset Review

Over the past few years, measures of ethnic fractionalization have been increasingly employed as a "standard" control variable in comparative politics. Fractionalization has been used to explain variation in levels of cross-national economic growth, corruption, levels of democracy, quality of governance, provision of public goods and conflict propensity (Easterly and Levine 1997, La Porta et al 1999). Agreeing on a standard index of ethnic fractionalization which can be consistently used, therefore, ought to be a pressing research project for the field. This article reviews the current state of the data on ethnic fractionalization and recommends a more concerted effort to create a long-term ethno-fractionalization dataset project.

While ethnic fractionalization is being increasingly used as an explanatory and control variable in data regressions, ethnicity as a theoretical concept has never been more contested. Despite the use of more standard datasets, such as Fearon (2003) and Alesina, Devleeschauer, Easterly, Kurlat and Wacziarg (2003), scholars continue to wonder how exact "ethnicity" can be defined as a theoretical concept, and whether it should even be used in our theoretical constructs at all (Chandra 2006). Even the authors of two of the best collections of ethnic datasets reviewed here cast theoretical doubts on their ethnic constructs. Alesina et al (2003) recognize ethnicity as a "vague and amorphous concept" and Fearon (2003) writes that it is "inherently slippery". There are several fundamental quandaries which must be addressed in the construction of a dataset on ethnicity, and

their resolution depends in large part on what one thinks "ethnicity" as a variable is capturing. The ethnic fractionalization datasets under review each respond to the basic quandaries by measuring ethnicity in a slightly different way, producing slightly different fractionalization indices, which, I argue, do not always measure up.

The first and most fundamental difficulty in constructing a dataset on ethnicity is deciding on who counts as an ethnic group. There are several reasonably objective attributes which can be used to distinguish ethnic groups from one another in a consistent way, such as language, race, religion, tribe, descent, nationality, or some combination of them. While one method, discussed below, is to disaggregate ethnic fractionalization indices into different attributes, the classic method is to create one ethnic fractionalization measure which differentiates ethnic groups by the "home" languages that they speak, the assumption being that language ties are the most salient feature that sets ethnic groups apart from one another.

If this is the agreed-on attribute which best distinguishes ethnicity, then it would seem natural for scholars to use the ethno-linguistic fractionalization index which does the best job counting up the world's languages, such as the *Ethnologue* index, arguably the most comprehensive and fine-toothed distinguisher of languages spoken in the world. The Ethnologue is compiled by anthropologists, linguists and geologists and lists over 6,800 distinct languages. However, no political scientist uses the Ethnologue's listing by itself as a measure of ethnic fractionalization. The case of Papua New Guinea, in some sense,

is the reason why. By the Ethnologue's counting, Papua New Guinea is the most ethno-linguistically diverse country of the world, with the greatest density of distinct languages per population. The problem is that such a "perfectly" fractionalized country has little relevance on political organization. While some scholars disagree with the choice, (see Reilly 2000) most political scientists are really only interested in ethnic groups which have some sort of political meaning. Conceivably, one could ethnically divide up the United States by family names and count up the numbers of Smiths, Jones, Johnsons and Browns. While an interesting exercise, this would not tell us much about how different groups of last names potentially aggregate into political decisions. So it is with language groups in Papua New Guinea which are so thoroughly disaggregated that it is nearly unthinkable that they would ever form the basis for political mobilization.

The difficulty, then, is in choosing which language groups do matter. For a first cut, Fearon (2003) chooses to count only ethno-linguistic groups which make up more than 1% of their perspective national population. Then, to deal with the problem that not all languages are equally distinct, he designs a way of categorizing these ethno-linguistic groups according to linguistic-cultural distance. Using his scale, a country with a population equally divided between Arabic and English speakers is of a higher fractionalizational order than one similarly divided between French and Italian speakers. In doing so, Fearon (2003) follows Scarritt and Mozaffar (1999) and Barrett (1982) before them in categorizing differing levels of ethnic aggregation. Scarritt and Mozaffar's index (1999) offer three levels of ethno-linguistic fractionalization. The first level

considers only those countries where two ethnic groups are politicized at the national level. The second level of aggregation considers "middle level" aggregations of ethnicity and the third level disaggregates significant ethno-political cleavages which divide up the middle level. Barrett (1982) offers 11 progressively detailed levels of aggregation, which, like a biological classification, starts with 5 races, splits into several skin colors, geographical races and sub-races, and is then followed by ethno-linguistic families (71), peoples (432), constituent peoples (8,990), languages/sub-peoples (7,010), dialects (17,000) and so on. As with many of databases, these illustrate typical problems of transparency and objectivity in coding which ethno-linguistic groups count and at what level. Barrett and his team, for example, are specifically putting together a *World Christian Encyclopedia*, and they claim to count languages as distinct only if those languages "need a separate bible translation" for evangelization purposes. While Barrett (1982) and other indices like his are impressively thorough, there are many subjective decisions adding up the languages, races and peoples in their various categories and one has to be skeptical about how meaningful some of these numbers are, and thus, the data analysis which are done on top of them.

As Barrett's (1982) classification implies, ethnicity does not only refer to language, but to race, color and often religion as well. While in places like Africa language seems to be the best way of dividing peoples into categories, in the United States or in South America, most people speak the same language even thought they may simultaneously belong to many different racial and religious categories. Some authors, therefore, add

seperate race and religion categories to their ethno-linguistic indexes. Alesina et al (2003), for example, include racial and religious indices, and Annett's (2001) index also includes a religious subcategory alongside ethno-linguistic ones.

An alternative method to counting which ethnic groups could be potentially politically relevant is to only count which ethnic groups are actively politically relevant. Thus, instead of distinguishing between race, religion or language, or attempt to make rough decisions based on anthropological categories, or count groups only above a certain threshold, Posner (2004) decides to count only the relevant share of ethnic groups he decides are politically relevant. His Politically Relevant Ethnic Group (PREG) fractionalization index codes ethnic groups which are currently, politically active and mobilized and he calculates the index in two time periods in order to track changes over time. For now, his dataset is available only with respect to sub-Saharan Africa, but could be potentially calculated worldwide.

With all these datasets, an additional problem has to do with sources used. Many of the earlier fractionalization measures, such as Hudson and Taylor (1972), Mueller and Murrell (1986), Easterly and Levine (1997), and La Porta et al (1999), used the infamous *Atlas Nirodov Mira (1964)*, a Cold War-era index created by Soviet ethnologists. This index was both discredited because it seemed to mis-specify certain ethnic groups (famously making no distinction between Hutu and Tutsi in Rwanda) and because ethnicity changes over time, in population shares and political significance. As Fearon

(2003) points out, while most political science scholars until the early 1990s considered Somalia to be ethnically homogenous, this is no longer the case, as a lower order of ethnicity has taken on saliency and Somalis distinguish themselves more often by these more disaggregated categories. While some of the newer datasets, Alesina et al (2003), Fearon (2003), Posner (2004), and Annett (2001) try to overcome this problem, they often end up using the same data sources as a baseline. Both Fearon (2003) and Alesina et al (2003) use the *Encyclopedia Britannica*, the *CIA World Factbook* and the Ethnologue project. Annett (2001) uses Barrett (1982) who also draws off of the Ethnologue project and the Encyclopedia Britannica. Posner (2004) translates and then uses the Atlas Nirodov Mira (1964) along with Morrison's (1989) *Black Africa: A Comparative Handbook*, which Fearon (2003) also references. All of these scholars supplement their missing "gaps" with newly available census data. However, despite using similar sources, the fractionalization measures vary with some significance. While they are generally close, as Posner (2004) reports for sub-Saharan Africa, Alesina et al (2003) correlates with Fearon (2003) at 0.73 and both of them correlate with Posner at under 0.54. Overall, Alesina et al (2003) correlates with Annett at 0.88. While these correlations are not bad, considering that they collectively employ similar sources and methods, the result is somewhat disappointing.

As most of these scholars recognize, ethnic fractionalization measures are still fairly weak. The variance between indices points out the need to have a more standardized and regular counting of ethnic fractionalization measures which can account for different

kinds of ethnicity, as well as change in ethnicity over time. One could theoretically begin constructing a standardized, multi-dimensional ethnic fractionalization measure using worldwide survey data based on ethnicity, language, tribe, religion and race, as Fearon (2003) suggests. This would allow scholars to classify ethnic groups over time according to how individuals categorize themselves with varying levels of significance for their multiple identities. There is some current work in this direction. Lind (2007) has recently used opinion polls to classify groups by size and distinctive cultural distance in the United States, while Dowd and Driessen (forthcoming) use Afrobarometer data to measure ethnic fractionalization by levels of ethnic voting in sub-Saharan Africa. Innovative work by Cederman and Giradin (2007) try to get at differences in kind by introducing a weighted measure in their fractionalization index for the ethnic group in power (in Eurasia and N. Africa)[1]. While all are innovative, these pieces still lack cross-national generalizability.

For range of data, rigor of method, and ease with which scholars can calculate ethnic fractionalization several different ways, Fearon (2003) and Alesina et al (2003) remain the standard, covering 160 and 190 countries, respectively, with each index measured in

[1] Cedermn and Giradin (2007) also attempt to overcome the shortcoming of the Herfindahl probability index, Ethno-linguistic Fractionalization = $1-\sum_{i=1}^{n} p_i^2$, which all these indices use. The index measures the probability that two randomly selected individuals from the entire population will be from different ethnic groups. However, as Fearon (2003) points out, the Herfindahl index cannot distinguish differences in kind of national ethnic structure. In nearly identitically fractionalized countries of 0.75, one country could have four ethnic groups of equal size and the other have one large ethnic group at 48% of the population and many smaller ones at 0.01%. Since at least Horowitz (1985) on, scholars have suspected that these two different national ethnic configurations have vastly different implications for national political outcomes.

at least 3 different ways. As the authors acknowledge, these datasets should not be the last word on ethnic fractionalization. Until a more systematic attempt is made to regularly measure ethnic fractionalization, however, their indices are a great help to scholars trying to pick apart the relationships between ethnicity and politics.

References:

Alesina et al. (2003). "Public Goods and Ethnic Divisions," *Journal of Economic Growth*. 8(2) 155-194.

Annett, A. (2001). " Social Fractionalization, Political Instability, and the Size of Government," *IMF Staff Papers*, 48(3), 561-592.

Atlas Narodov Mira. (1964). Moscow: Glavnoe Upravlenie Geodezii I Kartografii.

Barrett, D. (1982). *World Christian Encyclopedia.* New York: Oxford University Press.

Cederman, L. and L. Girardin. (2007). "Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies," *American Political Science Review.* 101(1) 173-185.

Central Intelligence Agency. (2000). *CIA World Factbook.* Washington DC: CIA Office of Public Affairs.

Chandra, K. (2006). "What is Ethnic Identity and Does it Matter?" *Annual Review of Political Science*, 9, 397-424.

Dowd, R. and M. Driessen. "Ethnically Dominated Party Systems and the Quality of Democracy: Evidence from Sub-Saharan Africa," Forthcoming Manuscript.

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

Encyclopedia Britannica. (2000). Chicago: Encyclopedia Britannica.

Fearon, J. (2003). "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth*. 8(2) 195-122.

Grimes, J. and B. Grimes. (1996). *Ethnologue: Languages of the World,* 13 edn. Dallas, TX: Summer Institute of Linguistics.

Horowitz, D. (1985). *Ethnic Groups in Conflic.* Berkeley, CA: University of California, Press.

La Porta, R, et al. (1999). "The Quality of Government," *Journal of Law, Economics and Organization* 15(1), 222-279.

Lind, J. (2007). "Fractionalization and Inter-Group Differences," *Kyklos.* 60(1) 123-139.

Morrison, D., R. Mitchell, and J. Paden. (1989). *Black Africa: A Comparative Handbook,* 2nd edn. New York: Paragon House.

Mueller, D. and P. Murrell (1986). "Interest Groups and the Size of Government," *Public Choice.* 48, 125-145.

Posner, D. (2004). "Measuring Ethnic Fractionalization in Africa," *American Journal of Political Science,* 48(4), 849-864.

Reilly, B. (2000). "Democracy, Ethnic Fragmentation, and Internal Conflict: Confused Theories, Faulty Data, and the 'Crucial Case' of Papua New Guinea," *International Security,* 25(3) 162-185.

Scarritt, J. and S. Mozaffar. (1999). "The Specification of Ethnic Cleavages and Ethnopolitical Groups for the Analysis of Democratic Competition in Africa," *Nationalism and Ethnic Politics* 5, 82-117.

Taylor, C. and M. Hudson (1972). *World Handbook of Political and Social Indicators*, 2nd edn. New Haven: Yale University Press.