A Generalized Index of Ethno-Linguistic Fractionalization*

Walter Bossert
Département de Sciences Economiques and CIREQ, Université de Montréal
walter.bossert@umontreal.ca

Conchita D'Ambrosio
Università di Milano-Bicocca and DIW Berlin
conchita.dambrosio@unibocconi.it

Eliana La Ferrara
Università Bocconi and IGIER
eliana.laferrara@unibocconi.it

June 2005

Abstract. The goal of this paper is to characterize a *generalized ethno-linguistic fractionalization index* that combines information on population shares with information on similarities among groups and individuals. We show that the generalized index is a natural extension of the *ethno-linguistic fractionalization index* and is also simple to compute. The paper offers some empirical illustrations on how the new index can be operationalized and what difference it makes as compared to standard indices. These applications pertain to the pattern of ethnic diversity in the United States across states and over time. *Journal of Economic Literature* Classification Nos.: C43, D63.

Keywords: Diversity, Ethno-Linguistic Fractionalization, Herfindahl Index.

* We thank Università Bocconi for its hospitality during the preparation of this paper. Financial support from the Polarization and Conflict Project CIT-2-CT-2004-506084 funded by the European Commission-DG Research Sixth Framework Programme and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged. We thank Joan Maria Esteban, Itzhak Gilboa, Debraj Ray for comments and suggestions.

1 Introduction

The traditional way of conceiving heterogeneity among individuals in Economics has been to think of income inequality, that is, individuals' differences in the command over economic resources. Many contributions have estimated the effects of inequality on all sorts of outcomes, and the literature on the measurement of inequality has proceeded on a parallel path, advancing to substantial degrees of sophistication. In recent times there has been a growing interest within Economics in the role that other types of heterogeneity, namely *ethnic or cultural diversity,* play in explaining socio-economic outcomes. A number of empirical studies have found that ethnic diversity is associated with lower growth rates (Easterly and Levine, 1997), more corruption (Mauro, 1995), lower contributions to local public goods (Alesina, Baqir and Easterly, 1999), lower participation in groups and associations (Alesina and La Ferrara, 2000) and a higher propensity to form jurisdictions to sort into homogeneous groups (Alesina, Baqir and Hoxby, 2004). For an extensive review of these and other contributions on the relationship between ethnic diversity and economic performance, see Alesina and La Ferrara (2005). Yet the literature on the measurement of ethnic — and other forms of non-income related — heterogeneity has received considerably less attention.

The measure of ethnic diversity used almost universally in the empirical literature is the so-called index of ethno-linguistic fractionalization (ELF), which is a decreasing transformation of the Herfindahl concentration index. In particular, if we consider a society composed of $K \geq 2$ different ethnic groups and let p_k indicate the share of group k in the total population, the resulting value of the ELF index is given by

$$1 - \sum_{k=1}^{K} p_k^2.$$

The popularity of this index in empirical applications can be attributed to two features. First, it is extremely simple to compute from micro as well as from aggregate data: all that is needed is the vector of shares of the various groups in the population. Second, ELF has a very intuitive interpretation: it measures the probability that two randomly drawn individuals from the overall population belong to different ethnic groups. On the other hand, the economic underpinnings for the use of this index seem underdeveloped. One of the few contributions that address this issue is Vigdor (2002), who proposes a behavioral interpretation of ELF in a model where individuals display differential altruism. He assumes that an individual's willingness to spend on local public goods depends partly on the benefits that other members of the community derive from the good, and that

the weight of this "altruistic" component varies depending on how many members of the community share the same ethnicity of that individual.

The implicit contention is often that different ethnic groups may have different preferences, and this would generate conflicts of interests in economic decisions. If this is the rationale for including ethnic diversity effects, then measuring diversity purely as a function of population shares seems a severe limitation. Presumably, people of different ethnicities will feel differently about each other depending on how similar they are. Similarity between groups could depend, for example, on other dimensions such as income, educational background, employment status, just to mention a few. If preferences might be induced by these other individual characteristics, then considering similarities between ethnic groups will give a better understanding of the potential conflict in economic decisions. Providing a measure of "fractionalization" that accounts for the degree of similarity among ethnic groups (for instance, similarity according to other characteristics) seems therefore a useful task.

The goal of this paper is to characterize a *generalized ethno-linguistic fractionalization index* ($GELF$) that combines information on population shares with information on similarities among groups. We show that the generalized index is a natural extension of ELF and is also simple to compute. The paper offers some empirical illustrations on how $GELF$ can be operationalized and what difference it makes as compared to the standard ELF index. These applications pertain to the pattern of ethnic diversity in the United States across states and over time. To our knowledge, this is the first attempt to propose a multi-dimensional fractionalization index. A similar aim drives Desmet, Ortuño-Ortín and Weber (2005). We will discuss the differences between the two approaches in the next paragraphs.

In the characterization we choose to be as general as possible and do not impose the group partition from the outset. In the discussion, though, we refer to ethnic groups to be in line with the strand of the literature to which we aim at contributing. Similarly, the empirical application makes use of these ethnic groups for comparison's purposes with more standard indices. But the way we think of the problem to be modelled is without such partition. Our reasoning proceeds as follows: imagine a society composed of individuals with personal characteristics, whatever they could be. Pick any two individuals, they could be perfectly identical according to the analized characteristics, completely dissimilar or similar to different degrees. For simplicity, normalize the similarity values to be in the interval $[0, 1]$ and assign the value one to perfect similarity, zero to maximum dissimilarity. Imagine to make these comparisons for all the possible couples of individuals. If the society

is composed of n individuals, the comparison process will generate n^2 similarity values. Stuck all those values in a matrix, the *similarity matrix*, where in each raw is contained the similarity value of one given individual with respect to all the others. Naturally, the main diagonal of such a matrix will have all entries equal to one – each individual is perfectly identical to itself; at the same time the matrix will be symmetric – the similarity value between, say, individual i and j, is equal to that between j and i. What is then a groups in this framework? Individuals belonging to the same group are identical in all respects. In our setting they have a similarity value of one among themselves *and* present the same degree of similarity with respect to all other individuals. With this process the group partition will emerge naturally from the similarity matrix without having to impose it in advance.

The generalized index that we are proposing could be applied to various areas in Economics. It is, indeed, an index of diversity, the complement to one of an index of concentration. And as the Herfindahl concentration index has a widespread use in areas spanning from academic research to antitrust regulation, so could be the complement to one of $GELF$. For example, since 1992 the US Department of Justice has used the Herfindahl index as a measure of market concentration to enforce antitrust regulation. According to the DOJ Horizontal Merger Guidelines of 1992, markets with an index of 0.18 or more should be considered "concentrated".

Our paper is related to several strands of the literature. First, it naturally relates to the above mentioned literature on ethnic diversity and its economic effects. While the bulk of this literature does not focus on the specific issue of measurement, a few contributions do. As the majority of applications have used language as a proxy for ethnicity, some authors have criticized the use of ELF on the grounds that linguistic diversity may not correspond to ethnic diversity. Among these, Alesina, Devleeschauwer, Easterly, Kurlat and Wacziarg (2003) have proposed a classification into groups that combines information on language with information on skin color. Note that this approach differs from ours because it aggregates information on the two dimensions *ex ante*, that is, it directly defines ethnic categories on the basis of two criteria (language and skin color) and then applies the ELF formula to the resulting number of groups. The relationship between this type of approach and ours is discussed more in depth in Subsection 2.3 below. Other authors, in particular Fearon (2003), have criticized standard applications of ELF on the grounds that they would fail to account for the *salience* of ethnic distinctions in different contexts. For example, the same two ethnic groups may be allies in one country and opponents in another, and using simply their shares in the population would fail to capture this.

We share Fearon's concerns on this point, and indeed we hope that our index can be a first step towards incorporating issues of salience in the measurement of diversity, albeit in a simplistic way. In particular, if one thinks that differences in income, or education, or any other measurable characteristic may be the reason why ethnicity matters only in certain contexts, our $GELF$ index already "weighs" ethnic categories by their salience. Turning to the notion of "distance" among ethnic groups, relatively little has been done. Using a heuristic approach, Laitin (2000) and Fearon (2003) rely on measures of distance between languages to assess how different linguistic groups are across countries. Caselli and Coleman (2002) stress the importance of ethnic distance in a theoretical model and propose to measure it using surveys of anthropologists.

Second, the paper relates to the literature on ethnic polarization. Montalvo and Reynal-Querol (2005) proposed an index of ethnic polarization, RQ, as a more appropriate measure of conflict than ELF itself. RQ aims at capturing how far is the distribution of the ethnic groups from the bipolar distribution, which represents the highest level of polarization. The authors also show that this index is highly correlated with ELF at low levels, uncorrelated at intermediate levels and negatively correlated at high levels. Desmet, Ortuño-Ortín and Weber (2005) focus on ethno-linguistic conflict that arises between a dominant central group and peripheral minority groups. To this aim the authors propose an index of peripheral ethno-linguistic diversity, PD, which can capture both the notion of diversity and of polarization. We will discuss the differences between the indices in Subsection 2.3.

Third, the measurement of diversity has been formally analyzed in different contexts as well. For example, Weitzman (1992) suggests an index that is primarily intended to measure biodiversity. Moreover, the measurement of diversity has become an increasingly important issue in the recent literature on the ranking of opportunity sets in terms of freedom of choice, where opportunity sets are interpreted as sets of options available to a decision maker. Examples for such studies include Pattanaik and Xu (2000), Nehring and Puppe (2002) and Bossert, Pattanaik and Xu (2003). A fundamental difference between the above-mentioned contributions and the approach followed in this paper is the informational basis employed which results in a very different set of axioms that are suitable for a measure of diversity. Both Weitzman's (1992) seminal paper and the literature on incorporating notions of diversity in the context of measuring freedom of choice proceed by constructing a ranking of *sets* of objects (interpreted as sets of species in the case of biodiversity and as sets of available options in the context of freedom of choice), whereas we operate in an informationally richer environment: not only whether

or not a group is present may influence the measure of diversity, but also the relative population shares of these groups along with the pairwise similarities among them.

The remainder of the paper is organized as follows. Section 2 contains our main theoretical results, namely, the axiomatic characterization of $GELF$. Section 3 provides some empirical illustrations of the different measures. Section 4 concludes with a summary of our results and possible extensions.

2 Similarity and fractionalization

In this section we propose a characterization of ELF and of its generalizations that take into account degrees of similarity between groups and individuals.

Let $\mathbb{N}$ denote the set of positive integers and $\mathbb{R}$ the set of all real numbers. The set of all non-negative real numbers is $\mathbb{R}_+$ and the set of positive real numbers is $\mathbb{R}_{++}$. For $n \in \mathbb{N} \setminus \{1\}$, $\mathbb{R}^n$ is Euclidean n-space and Δ^n is the n-dimensional unit simplex. Furthermore, $\mathbf{0}^n$ is the vector consisting of n zeroes. A *similarity matrix of dimension* $n \in \mathbb{N} \setminus \{1\}$ is an $n \times n$ matrix $S = (s_{ij})_{i,j \in \{1,\ldots,n\}}$ such that:

(a) $s_{ij} \in [0,1]$ for all $i, j \in \{1, \ldots, n\}$;
(b) $s_{ii} = 1$ for all $i \in \{1, \ldots, n\}$;
(c) $[s_{ij} = 1 \Rightarrow s_{ik} = s_{kj}]$ for all $i, j, k \in \{1, \ldots, n\}$.

The three restrictions on the elements of a similarity matrix have very intuitive interpretations. (a) is consistent with a normalization requiring that complete dissimilarity is assigned a value of zero and full similarity is represented by one. Clearly, this requires that each individual has a similarity value of one when assessing the similarity to itself, as stipulated in (b). Finally, (c) requires that if two individuals are fully similar, it is not possible to distinguish between them as far as their similarity to others is concerned. Because $i = j$ is possible in (c), the conjunction of (b) and (c) implies that a similarity matrix is symmetric, a self-evident requirement. Finally, (c) implies that full similarity is transitive in the sense that, if $s_{ij} = s_{ji} = s_{jk} = s_{kj} = 1$, then $s_{ik} = s_{ki} = 1$ for all $i, j, k \in \{1, \ldots, n\}$.

Let $\mathcal{S}^n$ be the set of all n-dimensional similarity matrices, where $n \in \mathbb{N} \setminus \{1\}$. We use I^n to denote the $n \times n$ identity matrix and $\mathbf{1}^n$ to denote the $n \times n$ matrix all of whose entries are equal to one. Clearly, both of these matrices are in $\mathcal{S}^n$, and they represent extreme cases within this class. I^n can be thought of as having maximal diversity: any

two individuals are completely dissimilar and, therefore, each individual is in a group by itself. $\mathbf{1}^n$, on the other hand, represents maximal concentration (and, thus, minimal diversity) because there is but a single group in the population all members of which are fully similar.

We let $\mathcal{S} = \cup_{n\in\mathbb{N}\setminus\{1\}}\mathcal{S}^n$, and a *diversity measure* is a function $D\colon \mathcal{S} \to \mathbb{R}_+$. The measure we suggest in this paper is what we call the *generalized ethno-linguistic fractionalization* ($GELF$) index G. It is defined by

$$G(S) = 1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$ (or any positive multiple; clearly, multiplying the index value by $\alpha \in \mathbb{R}_{++}$ leaves all diversity comparisons unchanged).

As an example, suppose a three-dimensional similarity matrix is given by

$$S = \begin{pmatrix} 1 & 1/2 & 1/4 \\ 1/2 & 1 & 0 \\ 1/4 & 0 & 1 \end{pmatrix}.$$

The corresponding value of G is given by

$$G(S) = 1 - \frac{1}{9}\left[1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{2} + 1 + 0 + \frac{1}{4} + 0 + 1\right] = 1 - \frac{1}{2} = \frac{1}{2}.$$

Before providing a characterization of our new index, we illustrate that it is indeed a generalization of the commonly-employed ethno-linguistic fractionalization (ELF) index. The application of ELF is restricted to an environment where the only information available is the vector $p = (p_1, \ldots, p_K) \in \Delta^K$ of population shares for $K \in \mathbb{N}$ predefined groups. No partial similarity values are taken into consideration—individuals are either fully similar or completely dissimilar, that is s_{ij} can assume uniquely the values one and zero respectively. Letting $\Delta = \cup_{K\in\mathbb{N}}\Delta^K$, the ELF index $E\colon \Delta \to \mathbb{R}_+$ is defined by letting

$$E(p) = 1 - \sum_{k=1}^{K} p_k^2$$

for all $K \in \mathbb{N}$ and for all $p \in \Delta^K$. Thus, ELF is one minus the well-known Herfindahl index of concentration.

In our setting, ELF environment can be described by a subset $\mathcal{S}_{01} = \cup_{n\in\mathbb{N}\setminus\{1\}}\mathcal{S}_{01}^n$ of our class of similarity matrices where, for all $n \in \mathbb{N} \setminus \{1\}$, for all $S \in \mathcal{S}_{01}^n$ and for all $i, j \in \{1, \ldots, n\}$, $s_{ij} \in \{0, 1\}$. By properties (b) and (c), it follows that, within this

subclass of matrices, the population $\{1,\ldots,n\}$ can be partitioned into $K \in \mathbb{N}$ non-empty and disjoint subgroups $N_1,\ldots,N_K$ with the property that, for all $i,j \in \{1,\ldots,n\}$,

$$s_{ij} = \begin{cases} 1 & \text{if there exists } k \in \{1,\ldots,K\} \text{ such that } i,j \in N_k; \\ 0 & \text{otherwise.} \end{cases}$$

Letting $n_k \in \mathbb{N}$ denote the cardinality of N_k for all $k \in \{1,\ldots,K\}$, it follows that $\sum_{k=1}^{K} n_k = n$ and $p_k = n_k/n$ for all $k \in \{1,\ldots,K\}$. For $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}_{01}^n$, we obtain

$$G(S) = 1 - \frac{1}{n^2}\sum_{k=1}^{K} n_k^2 = 1 - \sum_{k=1}^{K} p_k^2 = E(p).$$

For example, suppose that

$$S = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

that is we are analyzing a society composed of three individuals, two of which fully similar – hence belonging to the same group – the latter being completely dissimilar to everybody. The corresponding value of G is given by

$$G(S) = 1 - \frac{1}{9}\left[1+1+0+1+1+0+0+0+1\right] = 1 - \frac{5}{9} = \frac{4}{9}.$$

Because $S \in \mathcal{S}_{01}^3$, we can alternatively calculate this diversity value using ELF. We have $K = 2$, $N_1 = \{1,2\}$, $N_2 = \{3\}$, $p_1 = 2/3$ and $p_2 = 1/3$. Thus,

$$E(p) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 1 - \frac{5}{9} = \frac{4}{9} = G(S).$$

A second special case allows us to obtain population subgroups endogenously from similarity matrices even if similarity values can assume values other than zero and one. To do so, we define a partition of $\{1,\ldots,n\}$ into $K \in \mathbb{N}$ non-empty and disjoint subgroups $N_1,\ldots,N_K$. By properties (b) and (c), these subgroups are such that, for all $k \in \{1,\ldots,K\}$, for all $i,j \in N_k$ and for all $h \in \{1,\ldots,n\}$, $s_{ij} = s_{ji} = 1$ and $s_{ih} = s_{hi} = s_{hj} = s_{jh}$. Thus, for all $k,\ell \in \{1,\ldots,K\}$, we can unambiguously define $v_{k\ell} = s_{ij}$ for some $i \in N_k$ and some $j \in N_\ell$. Again using $n_k \in \mathbb{N}$ to denote the cardinality of N_k for all $k \in \{1,\ldots,K\}$, it follows that $\sum_{k=1}^{K} n_k = n$ and $p_k = n_k/n$ for all $k \in \{1,\ldots,K\}$. For $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}^n$, we obtain

$$G(S) = 1 - \frac{1}{n^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K} n_k n_\ell v_{k\ell} = 1 - \sum_{k=1}^{K}\sum_{\ell=1}^{K} p_k p_\ell v_{k\ell}.$$

Clearly, the ELF index E is obtained for the case where all off-diagonal entries of S are equal to zero.

For example, let

$$S = \begin{pmatrix} 1 & 1 & 1/2 \\ 1 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix},$$

that is we are analyzing a society composed of three individuals, two of which fully similar – hence belonging to the same group – the latter being partially, $\frac{1}{2}$ to be precise, similar to everybody. The corresponding index value is

$$G(S) = 1 - \frac{1}{9}\left[1+1+\frac{1}{2}+1+1+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+1\right] = 1 - \frac{7}{9} = \frac{2}{9}.$$

According to the method outlined above, we can alternatively partition the population $\{1,2,3\}$ into two groups $N_1 = \{1,2\}$ and $N_2 = \{3\}$. The population shares of these groups are $p_1 = 2/3$ and $p_2 = 1/3$. We obtain the intergroup similarity values $v_{11} = v_{22} = s_{11} = s_{22} = s_{12} = s_{21} = 1$ and $v_{12} = v_{21} = s_{i3} = s_{3i} = 1/2$ for $i \in \{1,2\}$, which leads to the index value

$$G(S) = 1 - \frac{1}{9}\left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}\right] = 1 - \frac{7}{9} = \frac{2}{9}.$$

We now turn to a characterization of $GELF$. Our first axiom is a straightforward normalization property. It requires that the value of D at $\mathbf{1}^n$ is equal to zero and the value of D at I^n is positive for all $n \in \mathbb{N}\setminus\{1\}$. Given that the matrix $\mathbf{1}^n$ is associated with minimal diversity, it is a very plausible restriction to require that D assumes its minimal value for these matrices. Note that this minimal value is the same across population sizes. This is plausible because, no matter what the population size n might be, there is but a single group of perfectly similar individuals and, thus, there is no diversity at all. In contrast, it would be much less natural to require that the value of D at I^n be identical for all population sizes n. It is quite plausible to argue that having more distinct groups each of which consists of a single individual leads to more diversity than a situation where there are fewer groups containing one individual each. Thus, we obtain the following axiom.

Normalization. For all $n \in \mathbb{N} \setminus \{1\}$,

$$D(\mathbf{1}^n) = 0 \quad \text{and} \quad D(I^n) > 0.$$

Our second axiom is very uncontroversial as well. It requires that individuals are treated impartially, paying no attention to their identities. For $n \in \mathbb{N} \setminus \{1\}$, let Π^n be the set of permutations of $\{1, \ldots, n\}$, that is, the set of bijections $\pi \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$. For $n \in \mathbb{N} \setminus \{1\}$, $S \in \mathcal{S}^n$ and $\pi \in \Pi^n$, S_π is obtained from S by permuting the rows and columns of S according to π. Anonymity requires that D is invariant with respect to permutations.

Anonymity. For all $n \in \mathbb{N} \setminus \{1\}$, for all $S \in \mathcal{S}^n$ and for all $\pi \in \Pi^n$,

$$D(S_\pi) = D(S).$$

Many social index numbers have an additive structure. Additivity entails a separability property: the contribution of any variable to the overall index value can be examined in isolation, without having to know the values of the other variables. Thus, additivity properties are often linked to independence conditions of various forms. The additivity property we use is standard except that we have to respect the restrictions imposed by the definition of $\mathcal{S}^n$. In particular, we cannot simply add two similarity matrices S and T of dimension n because, according to ordinary matrix addition, all entries on the diagonal of the sum $S + T$ will be equal to two rather than one and, therefore, $S + T$ is not an element of $\mathcal{S}^n$. For that reason, we define the following operation $\oplus$ on the sets $\mathcal{S}^n$ by letting, for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S, T \in \mathcal{S}^n$, $S \oplus T = (s_{ij} \oplus t_{ij})_{i,j \in \{1,\ldots,n\}}$ with

$$s_{ij} \oplus t_{ij} = \begin{cases} 1 & \text{if } i = j; \\ s_{ij} + t_{ij} & \text{if } i \neq j. \end{cases}$$

The standard additivity axiom has to be modified in another respect. Because the diagonal is unchanged when moving from S and T to $S \oplus T$, it would be questionable to require the value of D at $S \oplus T$ to be given by the sum of $D(S)$ and $D(T)$ because, in doing so, we would double-count the diagonal elements in S and in T. Therefore, this sum has to be corrected by the value of D at I^n, and we obtain the following axiom.

Additivity. For all $n \in \mathbb{N} \setminus \{1\}$ and for all $S, T \in \mathcal{S}^n$ such that $(S \oplus T) \in \mathcal{S}^n$,

$$D(S \oplus T) = D(S) + D(T) - D(I^n).$$

With the partial exception of the normalization condition (which implies that our diversity measure assumes the same value for the matrix $\mathbf{1}^n$ for all population sizes n), the

first three axioms apply to diversity comparisons involving fixed population sizes only. Our last axiom imposes restrictions on comparisons across population sizes. We consider specific replications and require the index to be invariant with respect to these replications. The scope of the axiom is limited to what we consider clear-cut cases and, therefore, represents a rather mild variable-population requirement. In particular, consider the n-dimensional identity matrix I^n. As argued before, this matrix represents an extreme degree of diversity: each individual is in a group by itself and shares no similarities with anyone else. Now consider a population of size nm where there are m copies of each individual $i \in \{1, \ldots, n\}$ such that, within any group of m copies, all similarity values are equal to one and all other similarity values are equal to zero. Thus, this particular replication has the effect that, instead of n groups of size one that do not have any similarity to other groups, now we have n groups each of which consists of m identical individuals and, again, all other similarity values are equal to zero. As before, the population is divided into n homogeneous groups of equal size. Adopting a relative notion of diversity, it would seem natural to require that diversity has not changed as a consequence of this replication. To provide a precise formulation of the resulting axiom, we use the following notation. For $n, m \in \mathbb{N} \setminus \{1\}$, we define the matrix $R^n_m = (r_{ij})_{i,j \in \{1,\ldots,nm\}} \in \mathcal{S}^{nm}$ by

$$r_{ij} = \begin{cases} 1 & \text{if } \exists h \in \{1, \ldots, n\} \text{ such that } i, j \in \{(h-1)m + 1, \ldots, hm\}; \\ 0 & \text{otherwise.} \end{cases}$$

Now we can define our replication invariance axiom.

Replication invariance. For all $n, m \in \mathbb{N} \setminus \{1\}$,

$$D(R^n_m) = D(I^n).$$

These four axioms characterize $GELF$.

Theorem 1 *A diversity measure $D\colon \mathcal{S} \to \mathbb{R}_+$ satisfies normalization, anonymity, additivity and replication invariance if and only if D is a positive multiple of G.*

Proof. That any positive multiple of G satisfies the axioms is straightforward to verify. Conversely, suppose D is a diversity measure satisfying normalization, anonymity, additivity and replication invariance. Let $n \in \mathbb{N} \setminus \{1\}$, and define the set $\mathcal{X}^n \subseteq \mathbb{R}^{n(n-1)/2}$ by

$$\mathcal{X}^n = \{x = (x_{ij})_{\substack{i \in \{1,\ldots,n-1\} \\ j \in \{i+1,\ldots,n\}}} \mid \exists S \in \mathcal{S}^n \text{ such that } s_{ij} = x_{ij} \text{ for all } i \in \{1, \ldots, n-1\} \text{ and for all } j \in \{i+1, \ldots, n\}\}.$$

Define the function $F^n\colon \mathcal{X}^n \to \mathbb{R}$ by letting, for all $x \in \mathcal{X}^n$,

$$F^n(x) = D(S) - D(I^n) \tag{1}$$

where $S \in \mathcal{S}^n$ is such that $s_{ij} = x_{ij}$ for all $i \in \{1, \ldots, n-1\}$ and for all $j \in \{i+1, \ldots, n\}$. This function is well-defined because $\mathcal{S}$ contains symmetric matrices with ones on the main diagonal only. Because D is bounded below by zero, it follows that F^n is bounded below by $-D(I^n)$. Furthermore, the additivity of D implies that F^n satisfies Cauchy's basic functional equation

$$F^n(x+y) = F^n(x) + F^n(y) \tag{2}$$

for all $x, y \in \mathcal{X}^n$ such that $(x+y) \in \mathcal{X}^n$; see Aczél (1966, p. ??). We have to address a slight complexity in solving this equation because the domain $\mathcal{X}^n$ of F^n is not a Cartesian product, which is why we provide a few further details rather than invoking the corresponding standard result immediately.

Fix $i \in \{1, \ldots, n-1\}$ and $j \in \{i+1, \ldots, n\}$, and define the function $f_{ij}^n\colon [0,1] \to \mathbb{R}$ by

$$f_{ij}^n(x_{ij}) = F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$$

for all $x_{ij} \in [0,1]$, where the vector $(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$ is such that the component corresponding to ij is given by x_{ij} and all other entries (if any) are equal to zero. Note that this vector is indeed an element of $\mathcal{X}^n$ and, therefore, f_{ij}^n is well-defined. The function f_{ij}^n is bounded below because F^n is and, as an immediate consequence of (2), it satisfies the Cauchy equation

$$f_{ij}^n(x_{ij} + y_{ij}) = f_{ij}^n(x_{ij}) + f_{ij}^n(y_{ij}) \tag{3}$$

for all $x_{ij}, y_{ij} \in [0,1]$ such that $(x_{ij} + y_{ij}) \in [0,1]$. Because the domain of f_{ij}^n is an interval and f_{ij}^n is bounded below, the only solutions to (3) are linear functions; see Aczél (1987, p. ??). Thus, there exists $c_{ij}^n \in \mathbb{R}$ such that

$$F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1}) = f_{ij}^n(x_{ij}) = c_{ij}^n x_{ij} \tag{4}$$

for all $x_{ij} \in [0,1]$.

Let $S \in \mathcal{S}^n$. By additivity, the definition of F^n and (4),

$$F^n\left((s_{ij})_{\substack{i\in\{1,\ldots,n-1\}\\ j\in\{i+1,\ldots,n\}}}\right) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} F^n(s_{ij}; \mathbf{0}^{n(n-1)/2-1}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f_{ij}^n(s_{ij}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij}$$

and, defining $d^n = D(I^n)$ and substituting into (1), we obtain

$$D(S) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij} + d^n. \tag{5}$$

Now fix $i, k \in \{1, \ldots, n-1\}$, $j \in \{i+1, \ldots, n\}$ and $\ell \in \{k+1, \ldots, n\}$, and let $S \in \mathcal{S}^n$ be such that $s_{ij} = s_{ji} = 1$ and all other off-diagonal entries of S are equal to zero. Let the bijection $\pi \in \Pi^n$ be such that $\pi(i) = k$, $\pi(j) = \ell$, $\pi(k) = i$, $\pi(\ell) = j$ and $\pi(h) = h$ for all $h \in \{1, \ldots, n\} \setminus \{i, j, k, \ell\}$. By (5), we obtain

$$D(S) = c_{ij}^n + d^n \quad \text{and} \quad D(S_\pi) = c_{k\ell}^n + d^n,$$

and anonymity implies $c_{ij}^n = c_{k\ell}^n$. Therefore, there exists $c^n \in \mathbb{R}$ such that $c_{ij}^n = c^n$ for all $i \in \{1, \ldots, n-1\}$ and for all $j \in \{i+1, \ldots, n\}$, and substituting into (5) yields

$$D(S) = c^n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} + d^n$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Normalization requires

$$D(\mathbf{1}^n) = c^n \frac{n(n-1)}{2} + d^n = 0$$

and, therefore, $d^n = -c^n n(n-1)/2$ for all $n \in \mathbb{N} \setminus \{1\}$. Using normalization again, we obtain

$$D(I^n) = -c^n \frac{n(n-1)}{2} > 0$$

which implies $c^n < 0$ for all $n \in \mathbb{N} \setminus \{1\}$. Thus,

$$D(S) = c^n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} - c^n \frac{n(n-1)}{2} \tag{6}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$.

Let n be an even integer greater than or equal to four. By replication invariance and (6),

$$D(R_{n/2}^2) = c^n \frac{n}{2} \left(\frac{n}{2} - 1 \right) - c^n \frac{n(n-1)}{2} = -c^2 = D(I^2).$$

Solving, we obtain

$$c^n = 4 \frac{c^2}{n^2}. \tag{7}$$

Now let n be an odd integer greater than or equal to three. Thus, $q = 2n$ is even, and the above argument implies

$$c^q = 4 \frac{c^2}{q^2} = \frac{c^2}{n^2}. \tag{8}$$

Furthermore, replication invariance requires

$$D(R_2^n) = D(R_2^{q/2}) = c^q \frac{q}{2} - c^q \frac{q(q-1)}{2} = -c^n \frac{n(n-1)}{2} = D(I^n).$$

Solving for c^n and using the equality $q = 2n$, it follows that $c^n = 4c^q$ and, combined with (8), we obtain (7) for all odd $n \in \mathbb{N} \setminus \{1\}$ as well.

Substituting into (6), simplifying and defining $\alpha = -2c^2 > 0$, it follows that, for all $n \in \mathbb{N} \setminus \{1\}$ and for all $S \in \mathcal{S}^n$,

$$\begin{aligned} D(S) &= 4\frac{c^2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij} - 2\frac{c^2}{n^2}n(n-1) \\ &= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} s_{ij} - 2c^2 + 2\frac{c^2}{n} \\ &= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} s_{ij} - \frac{1}{n}\right] \\ &= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\right] \\ &= \alpha G(S). \ \blacksquare \end{aligned}$$

2.1 Discussion

In this section we discuss the differences between $GELF$ and related indices proposed in the literature, namely the indices of ethno-linguistic fractionalization (ELF), ethinic polarization (RQ), peripheral diversity (PD). ELF can be interpreted as one minus a weighted sum of population shares p_i, where the weights are these shares themselves. $GELF$ is based on a natural generalization of that idea. In the presence of similarities among different individuals belonging to different groups, the weight assigned to p_i should not be merely p_i itself but the similarities of individuals in group i to the individuals in other groups should be taken into consideration as well when constructing these weights. In $GELF$ each individual counts in two capacities. Through its membership in its *own* group, an individual contributes to the population share of the group. In addition, there is a secondary contribution via the similarities to individuals of *other* groups.

The index of ethnic polarization RQ (see Montalvo and Reynal-Querol, 2005) shares a structure similar to that of ELF and of $GELF$:

$$RQ = 1 - \sum_{k=1}^{K}\left(\frac{1/2 - p_k}{1/2}\right)^2 p_k$$

for all $p \in \Delta$. As is the case for ELF and $GELF$, RQ employs a weighted sum of population shares. The weights employed in RQ capture the deviation of each group

from the maximum polarization share 1/2 as a proportion of 1/2. As is the case for ELF, underlying that formula is the implicit assumption that any two groups are either completely similar or completely dissimilar and, thus, the weights depend on population shares only.

The index of peripheral diversity PD (see Desmet, Ortuño-Ortín and Weber, 2005) is a specification of the original Esteban and Ray (1994) polarization index. It is derived from the alienation-identification framework proposed by Esteban and Ray (1994), applied to distances between language spoken – as opposed to income distances of the original contribution – and distinguishing between the effective alienation felt by the dominant group and that of the minorities. In particular, written in the setting of our paper, the index has the following expression:

$$PD = \sum_{k=1}^{K} \left[p_k^{1+\alpha} \left(1 - s_{0k}\right) + p_k p_0^{1+\alpha} \left(1 - s_{0k}\right) \right],$$

where α is a parameter indicating the importance given to the identification component, 0 is the dominant group and the other K are minority groups. When $\alpha < 0$ PD is an index of peripheral diversity, otherwise, when $\alpha > 0$, PD is an index of peripheral polarization. The structure of this index is different from that of the previous ones. As $GELF$ it does incorporate a notion of dissimilarity between groups, given by the complement to one of the similarity value. On the other hand, as opposed to the previous indices, the identification component plays a crucial role enhancing (when $\alpha > 0$) or diminishing (when $\alpha > 0$) the alienation produced by distances between groups. The additional difference with the other indices discussed in this section is the dinstinction between the dominant groups and the minorities.

SHALL WE DELETE WHAT IS BELOW HERE UP TO THE END OF THE PARAGRAPH? I WOULD DO IT. Suppose we have information on two characteristics of the population, say, ethnicity and income, and assume for simplicity that there are only two ethnic groups A and B and two income classes H (for high) and L (for low). $GELF$ allows us to use one characteristic to construct population shares and another characteristic to compute a "similarity value" that will then be used as a weight. In doing so, an asymmetric structure can be imposed because the dimension of primary interest (the one on which shares are computed) needs to be selected on a priori grounds. An alternative approach would be to use both dimensions in a symmetric way and define "raw groups" based on the intersection of the different dimensions (as proposed by Alesina, Devleeschauwer, Easterly, Kurlat and Wacziarg, 2003).

In the above example, we would have four groups: AH, AL, BH, BL, and on these finer partition we could compute the standard ELF index. We find the latter approach less advantageous compared to ours in the sense that it imposes *independence* across the resulting set of groups. In particular, in computing ELF the raw groups AH and AL would be treated as "different" in the same way as AH and BL are. This seems undesirable because AH and AL share the same ethnicity, while AH and BL do not share any of the two dimensions. Our approach does not imply this type of independence across all dimensions.

The application of $GELF$ is not restricted to situations where the similarity values are given by a second characteristic—*any* definition of the notion of similarity can be accommodated.

3 Empirical illustrations

In this section we provide an application of $GELF$ to the pattern of ethnic diversity in the United States across states and over time.

The data sets used are the March Supplement of the Current Population Survey (CPS), and the 1990 Census. From the CPS we only use the years from 1989 to 1995 because of changes in the classification of ethnicity before and after this interval. Although the resulting picture may only be representative of the pattern of ethnic diversity in the US in the Nineties, the advantage of our choice is that in this time framework we have the moste detailed disaggregation available into racial groups. In particular, both in the CPS and in the Census the population is divided into five racial groups: (i) White; (ii) Black; (iii) American Indian, Eskimo or Aleutian; (iv) Asian or Pacific Islander; and (v) Other. The last category includes any other race except the four mentioned. For these years the CPS and the Census do not identify Hispanic as a separate racial category. However, Alesina, Baqir and Easterly (1999), who construct ELF from the same five categories, have verified that the category Hispanic (obtained from a different source) has a correlation of more than 0.9 with the category Other in the Census data.

We analyze different dimensions of similarity across ethnic groups: household income, education and employment status of the head of the household. We use the two data sources to exploit their comparative advantage in terms of cross-sectional versus time-series coverage. In particular, using Census data, we explore differences across states and we illustrate applications in which aggregate data on mean characteristics are available. Using the CPS, on the other hand, we analyze trends in diversity over time and we

illustrate a methodology that requires the availability of the entire distribution of characteristics, as is typically the case with individual survey data. We do not rely on the CPS for analyzing cross-states differences because of the small sample size of certain racial groups in several states.

3.1 *GELF* and similarity of distributions

The household income series from the CPS allows us to explore in detail the similarity among racial groups in this respect. We first estimate non-parametrically the distributions of household income by race. In particular, the estimation method applied in the paper is derived from a generalization of the kernel density estimator to take into account the sample weights, θ_k^i, attached to each observation k in each group i, namely, from the *adaptive* or *variable* kernel.

The adaptive kernel is built with a two-stage procedure: a density is determined in the first stage in order to obtain the optimal bandwidth parameter; in the second stage, the final density is computed for each race. The estimate of the density function of income for each group i, $\widehat{f^i}$, is determined directly from the data $\{y_1^i, \ldots, y_{N_i}^i\}$ of the sample of size N_i for group i, without assuming its functional form a priori. The only assumption made is that there exists a density function f^i from which the sample is extracted. More precisely, the estimated density in the first stage is

$$\widetilde{f^i}\left(y_j^i\right) = \sum_{k=1}^{N_i} \frac{\theta_k^i}{h_{N_i}} K\left(\frac{y_j^i - y_k^i}{h_{N_i}}\right)$$

for all $j \in \{1, \ldots, N_i\}$, and the final estimate is

$$\widehat{f^i}\left(y_j^i\right) = \sum_{k=1}^{N_i} \frac{\theta_k^i}{h_{N_i}\lambda\left(y_k^i\right)} K\left(\frac{y_j^i - y_k^i}{h_{N_i}\lambda\left(y_k^i\right)}\right)$$

for all $j \in \{1, \ldots, N_i\}$, where h_{N_i} is the bandwidth parameter, K is the kernel function,

$$\lambda\left(y_k^i\right) = \left\{\frac{\widetilde{f^i}\left(y_k^i\right)}{g}\right\}^{-\frac{1}{2}}$$

and g is the geometric mean of $\widetilde{f^i}\left(y_k^i\right)$. The sample weights are normalized to sum to one so that $\sum_{k=1}^{N_i} \theta_k^i = 1$.

With the estimated densities of household income by race, we measure the overlap among them implying that two racial groups whose income distribution perfectly overlaps

are considered perfectly similar. The measure of overlap of the income distributions applied is the Kolmogorov measure of variation distance

$$Kov_{ij} = \frac{1}{2} \int \left| \widehat{f}^{i}(y) - \widehat{f}^{j}(y) \right| dy.$$

The Kolmogorov measure of variation distance is a measure of the lack of overlap between groups i and j. The value of Kov_{ij} is equal to zero if $f^{i}(y) = f^{j}(y)$ for all $y \in \mathbb{R}$ and equal to its maximum one if $f^{i}(y)$ and $f^{j}(y)$ do not overlap. The distance is sensitive to changes in the distributions only when both take positive values, being insensitive to changes whenever one of them is zero. It will not change if the distributions move apart, provided that there is no overlap between them or that the overlapping part remains unchanged. The resulting measure of similarity between any two groups i and j is

$$s_{ij} = 1 - Kov_{ij}.$$

3.2 $GELF$ and similarity of means

As a second step of our empirical application, we compute a crude measure of similarity based on the expected value of the distribution of a second dimension analyzed. This is to illustrate the performance of $GELF$ in case of grouped data or poor availability of information in the data set.

We can measure similarity with respect to continuous or to categorical variables. For continuous variables such as household income, we indicate by λ^{i} the sample mean of the distribution of income of group i, and by λ the overall sample mean. Then similarity between any two groups i and j is

$$s_{ij} = 1 - \left| \frac{\lambda^{i}}{\lambda} - \frac{\lambda^{j}}{\lambda} \right|, \tag{9}$$

where the difference between the groups is measured as the distance between the means as a proportion of the overall mean to eliminate the effect that changes in the latter over the years might have. When available, sample weights are used in the computation of the sample means.

In cases where the absolute value of $\frac{\lambda^{i}}{\lambda} - \frac{\lambda^{j}}{\lambda}$ in expression (9) is greater than one for one or more pairs of groups, one can adopt a different normalization. In particular, denote by λ_{Max} the maximum *average* income among all groups in all states, and by λ_{Min} the minimum. Then we can compute s_{ij} alternatively as

$$s_{ij} = 1 - \left| \frac{\lambda^{i} - \lambda^{j}}{(\lambda_{Max} - \lambda_{Min})} \right|. \tag{10}$$

Note that expression (10) is bounded between zero and one by construction. In our empirical illustrations, we employ the formula (9) when using CPS data and (10) when using Census data. The reason is that taking Max and Min from individual pairwise comparisons in the CPS would generate unrealistically large differentials, while using mean values allows us to eliminate some noise. On the other hand, normalizing by mean US income in the application that uses Census data would result in some differentials for example, between Asian/Pacific Islanders and other groups in West Virginia) that are greater than the average income in the US, hence in negative values for (9).

For education, we create a dummy variable that assumes the value one if the head of the household has at least a high school degree (Education HS), respectively at least a bachelor degree (Education BA). For employment, the value of the dummy variable is one if the household head is not unemployed or not in the labor force (Employment 1), respectively not unemployed (Employment 2). Indicating by δ^i the sample means of these variables for group i, that is the share of the population assuming value one, similarity between any two groups i and j is

$$s_{ij} = 1 - \left|\delta^i - \delta^j\right|.$$

Again, sample weights are used in the computations for these variables.

3.3 Results

We discuss our results starting with computations based on the CPS and on the Kolmogorov measure of variation distance (for the case of income), and then turning to Census data and to similarity of means.

[Insert Figures 1 to 4]

For the CPS, the results are presented in Figures 1 to 4. The values of the indices are normalized to 100 in 1989 to facilitate the analysis of the pattern over time. In Figure 1 we plot all the indices together, including RQ, while in the following figures we isolate the series depending on the additional characteristic analyzed. The indices based uniquely on the population shares, ELF and RQ, show an increasing trend over time, steeper for ELF. With $GELF$, the inclusion of the similarities has, in the majority of the cases, the effect of changing the increasing trend to a flat one with a path exhibiting oscillations around this stable trend.

Household income has a different influence on the pattern depending on the quality of the measure of similarity computed. When the data allow the entire distribution of

income of the groups to be compared, as is the case of CPS with the Kolmogorov measure of variation distance, we observe an oscillating path with the value in 1995 being almost identical to the one at the beginning of the period. Mean income, on the other hand, is very sensitive to extreme values and the changes of the minimum and maximum incomes reported in the sample of the most numerous groups, namely white and black, have a great impact on the index.

The results obtained with inclusion of employment status of the household head mimic closely, in the first half of the period in particular, those obtained with income based on the Kolmogorov measure of variation distance. Income reported in each year is the amount of money received in the preceding calendar year, hence the downturn observed for income in 1992 according to the Kolmogorov measure is contemporaneous of the downturn in employment according to both series in 1991.

The characteristic that generates the path of $GELF$ most similar to ELF (and RQ) is Education BA, though with a somewhat higher variation around the increasing trend. Education HS, on the other hand, shares the same increasing trend, with a more pronounced cycle.

[Insert Tables 1 to 3]

We next move to cross-states comparisons in our indices built from Census data. The results are presented in Tables 1 to 3 for income, education and employment, respectively. Table 1 lists the US states in decreasing order of the standard ELF index. According to ELF, the five most fractionalized states are Hawaii, California, Washington DC, Mississippi and Louisiana; the five least fractionalized states are West Virginia, Iowa, New Hampshire, Maine and Vermont. When we "correct" this index to account for similarity in incomes, some of the rankings change dramatically. First of all, Hawaii moves from the first to the 38th most fractionalized state; secondly, California moves from the 2nd to the 11th position. In other words, while these states may be very diverse in terms of pure racial shares, the distribution of income across races is relatively more equal than elsewhere, and they turn out to be less fractionalized under $GELF$. States that decreas in the ranking under $GELF$ include New Mexico, Oregon and Washington State. On the other hand, there are states that rank relatively low under ELF but move up the ranking under $GELF$: the most notable example is Connecticut, going from the 26th position under ELF to the 15th under $GELF$. Other notable cases include New Jersey and Massachusetts.

We then consider similarity in education (Table 2), as measured by the fraction of people aged 25 and above with high school degree or more (columns 3 and 4) or bachelor's degree and above (columns 5 and 6). Again, we see some states that become less fragmented under $GELF$ and some that become relatively more fragmented, but the results are sensitive to the threshold used for educational achievement. Among the states that unambiguously become less fragmented under both measures of education are Oklahoma and Kentucky; states that are unambiguously more fragmented, on the other hand, include Illinois, Arizona, Connecticut, Colorado, Pennsylvania, Massachusetts and Wyoming.

Finally, Table 3 reports results with respect to two measures of employment. Employment1 classifies as employed all the population above 16 but those "unemployed" or "not in labor force". Employment2 classifies as employed all the population above 16 except those "unemployed". Again, re-rankings among States are somewhat sensitive to the definition adopted, but there are some clear patterns. States like California, Maryland and Delaware are places where the incidence of unemployment among races is much more similar than elsewhere, and their rank goes down when we consider $GElF$ instead of ELF. On the other hand, in states like Alaska, Illinois, Michigan, South Dakota, Montana and North Dakota, racial differences are amplified by employment differentials.

4 Concluding remarks

– Summary of results.

The *generalized ethno-linguistic fractionalization index* characterized in the paper combines information on population shares with information on similarities among groups. The concept of similarity has not been derived in the theoretical section; we assume that it is known to what degree any two groups are similar. In the application to the US we choose as dimensions of similarities across ethnic groups household income, education and employment status of the head of the household since we believe that these are important aspects of the US economy that could influence the behaviour of individuals. This need not necessarily be the case for other countries. For example, in less developed countries, it might be more important to consider the amount of natural resources, the quality of the land or a combination of characteristics. Allowing any possible concept of similarity has the advantage of letting the researcher free to pick the most appropiate in the context analyzed. In addition, and most importantly, our index allows to incorporate a multidimensional concept of similarity, as opposed to the single dimension of our

application.

– Other applications: I.O. literature.

References

[1] Alesina, Alberto, Reza Baqir and William Easterly (1999), "Public Goods and Ethnic Divisions", *Quarterly Journal of Economics*, 114, 1243–1284.

[2] Alesina, Alberto, Reza Baqir and Caroline Hoxby (2004), "Political Jurisdictions in Heterogeneous Communities", *Journal of Political Economy*, 112, 348–396.

[3] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), "Fractionalization", *Journal of Economic Growth*, 8, 155–194.

[4] Alesina, Alberto and Eliana La Ferrara (2000), "Participation in Heterogeneous Communities", *Quarterly Journal of Economics*, 115, 847–904.

[5] Alesina, Alberto and Eliana La Ferrara (2005), "Ethnic Diversity and Economic Performance", *Journal of Economic Literature*, forthcoming.

[6] Bossert, Walter, Prasanta K. Pattanaik and Yongsheng Xu (2003), "Similarity of Options and the Measurement of Diversity", *Journal of Theoretical Politics*, 15, 405–421.

[7] Caselli, Francesco and Wilbur J. Coleman (2002), "On the Theory of Ethnic Conflict", unpublished manuscript, Harvard University.

[8] Desmet, Klaus, Ignacio Ortuño-Ortín and Shlomo Weber (2005), "Peripheral Diversity and Redistribution", CEPR, Discussion Paper No.5112.

[9] Easterly, William and Ross Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, 111, 1203–1250.

[10] Esteban, Joan-Maria and Debraj Ray (1994), "On the Measurement of Polarization", *Econometrica*, 62, 819–851.

[11] Fearon, James D. (2003), "Ethnic and Cultural Diversity by Country ", *Journal of Economic Growth*, 8, 195–222.

[12] Laitin, David (2000), "What is a Language Community?", *American Journal of Political Science, 44, 142–154.*

[13] Mauro, Paolo (1995), "Corruption and Growth", *Quarterly Journal of Economics*, 110, 681–712.

[14] Montalvo, Jose G. and Marta Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict, and Civil Wars", *American Economic Review*, forthcoming.

[15] Nehring, Klaus and Clemens Puppe (2002), "A Theory of Diversity", *Econometrica*, 70, 1155–1198.

[16] Pattanaik, Prasanta K. and Yongsheng Xu (2000), "On Diversity and Freedom of Choice", *Mathematical Social Sciences*, 40, 123–130.

[17] Prakasa Rao, B.L.S. (1983), *Nonparametric Functional Estimation*, Academic Press, Orlando.

[18] Vigdor, Jacob L. (2002), "Interpreting Ethnic Fragmentation Effects", *Economics Letters*, 75, 271–76.

[19] Wand, Matt P. and M.Chris Jones (1995), *Kernel Smoothing*, Chapman & Hall, London.

[20] Weitzman, Martin (1992), "On Diversity", *Quarterly Journal of Economics*, 107, 363–405.