

A Constructivist Dataset on Ethnicity and Institutions (CDEI)

Kanchan Chandra
New York University
kanchan.chandra@gmail.com

Chapter draft prepared for Rawi Abdelal, Yoshiko Herrera, Alastair Ian Johnston and Rose
McDermott eds,
Identity as a Variable
August 22 2005

A Constructivist Dataset on Ethnicity and Institutions (CDEI)

Constructivism – the principal theoretical revolution in the study of ethnic identities over the last thirty years – has established that individuals have multiple ethnic options which a choice of which one to activate in any given context, and that the ethnic identities they activate can change over time, often endogenously to political and economic outcomes. The implication for our data collection efforts is that they must make a distinction between ethnic “structure” (the set of potential ethnic identities that characterizes a population) and ethnic “practice” (the set of identities actually activated by that population), must accommodate the possibility of the multiplicity of identities in both structure and practice, and must be sensitive to context and time in collecting these data. But our cross-national datasets on ethnic groups, and the measures constructed on the basis of these datasets, are resolutely primordialist – they do not distinguish between structure and practice, do not accommodate the possibility of multiplicity, and are not sensitive to time and context. The lack of theoretically justified data hampers the production of high quality research on the origin and effect of ethnic diversity, and related concepts. Although there is now a steadily increasing number of studies on the role of ethnicity in determining economic growth, the consolidation of democracy, the distribution of public goods, the study of large-scale violence and the shape of party systems, the poor quality of the data on which these studies are based means that we cannot be confident of their results.

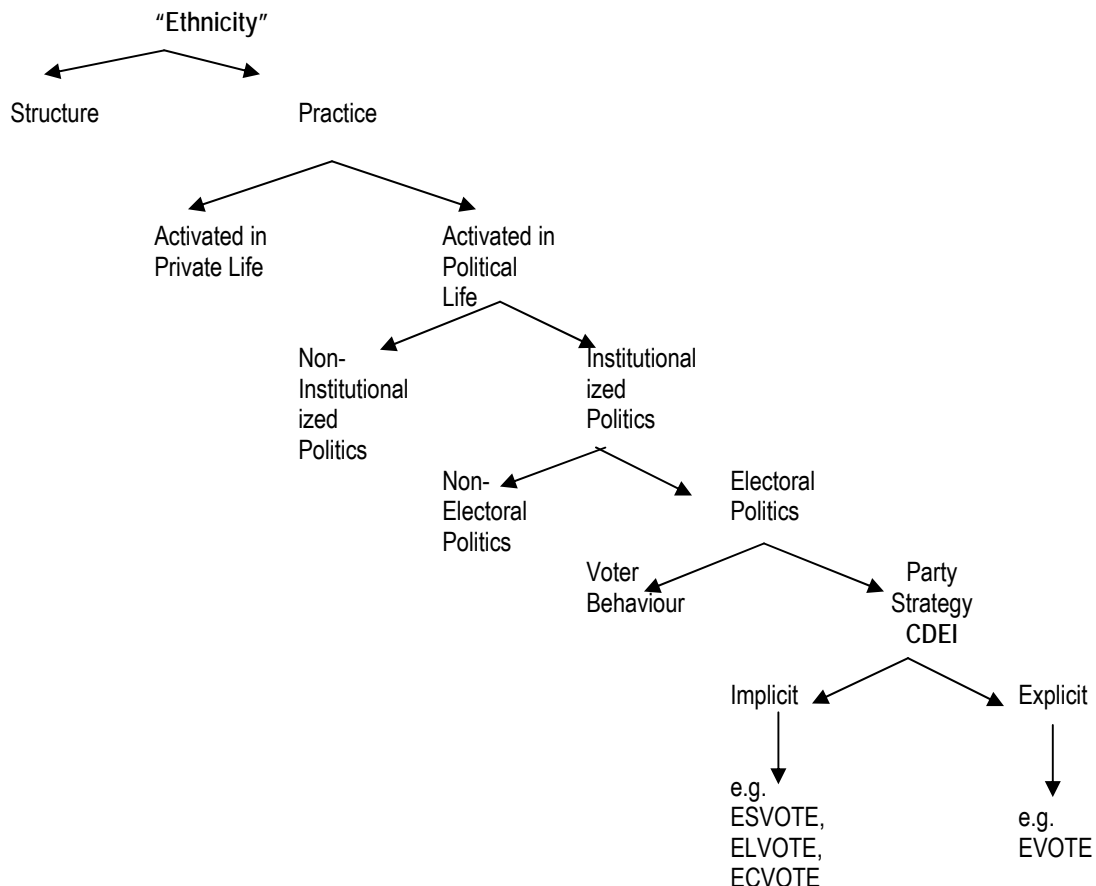
This chapter describes a new, constructivist dataset on several concepts related to ethnic identity and institutions – referred to below as CDEI (Constructivist Dataset on Ethnicity and Institutions) -- which I have been building in collaboration with MIT graduate students. CDEI, which currently covers 100 countries for the year 1996, generates a range of variables related to ethnicity and institutions constructed explicitly on constructivist insights (Chandra, Gisselquist, Metz, Wendt and Ziegfeld 2005). But I focus here on one key variable – EVOTE, or the percentage of the vote captured by ethnic parties in each country for the year 1996. EVOTE is based on a content analysis of party rhetoric – that is, what parties actually say to voters rather than what they write in their manifestos – in the election campaign closest to but before 1996. My own interest in collecting data on EVOTE is to account for variation in the performance of ethnic parties across political systems and to test for the relationship between the emergence of ethnic parties and democratic consolidation. But EVOTE, and the archives that support it, has the potential to illuminate a very diverse set of research agendas related to the origins or effects of politically activated ethnic identities, especially when expanded over time.

Section 1 describes the variables contained in CDEI and their relation to the umbrella concept of “ethnicity.” Section 2 describes EVOTE specifically and the procedure for coding EVOTE. Section 3 contrasts CDEI and EVOTE with the primordialist bases of previous cross-national datasets on ethnic groups, and the Index of Ethnolinguistic Fractionalization (ELF), the main measure constructed using these data. Section 4 describes how CDEI can be used to answer questions about the rise of ethnic parties and democratic stability. Section 5 identifies some of the broader questions that CDEI and EVOTE can be used to answer in social science research. Section 6 raises some frequently asked questions about CDEI and EVOTE. Section 7 concludes by locating CDEI and EVOTE within a broader family of efforts by comparative political scientists to collect data on different aspects of “ethnicity” from a constructivist perspective.

1. CDEI (Constructivist Dataset on Ethnicity and Institutions)

CDEI starts with an obvious point, worth stating only because it has so often been ignored: “Ethnicity” is a big concept, much like “politics.” Big phenomena are best captured, not by big concepts and measures, but by the proliferation and conjunction of several small ones. For instance, if we wanted to study how “politics” matters in explaining some outcome of interest, we would not construct datasets on and measures of “politics” in general. Rather, we would (and do) identify narrow concepts that name some specific aspect of political structure --- for instance, democratic v/s dictatorial regimes, presidential v/s parliamentary regimes, the effective number of parties in a political system – or political practice – for instance, the content of campaign rhetoric, the allocation of budgets, the degree of politically motivated violence - and collect data and design measures that operationalized these concepts. By the same logic, concepts, measures and data designed to explore the role of “ethnicity” must be narrow, tailored to specific questions and contexts.

Accordingly, CDEI collects data on over forty variables related to ethnicity, each of which captures only one aspect of the large number of ways in which ethnic identity might manifest itself in structure or practice and all of which, taken together, do not exhaust the ways in which we can measure the role of ethnicity. In particular, it focuses on collecting data and designing measures for the explicit and implicit ways in which ethnic identity is activated by political parties in election campaigns. The relation of the variables generated by CDEI to the broader concept of ethnicity is summarized in the chart below:



At the broadest level, we can imagine the term “ethnicity” to encompass two families of concepts – the “structure” of ethnic identities, and the “practice” of ethnic identification. Ethnic “structure” refers to that set of identities that are considered commonsensically real by a population, whether or not individuals actually identify with them. Ethnic “practice” refers to the act of actually using one or more identities embedded in this structure to guide behaviour. Ethnic “practice” in other words refers to the set of “activated” identities that individuals actually employ in any given context. The set of “activated” ethnic identities for any given country is typically a subset of the identities contained in the ethnic “structure.” For instance, the identities Scandinavian-American and “African-American” are both embedded in the ethnic structure of the U.S. to the extent that both are among the set of commonsensical identities that people would acknowledge as real in the US. But the identity “African-American” belongs in the subset of activated identities, while the identity “Scandinavian-American, arguably, does not.

The ethnic identities activated in practice differ according to whether they are activated in private life, and the set of ethnic identities that are activated in political life. By private life, I mean that aspect of life which concerns individuals alone, or their immediate family and friends. By political life, I mean that aspect of life which concerns collective action by individuals who are not bound by immediate personal ties. Caste identity in Sri Lanka (e.g. Goyigama) is an example of an identity that informs private actions such as the choice of a marriage partner. But religious and linguistic identities (e.g. Buddhist or Sinhala) rather than caste are the principal identities invoked in collective action (Rajasingham 1999, Tambiah 1986, Chandra 2005a).

Among the set of identities that are politically activated, we can distinguish again between identities that are activated in institutionalized politics, in parliament, party politics, the legal system and so on, and identities that are activated in non-institutionalized contexts, such as civil war, riots and social movements. In many countries, the set of identities activated in both contexts may be identical. But in others, especially in states which outlaw certain types of political participation, they can diverge. In Indonesia, for instance, institutionalized participation by political parties activates religious identities (Muslim and Christian), while regional identities are more likely to be found the arena of non-institutionalized politics (CDEI data).

Among identities that are activated in institutionalized politics, we can distinguish further between identities mobilized in electoral contexts, party politics and voting behaviour, and in non-electoral contexts, in the corridors of parliament, the military, the judiciary and the bureaucracy. In Uganda, for instance, the identity of “Nubian” was an identity mobilized principally in the military and the corridors of the bureaucracy of Idi Amin’s regime, while the identities Baganda or Catholic have frequently been mobilized in the course of electoral politics (Kasfir 1976, Kasfir 1979).

Among the identities mobilized in electoral politics, we can distinguish between identities that drive voter behaviour and those that drive party strategy. In principle, we should expect there to be some connection between the identities that parties activate and those that condition voter behaviour. But the two concepts are analytically distinct, and may sometimes diverge. In South Africa, for instance, the African National Congress mobilizes voters based on multi-ethnic appeals targeted to all South Africans, and, in some contexts, on appeals based on the racial category “Black.” But voters often vote for it, not as “South Africans” or “Blacks” but on the basis of their particular tribal identities (Xhosa, or Zulu and so on).

Finally, among the identities mobilized by political parties, we can distinguish between identities mobilized implicitly and identities mobilized explicitly. The Willie Horton advertisements used by the Republican Party in its 1988 presidential campaign in the US are an example of an implicit appeal to race (Mendelberg 2001). By contrast, the election campaign run by Slobodan Milosevic in 1992, in the immediate aftermath of Communist rule, made an explicit appeal to Serbian identity (CDEI data).

The set of identities mobilized by political parties currently forms the principal domain of CDEI. The forty-eight variables currently included in CDEI can be divided into three families:

(1) Seventeen variables on the explicit mobilization of ethnic identities by political parties: These variables are all derived from a classification of all political parties in the dataset into “ethnic,” “multi-ethnic” and “non-ethnic” parties. The most important of these is EVOTE – the aggregate percentage of the vote captured by ethnic parties in each country. Others in this family include variables coding the aggregate vote captured by non-ethnic and multi-ethnic parties, names and sizes of the ethnic groups explicitly mobilized by political parties, the types of identities explicitly mobilized by political parties, the number of identity types mobilized in each country, the proportion of an explicitly mobilized ethnic group that votes for its “own” party and so on.

(2) Seventeen variables on the implicit mobilization of ethnic identities by political parties: These variables are informed by the insight that political parties may often activate ethnic identities without explicitly naming any ethnic category. Such implicit mobilization may be based on “coded” appeals. In the 2002 race for the position of Mayor in Newark, for instance, the incumbent’s campaign slogan was “The Real Deal.” In a context in which race mattered, and both candidates were black, the slogan was an implicit attempt by the incumbent to cast doubt on the authenticity of his competitor as a representative of “black” interests. Implicit mobilization may also be based on the identity of the candidates and leaders of a political party. In India, for instance, it is routine to distribute party tickets according to complicated formulas of ethnic balancing – but the official manifestos of parties engaged in such detailed ethnic balancing often do not say a word about ethnicity explicitly. Finally, implicit mobilization can also be based on the arena of contestation: parties can convey which ethnic groups they are for simply by confining their message to voters from this ethnic group rather than explicitly championing the cause of the group. The nature of implicit appeals makes it difficult to code them based on party rhetoric – they are either designed to be open to multiple interpretations, or are invisible in speech, relying on the context to reveal them. However, we try to get at them through including three other variables in CDEI. ESVOTE measures the aggregate percentage of vote obtained by political parties which have an ethnically identified support base. The logic behind ESVOTE is that if a party is making an implicit ethnic appeal, it should show up in the nature of its support base even if we cannot code it in its message. ELVOTE measures the aggregate percentage of vote obtained by political parties which have an ethnically identified leadership. ELVOTE tries to capture the surreptitious signals sent by political parties who court ethnic groups by giving their representatives positions of power. And ECVOTE - - the aggregate percentage of vote obtained by political parties which have an ethnically identified arena of contestation whether or not they make an explicit ethnic appeal – tries to capture the implicit signal sent by parties who court ethnic groups by choosing their audience rather than choosing their words. CDEI also includes variables measuring the name, size, type and number of types of ethnic identity category that are implicitly mobilized in each country.

(3) Fourteen “general” variables on each country or party: These include the year of the election, the year of founding of the political party, whether or not there are laws preventing parties from making explicit appeals, whether or not the election is boycotted by any significant party, whether or not the election in question is a “founding” election, the percentage of votes obtained by each party, the percentage of seats obtained by the party and so on. These variables are useful in their own right and in the interpretation of the variables on explicit and implicit mobilization.

We currently expect to expand CDEI in three directions. First, we expect to collect data on variables located on other branches in the conceptual tree drawn above for the cross-national dataset. In particular, we have begun to collect data on ethnic structure (STRUCTURE), by summarizing all mentions of ethnic identities that are mentioned in our source materials as commonsensically describing a country’s population, whether or not these identities are actually mobilized by political parties. But the coding of this variable is in its very early stages. Second, we expect to expand the dataset over time to include the years 1956, 1966, 1976, 1986 and 2006, so that we will obtain a time-series coding for six decades. Finally, we expect to expand the country coverage to include all countries in each year for which we are able to obtain vote share data.

2. EVOTE (The aggregate percentage of votes won by ethnic parties)

EVOTE is constructed as follows: first, we classify each political party in each country for which we can obtain data as “ethnic,” “multi-ethnic” or “non-ethnic” based on its campaign in the legislative election closest to but before 1996. Then, we add up the total percentage of votes obtained by all ethnic parties in a given country. Thus, EVOTE for Country A is constructed as follows: EVOTE96 (Country A) = Vote for Ethnic Party 1 (Country A) + Vote for Ethnic Party 2 (Country A) + Vote for Ethnic Party 3 (Country A) + Vote for Ethnic Party N (Country A). In principle, the value of EVOTE could range between 0 (for countries with no ethnic parties) to 100% (for countries in which all votes are captured by ethnic parties). In reality, EVOTE for the year 1996 ranges from 0% (e.g. in Greece) to 85.63% (in Yugoslavia), with a mean value of 12.95%. The table below summarizes the regional distribution of parties and countries on which we have data on EVOTE. A full description of the patterns revealed in EVOTE is contained in Chandra, Gisselquist, Metz, Wendt and Ziegfeld 2005.

Regional Distribution of Parties and Countries in CDEI for which we have data on EVOTE

	No. of Countries	% of countries	No. of Parties	% of parties
Latin America	17	17	143	10.62
Europe	31	31	554	41.16
Asia	16	16	183	13.6
Post-Soviet	10	10	202	15.01
Africa	22	22	215	15.97
Middle East	2	2	33	2.45
North America	2	2	16	1.19
Total	100	100	1346	100

The classification of parties which is the foundation of EVOTE is based on the definitions proposed in Chandra (2004a, 3), according to which an ethnic political party is “a party that represents itself

to voters as the champion of the interests of one ethnic category or set of categories to the exclusion of another or others, and makes such a representation central to its strategy of mobilizing voters.” The key aspects of this definition are exclusion – an ethnic party must make an appeal on behalf of some ethnic group(s) that excludes others – explicitness – the appeal must be open – and centrality – the appeal must be central to its mobilizing strategy. A multi-ethnic party also makes an open appeal related to ethnicity central to its mobilizing strategy but assumes a position of neutrality/equidistance towards all relevant groups. In other words, it differs from an ethnic party only in its inclusiveness. A non-ethnic party is one that does not make an ethnic appeal central to its mobilizing strategy. Note that these definitions classify parties based on their message. And because messages can change across elections, they are time-sensitive: a party classified as an ethnic party in one election need not be classified the same way in subsequent elections.

An ethnic group, in turn, is defined as an impersonal social category in which membership is determined by inherited attributes and which comprises a subgroup of a country’s population (Chandra 2004b). This typically includes most (but not all) groups defined by caste, tribe, nationality, race, religion, region, and language. Note that there is nothing in this definition which requires individuals in the same ethnic group to share a common sense of group identity or even a consciousness of the existence of other group members. I use the term ethnic “category” and ethnic “group” interchangeably to emphasize this point. Considered in relation to the definition of identity laid out by Abdelal et al, this is a minimalist definition of ethnic identity, which distinguishes between ethnic identity categories based purely on their “constitutive norms” (i.e. rules of membership), and not on their shared goals, frames of comparison, and cognitive content, or the degree of contestation over any of these things.

The coding of the parties is based on a content analysis of the election campaign of the party in question using four sources: the Europa World Yearbook, the Political Handbook of the World, news sources from FBIS (Foreign Broadcast Information Service), and Lexis-Nexis searches. For each party, we obtain a sample of campaign materials (speeches at election rallies, policy pronouncements, and so on) as reported in FBIS and Lexis Nexis for a period up to three months before the election date. These include reports from the international media and translations of local news reports from newspapers, radio and TV. These samples have three advantages: (1) They are primary sources that report what parties are actually saying to voters rather than what they print in their manifestos. (2) Many of these sources are translations of what parties say to voters in local languages (3) They are time-sensitive sources that report party statements for the year of election. Where the samples are too small to permit reliable codings, we turn to local newspapers and secondary sources as a last resort. These sources give us a sample of articles for the election platform of each party individually. We archive the materials for each party for each country after completing the coding. This archive, composed of a uniform set of source materials for each observation (country or party), makes it possible both to double check old variables as we proceed, and to construct new variables as they become important.

The coding is based on a protocol that establishes rules for the identification of an appeal as ethnic, explicit and central. In contrast to the analyses described in Abdelal et al and in other chapters in this volume, the content analysis is qualitative rather than quantitative. Thus, rather than establishing centrality simply by counting the number of times an issue is mentioned, the protocol identifies rules of interpretation for centrality – an ethnic appeal can be central based on the frequency of an issue, but also on the way in which an ethnic appeal is used. For instance, a

party which associates an ethnic category with ownership of the state would be coded as an ethnic party even if statements to this effect were not frequent, based on the reasoning that once such a statement is made, it colours the interpretation of other statements. If, based on the content analysis, we find that a political party makes an open and exclusive appeal to some ethnic category or set of categories, and that such an appeal is central to its campaign, we code it as an ethnic party. If we find that a political party makes an open and inclusive appeal to all ethnic categories that define a population and makes such an appeal central to its election campaign, we code it as a multi-ethnic party. And if we find that a political party does not make an open or a central appeal to an ethnic category, whether exclusive or inclusive, we code it as a non-ethnic party. We document each coding for each party on each country, compare it with codings in other datasets where available, check for consistency across coders, and assign a reliability score to the coding (1 = high certainty, 2 = moderate certainty, 3 = low certainty) based on the quality of information in the sample. Since countries can sometimes have over a hundred parties, each with a separate sample, this level of documentation adds considerably to the time that constructing this dataset requires. But it is important if other researchers to replicate our efforts, and estimate the bias and error in the data.

Consider the case of India, as an example of our coding procedures. Hundreds of parties competed in India in the 1991 parliamentary elections (the elections closest to but before 1996) but most of them obtained a miniscule percentage of the vote. We obtained disaggregated data on all parties that obtained at least .01% of the vote, thus including 66 parties in our dataset. We then coded each of these 66 parties based on a content analysis of its party platform. Of the 66 parties, we coded 13 parties, accounting for 51.81% of the vote as non-ethnic, 18 parties, accounting for 38.95% of the vote as ethnic, and were not able to find sufficient articles on election platform to code the remaining 35 parties, accounting for 10.24% of the vote (these were very small parties, with a mean vote of .14%).

Once we have coded EVOTE, we can also obtain data on other variables from the same source materials in relatively short order. We can immediately record the names and population of all explicitly mobilized categories. In India, for instance, ethnic parties, taken together, explicitly mobilized the following ethnic categories: Hindus (82%), Muslims (12.12%); Sikhs (1.94%); OBCs (52%), Scheduled Castes (16.48%), Jharkhandis (3.18%), Assamese (2.64%); Tamils (6.6%). We can also use the archive of materials on each country (which typically includes background information on the country from Europe and Political Handbook of the World, third party assessments of the main parties, as well as specific information on the rules governing the elections from the local news sources) to code the variables on the implicit mobilization of ethnic identities, on laws governing ethnic appeals, on party registration procedures and so on.

3. What makes CDEI and EVOTE constructivist?

In order to highlight the constructivist features of CDEI and EVOTE and CDEI, we must first relate them to previous datasets that also collect data on and construct measures for some aspect of "ethnicity."

There are currently three cross-national dataset on ethnic groups (1) The Atlas Narodov Mira, published by Soviet Ethnographers in 1964 (Miklukho-Maklai Ethnological Institute 1964) (2) a dataset on ethnic groups in 190 countries published by Alesina et al in 2003 (Alesina et al 2003) and (3) a comparable count of ethnic groups in 160 countries published by James Fearon in the same year (Fearon 2003). The principal measure constructed on the basis of these data is the

Index of Ethnolinguistic Fractionalization (ELF), which measures the degree of ethnic diversity in a population. The ELF Index is calculated according to the formula $1 - \sum s_i^2$, where s_i is the proportion of the i th activated ethnic category, $i=\{1, 2, \dots, n\}$, where the ethnic groups are mutually exclusive (i.e. if you are in ethnic group 1, you are not in ethnic groups 2-n) and exhaustive (every member of the population is in some ethnic group). It measures the probability that two randomly chosen individuals from a country's population belong to different groups. Thus, a society with two groups, a majority of 80% and a minority of 20%, would have an ELF score of $1 - (.64 + .04) = .32$. A society with several small groups of 25% each would have a higher ELF score of $1 - (.0625 + .0625 + .0625 + .0625) = .75$.

Below, I raise at least four questions suggested by a constructivist perspective about what these data and measures refer to: (1) Do they measure ethnic "structure" or ethnic "practice"? (2) If "structure," then why do the datasets ignore the problem of multidimensionality, and what does the ELF index mean, given such multidimensionality? (3) If "practice," they why do these datasets ignore the problems of overlap and incompleteness? And how can the ELF index, which requires the assumptions of mutual exclusiveness and exhaustiveness, be meaningful? (4) Whether structure or practice, what time period and context do these data refer to? None of these questions would be worth raising in a primordialist world, in which there is no distinction between structure and practice, individuals have unidimensional ethnic identities which are mutually exclusive and exhaustive, and retain these identities in all times and contexts. But they are critical questions from a constructivist point of view. Although the authors of the more recent datasets acknowledge that some of these questions can create problems for a data-collection exercise in theory, they do not provide answers about how they actually solve these problems in their coding criteria. Without answers to these questions, these data cannot be satisfactorily interpreted, replicated and extended. This section elaborates on these problems and shows how CDEI and EVOTE address them.

(i) Do the data and the ELF measure refer to ethnic "structure" or ethnic "practice"? Do the three previous datasets count the ethnic groups which are commonsensically real or the ethnic groups which are activated? And does the ELF index measure the degree of ethnic heterogeneity given the set of commonsensically real identities, or the degree of ethnic heterogeneity given the set of activated ones?

The approach taken by Atlas Narodov Mira is not spelled out. Alesina et al refer to ethnic groups as the product of "persistent identification" (Alesina et al, 161), thus conflating structure and practice. Fearon, by contrast, makes an explicit distinction between the two, and aims to code for ethnic structure by trying to capture commonsensically real identities -- "how people in the country mentally divide the social terrain in ethnic terms" (Fearon 2003, 203) -- and not ethnic practice, at least in the political realm. But the criteria for operationalizing this distinction in the coding procedures are not laid out, and if we look at the data, we find that all three datasets veer inconsistently between collecting data on ethnic structure and ethnic practice.

Consider the example of Albania. Each of the three datasets code ethnic groups in Albania as "Albanian", "Greek" and "Macedonian." But why not include ethnic groups based on religion (Catholic, Orthodox and Muslim) or dialect (Gheg-speakers, concentrated in the north and Tosk speakers, concentrated in the south) in their count? These other identities appear also to be commonsensically real. The principal distinction between the excluded and included groups seems to be that the groups Albanian, Greek and Macedonian have more political resonance in the

present than the others. Inadvertently, then, the datasets appear to be coding practice rather than structure in the case of Albania.

But in Italy, the pattern is reversed. The Atlas Narodov Mira codes Italy as 98% Italian, with a range of smaller groups including Austrians, French, Slovenians, Albanians making up the remaining 2%. Fearon codes Italy as an almost entirely homogeneous country, with a 98% majority of “Italians.” Alesina et al code Italy in a comparable way, as 94% Italian, 2.7% Sardinian, 1.3% “Rhaetian”, and 1.9%. But if we look at politically mobilized identities in Italy, at least in electoral politics, we see several other identities mobilized by political parties, including regional identities (“North and “South”), sub-regional identities (Milan and Lombardy) and racial identities (“native” versus immigrant Italians). Germany, similarly, is consistently coded by all three datasets as having an overwhelming German majority (ranging from 91% in Alesina to 98.8% in Narodov Mira), with several tiny minorities. But why only count groups based on nationality or race in Germany and not groups based on region -- East Germans, for instance, or Bavarians -- which CDEI reveals are identities activated by political parties? By excluding these identities, the datasets all appear to be attempting to capture some undefined notion of structure rather than practice.

CDEI corrects for this problem by coding variables based on ethnic practice, consistently separating this from ethnic “structure,” which we are now beginning to code.

(2) If “structure,” then how to address the problem of multi-dimensionality and level of aggregation in datasets and measures? The structure of ethnic identities in most countries is multi-dimensional, although the number and type of dimensions can vary. The set of identities that are commonsensically real in the US, for instance, includes identities based at least on the dimensions of race, nationality, region, religion and tribe. In India, it includes identities based on the dimensions of caste, language, tribe, region and religion. In Zambia, it includes identities based on the dimensions of tribe and language. In South Africa, it includes identities based on the dimensions of race and tribe. In Malaysia, it includes identities based on the dimensions of race, language, region, religion and tribe.

Further, categories on each dimension are arrayed at multiple levels –which level should we count on (Laitin and Posner 2001)? When faced with the dimension of tribe in India, should we count categories at the highest level of aggregation (e.g. Scheduled Tribe) or at the lowest level (Santhal, Munda, Bhil and so on) or somewhere in between? When faced with the dimension of religion in the U.S., should we count based on meta categories (Christian, Muslim, Jewish) or micro-categories (Methodist, Baptist, Presbyterian, Shia, Sunni, Ismaili, Hasidic, Orthodox and so on)?

Alesina et al and Fearon acknowledge these problems and discuss them at some length. But they do not furnish the decision rule that they employ to solve them, and we cannot infer the rule from actually looking at the data. To illustrate, consider the coding for India across the three datasets and CDEI:

	Groups Included in Count	ELF	Size of Largest Group
Atlas Narodov Mira	Hindi speaking peoples of North India (.25), Bihars, Maraths, Bengals, Gujarats, Rajastans, Oriya, Panjabs, Assams, Kumaoni, Kashmirs, Bhils, Gujars, Sindhi,	.89	.25

	Gurkhi, Pars, Jhats, Shina, Kho, Kohians, English, Jews, Pushtuns, Portuguese, Telugs, Tamils, Kannara, Malayali, Gondi, Tulu, Oraoni, Kandhi, Kodagu, Badaga, Irula, Urali, Maler, Mannans, Malavedans, Kurumba, Kadari, Paniabs, Toda, Kota, Chenchu, Santals, Munda, Ho, Savara, Korku, Bhumidji, Kharia, Gadaba, Djuangs, Minipuri, Naga, Garo, Balti, Lushei, Kachars, Tipera, Mikiri, Kirats, Kuki, Tamangs, Ladahs, Bhoti, Thado, Miri, Abor, Mishmi, Dafla, Limbu, Lepcha, Kanauri, Lahauli, Gurungs, Nevars, Magars, Sherps, Sunvars, Burmese, Kachins, Chinese, Khamti, Khasi, Nikobars, Burishs, Andamanese. Put in data on numbers.		
Alesina et al	Indo-Aryan (.72); Dravidian (.25), Other (3)	.42	.72
Fearon	Speakers of Hindi (.39), Bengali (.08), Telugu (.08), Marathi (.07), Tamil (.06), Gujarati, Malayalam (.05), Kannada (.04), Oriya (.03), Punjabi (.03), Sikhs (.02), Assamese (.01)	.81	.39
CDEI	Hindus (.82), Muslims (.12); Sikhs (.019); OBCs (.52), Scheduled Castes (.16), Jharkhandis (.031), Assamese (.26); Tamils (.066).	NA (EVOTE = .3895)	.82

The Atlas Narodov Mira counts a cluster of groups based on several dimensions -- language (e.g. Hindi speakers and Tamils), nationality (Portuguese, English), Tribe (Santals, Munda), Religion (Jews), Caste (Gujars, Jats). But the choice of groups included from each dimension is arbitrary, and I cannot discern the logic to the level of aggregation chosen.

Alesina et al include two groups from the highest level of aggregation on the dimension of language (Indo-Aryan and Dravidian). But, given that they are concerned with a count of "ethnic" groups, it is not clear why they choose the dimension of language rather than the dimensions of caste, religion, region, and tribe. And it is not clear why they choose the highest level of aggregation in this case, in express contradiction of their intention to collect data at as disaggregated a level as possible (Alesina et al 160). (In the same dataset, Alesina et al collect data separately on the dimensions of religion and language, but never make clear how these dimensions are separate from rather than contained within the concept of "ethnic" identities).

Fearon reports the ethnic structure as made up of several groups defined on the dimension of language -- Speakers of Hindi (.39), Bengali (.08), Telugu (.08), Marathi (.07), Tamil (.06), Gujarati, Malayalam (.05), Kannada (.04), Oriya (.03), Punjabi (.03), Assamese (.01), and one group on the dimension of religion -- Sikhs (.02). But why not include other groups on the dimensions of religion (Hindus and Muslims, for instance) or tribe (Scheduled Tribes and others) or caste (at the highest level of aggregation, this would include Upper Castes., Backward Castes and Scheduled Castes)? And on the dimension of language, why not include groups at a higher level of aggregation (e.g.

Indo-Aryan languages (including Hindi, Bengali, Gujarati, Punjabi) and Dravidian languages (including Telugu, Tamil, Kannada and Malayalam)?

In an improvement over the other two datasets, Fearon does provide a conceptual justification for his count. He attempts to include in his count groups that fulfil as many of the following prototypical criteria as possible: (1) Membership is reckoned primarily by descent (2) Members are conscious of group membership (3) Members share distinguishing cultural features (4) These cultural features are valued by a majority of members (5) The group has or remembers a homeland (6) The group has a shared history as a group that 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.” (Fearon 2003, 201).

But I cannot tell how these conceptual criteria are operationalized in the coding process. Determining whether the members of a group have a factual rather than a fictitious history, or whether they value distinguishing cultural features or whether they are conscious of group membership is no easy task, even for those who specialize in a particular country or a particular group. How might a coder make these decisions, and how might others replicate them? Nor is it clear how many prototypical criteria a group must satisfy in order to be included, or how a coder should decide between multiple candidate groups on multiple dimensions that fit the prototypical criteria. Why, for instance, was the category “Jat” (included in Atlas Narodov Mira but not in Fearon), which appears to meet criteria 1-6, but not 7 not chosen over the category “Punjabi,” which appears to meet criteria 1, 3, 4 and 5, but not 2, and 6 and, arguably 7? Some of the groups included in Fearon’s count do not seem to meet several of the conditions. “Hindi-speakers,” for instance, are not a “group” in which members are conscious of group membership, share distinguishing cultural features that are valued by a majority of members, and has or remembers a homeland. Further, several groups excluded from this count also appear to meet several of the prototypical criteria, such as Indo-Aryans and Dravidians, Hindus and Muslims, and Scheduled Tribes.

These differences are consequential: the ELF index jumps from .42 to .89 depending upon the data chosen. And the size of the largest ethnic group, also a common measure constructed on the basis of these datasets, ranges from .25 to .72.

The multidimensionality of ethnic identities also poses a problem for the ELF index as a measure of ethnic diversity. The ELF index can only give us a measure of ethnic diversity based on a set of mutually exclusive groups chosen from within or across dimensions. But it cannot be used to capture the ethnic diversity of a country if we take the multiple dimensions together, along with the relationship between them, into account. To illustrate the problem this poses, compare India and Zambia. I noted above that at least five dimensions of ethnic identity are commonsensically real in India (Chandra 2004a), with approximately two to seven categories arrayed on each, if we confine ourselves to the highest level of aggregation. In Zambia, only two dimensions of identity are salient: tribe and language (Posner 2005). The dimension of language currently has four groups arrayed on it at the highest level of aggregation: {Bemba-speakers, Nyanja-speakers, Tonga-speakers, Lozi-speakers}. The dimension of tribe has roughly seventy: {Chewa, Tembuka, Bemba, and so on...} Which country is more diverse, given variation in the number of dimensions and categories in each? We cannot use the ELF index to tell us.

The CDEI variable on ethnic “structure,” aims to collect separate data on all dimensions of ethnic identity that are commonsensically real in a given country, at all levels of aggregation. We also propose a new measure, a “Multidimensional ELF Index (MELF) that allows comparison in the degree of ethnic diversity across countries taking these multiple dimensions into account, and allowing for differences in the number of dimensions and the number of categories on each dimension, across countries (Chandra 2005b). But in the realm of ethnic “practice,” currently our main concern, the categories and dimensions counted in CDEI are those that are named by the parties in question. Thus, the CDEI count of explicitly activated ethnic categories in India is based on all the categories explicitly mobilized by political parties in the 1991 parliamentary election campaign (the national legislative election closest to but before 1996). (We could in principle also construct a list of all categories activated in an election campaign, explicitly or implicitly). These data cannot be used to calculate an ELF index, since they are neither mutually exclusive nor exhaustive. But they can be used to calculate EVOTE (.3895), and they can be used to measure the size of the largest activated ethnic category (.82).

(iii). If practice – the set of identities that are politically (or otherwise mobilized) – then how do we account for the problem of overlap and incompleteness in our datasets and measures? What if we take the datasets to report data not on the groups embedded in an ethnic structure but on ethnic groups as they are activated in practice, in politics or otherwise? In this case, we run into the problems of overlap and incompleteness. Take, first, examples of overlap. We have no reason to expect that the ethnic categories that individuals activate in practice should be mutually exclusive. Indeed, in many of the cases I have looked at in CDEI, the mobilized categories are overlapping. In CDEI’s count of politically activated ethnic categories in India, summarized in the table above, the categories Hindu, Muslim and Sikh are mutually exclusive in relation to each other, but overlap with the categories OBCs, Scheduled Castes, Jharkhandis, Assamese and Tamils. Sometimes, there are cases of total overlap, so that one category is entirely nested within another. In Belgium, for instance, among the categories mobilized by political parties are native Belgians (v/s immigrants), who constitute 91% of the population, and French speakers, who constitute 42% of the population, who are largely contained within the native-Belgian category.

Take, next, examples of incompleteness. There is no rule that individuals in a population should all activate their ethnic identities exclusively. Indeed, it is only a few very polarized countries at particular points in time, such as Yugoslavia in 1992, where almost the entire population lines up behind an ethnic identity – but even in such countries, the ethnic identification may not be complete. CDEI shows that 86% of the population in Yugoslavia voted for ethnic parties in 1992, leaving a minority of voters who voted for other types of parties. In other countries, we typically see several types of identities activated in practice. In the recent U.S. presidential elections, for instance, some voters activated their class identities (e.g. middle-class), others their party identities (Republican or Democrat), others identities based on age (e.g. pensioners) and still others their racial identities (e.g. Black).

Given the problems of overlap and incompleteness, it is not clear, how the three datasets produce a count of groups in all countries that is conveniently mutually exclusive and exhaustive. Further it is not clear what the ELF index, which requires mutual exclusiveness and exhaustiveness, means.

CDEI addresses these problems by constructing a variable – EVOTE –which does not impose any assumptions about mutual exclusiveness and exhaustiveness. For instance, in the Indian case, the

value of EVOTE is .3895. This proportion is unaffected by whether the ethnic parties in question mobilize mutually exclusive or overlapping categories. In India, it so happens that the parties activate overlapping categories. But the value of EVOTE would be the same even if the parties in question activated mutually exclusive categories: we might get the same value of EVOTE if a party mobilizing Hindus obtained 30% of the vote and a party mobilizing Muslims obtained 8.95% of the vote. Similarly, EVOTE also does not impose the requirement that the categories activated by political parties be complete, since it is the votes won by the parties that mobilize each category that are added, not the proportion of the population made up by the categories themselves. Rather, we it allows us to observe such completeness in the data. In countries in which all individuals activate ethnic identities in their voting behaviour, the value of EVOTE would be 100%. In countries in which only some individuals activate ethnic identities in their voting behaviour, the value of EVOTE would be less than 100%.

(iv) Whether structure or practice, what time period and context do the data refer to?

Finally, a key constructivist insight is that both the structure of ethnic identities, and the set of politically mobilized identities can change over time and by context. Ideally, this insight suggests that we should collect time-series data on ethnic identities, just as we collect time-series data on regime type, per capita income and other variables which are dynamic in nature. And cross-sectional datasets should at a minimum explicitly identify the time and context in which it is collected.

But the three previous datasets do not date their sources and locate them in a particular context. This creates a problem in using the data: how can we use it for an analysis of democratic stability in 2000 if we do not know whether the ethnic groups in question were counted in 2000, 1950 or 1900, and we do not know whether this count reflects the groups relevant in electoral politics, or anti-colonial mobilization, or violence? It also makes it difficult to expand the dataset – if we do not know which time period and context these data refer to, how can we know which time periods and contexts to add?

CDEI and EVOTE have yet to be expanded over time. But the data are restricted to a single context – electoral politics at the national level – and to a specific time period – the legislative election closest to but before 1996. We do not to incorporate data on political parties which does not apply to the election campaign for the election under study. A party might well have been an ethnic party in some previous election, but if we do not find evidence of an explicit ethnic appeal in that particular election campaign, we do not code it as ethnic regardless of its political history. This means that the data can be located at a particular point in time and that we can expand the dataset to include other points in time in a meaningful way.

These criticisms of previous datasets highlight, not that these previous datasets are “wrong,” but that we do not know whether they are wrong or right. When the counts in the three datasets disagree, as in the case of India, we do not know why and have no criteria by which to determine which dataset to trust. And when they agree, as in the case of Italy and Germany, we do not know if they are right and what they are right about, given that all three datasets inexplicably exclude the same identities. The ELF indices based on the three datasets are moderately well correlated (Fearon 2003, 214), but we cannot tell whether the correlation is an indication that the data are correct or an indication of systematic bias. At least among the Alesina and Fearon datasets, agreement when it occurs may simply be a consequence of the fact that both rely on some of the same data sources (the CIA World Factbook and the Encyclopedia Britannica). In comparison, the

advantage of CDEI in an analysis of democracy (and in other analyses described below) is not that the variables it generates are “correct” or suitable for all purposes. CDEI has both bias and measurement error, which I discuss in the conclusion and in the supporting documentation. But the coding criteria are sufficiently transparent for other researchers to make judgments about what these biases and errors might be and how to compensate for them.

4. Using these Data to Investigate the Relationship Between Ethnic Parties and Democratic Stability

My purpose in building and expanding CDEI is to test competing propositions on the relationship between the rise of ethnic parties and democratic stability. According to a classic proposition in empirical democratic theory, ethnically divided societies invariably give rise to ethnic political parties, and ethnic political parties destabilize a democratic system. (Horowitz 1985, Rabushka and Shepsle 1972). The logic of this proposition goes as follows: (1) The rise of even a single ethnic party infects the rest of the party system (2) Ethnic political parties engage in outbidding behaviour, with each side seeking to exclude the supporters of the other from electoral competition (3) Sooner or later, democracy is subverted, either because the winning party tampers with the rules of free and fair competition or because the losing side engages in pre-emptive violence to prevent exclusion. This proposition has gained wide currency, not only among democratic theorists, but among governing elites in multi-ethnic democracies, who try hard to prevent the emergence of such parties, sometimes by outlawing them altogether.

Multi-ethnic democracies are also believed to be under threat for reasons not explicitly lined to political party behaviour: (1) Because they do not possess the minimal sense of political community necessary for democracy to function (Mill 1991); (2) Because the demands made by ethnic groups are more intractable than demands made by non-ethnic groups (Rustow 1970, Horowitz 1985); and (3) Because they are more likely to produce incipient nations than societies divided on class lines (Geertz 1973). These additional propositions all suggest that the emergence of ethnic parties is not necessary to destabilize democracy in multi-ethnic societies. But they imply that ethnic parties are sufficient to destabilize democracy, since an ethnic party system should surely activate one or more of these mechanisms by activating ethnic differences.

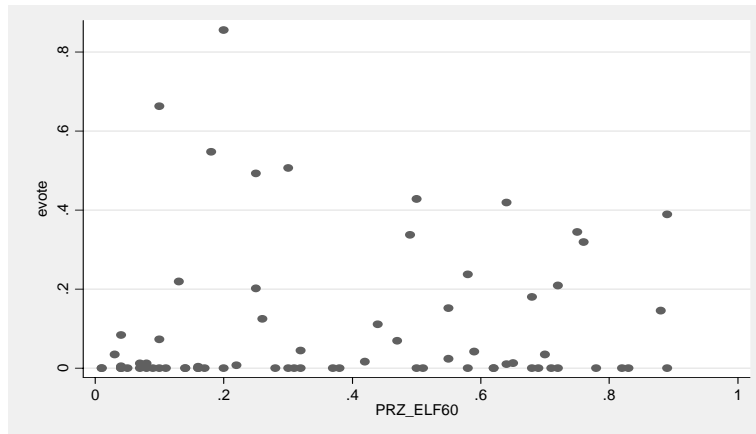
But other work, including my own, suggests a different hypothesis about the effect of ethnic parties. Based on a study of the anomalous behaviour of ethnic parties in India, I argue that political systems that encourage the proliferation of ethnic parties can safeguard democratic stability by making it more profitable for losing political entrepreneurs to play within the rules of the game and manufacture new electoral majorities than subvert the system through violence (Chandra 2005a, Chandra 2004a).¹ A new manuscript by Johanna Birnir makes the argument that institutionalized access – a concept that includes representation in political parties -- reduces the tendency of ethnic groups to engage in conflict (Birnir forthcoming). And earlier work by Arend Lijphart implies that ethnic parties that enjoy the support of their followers can play a benign role in democratic politics by negotiating and guaranteeing consociational bargains (1977). These works all suggest a

¹ Chandra 2004a also argues that successful ethnic parties are not invariably produced by ethnic divisions but result from a conjunction of demographic, electoral and organizational variables. I use CDEI data to test this hypothesis in other research (Chandra et al 2005). But here, I am concerned principally with the effect rather than the origin of ethnic parties.

positive correlation between successful ethnic parties and democratic stability, although each isolates a different mechanism explaining this correlation.

Neither hypothesis has been subjected to empirical tests for the reason that we have not so far had cross-national data so far on ethnic political parties. The studies that we do have test for the relationship between “ethnic diversity,” measured using the ELF index, constructed on the basis of cross-national data on ethnic groups. In principle, these data could have served as a partial test of the hypotheses outlined above, by establishing at least whether there is a negative relationship between ethnic diversity and democratic breakdown. But these data, and the ELF index, are unreliable for reasons that I will detail at some length in the subsequent section. Partly because of the poor quality of these data, studies of the relationship between ethnic diversity and democratic stability are inconclusive. In the principal work on the determinants of democratic consolidation, Przeworski et al find that ethnic diversity destabilizes all types of regimes, democracies or dictatorships (Przeworski et al 2000, 125). However, they note that, given the quality of the data, they are not confident of these results, and note that these results are not robust to model specifications. Studies of the onset of civil war, a subject related to democratic destabilization, are similarly inconclusive. They find that different measures of ethnic diversity are negatively related, positively related, and not significantly related to the probability of civil war onset (Collier n.d., Collier and Hoeffler 2001, Fearon and Laitin 2003, Sambanis 2001).

I expect to test the two hypotheses outlined above by replicating the test for democratic consolidation conducted by Przeworski et al, with some important modifications in both model specification and data (Przeworski et al 2000, Table 2.17, 124). Using a time-series cross-sectional dataset with annual observations on 141 countries, Przeworski et al used a dynamic probit model to estimate the effect of the level of economic development on democratic stability, controlling for ethnic diversity, culture, political history and the influence of the international environment. The variables they used to measure the effect of ethnic diversity are ELF or RELDIF (religious fractionalization, calculated on the basis of the proportion of Catholics, Protestants and Muslims in a country's population). Once the time-series version of CDEI is complete, I expect to replicate their analysis by replacing the ELF and RELDIF variable with EVOTE (and also to experiment with supplementing, or replacing EVOTE with the several CDEI variables measuring the strength of political parties that mobilize ethnic identities implicitly). At least on the basis of the cross-sectional data collected so far, there appears to be no significant relationship between EVOTE and the ELF measure used by Przeworski et al. The correlation between EVOTE and ELF is only .07 (the correlation between EVOTE and ELF based on the Fearon and Alesina et al datasets described in Section 4 is similarly weak). The relationship between the two measures is summarized in the chart below.



Apart from replacing the ELF index with EVOTE, I expect to modify the analysis in two other ways: (1) By introducing variables on institutional structure drawn from existing datasets (Golder 2005, Rodden 2004, Rodden and Wibbels 2002, Keefer 2002). These include variables on electoral rules, district magnitudes, federal systems, and presidential and parliamentary regimes, which I expect to consider separately, and in combination, in a four-point scale of “consociationalism” (Chandra 2005b) (2) By employing several different specifications of the dependent variable, using at least the following sources of data: (1) The Polity IV scale of democracy. (2) The Fearon-Laitin on the onset of civil war. (3) The Banks and MAR data on civil violence. The first is important to do because we have a theoretical literature that indicates that the effect of ethnic divisions and ethnic parties on democratic stability may be mediated to some extent by institutional designs. The second is important to do to see if the results are robust across different specifications of the dependent variable.

Once the correlation between the rise of ethnic parties and democratic stability has been established, I expect to conduct further tests for the mechanism explaining this correlation, using the time-series data on EVOTE and other variables to examine whether ethnic parties “infect” a party system (this can be investigated by seeing whether ethnic parties at any one point in time are positively correlated with the vote for ethnic parties at a later point), whether particular types of ethnic identity (e.g. religion) are more likely to be associated with democratic breakdown than others (this can be tested by considering the effect of the politicization of individual identity types on democratic stability), whether exclusion from participation in government for ethnic parties is especially destabilizing (this can be tested by coding a new variable for the number of times an ethnic party has been included in government), and so on.

5. Using the Data to Ask Other Social Science Questions

Although my own interest in constructing EVOTE and CDEI is to investigate the relationship between the rise of ethnic parties and democratic stability, these data are of value to social scientists interested in exploring the effect of the politicization of ethnic identities, at least as measured through the party system, on a wide range of outcomes. Other social scientists can use the data in three ways: (1) Take one or more of the 48 off-the-shelf variables contained in CDEI for use in their own analyses, combining them with data already collected from other sources (2) Modify and recode one or more of these 48 variables for their own use (3) Use the archive of materials that informs CDEI to generate additional variables. In each case, the dataset can be organized by country or party or group.

One important set of questions that can be answered using the off-the-shelf variables in CDEI include the following: What is the effect of the explicit politicization of ethnic divisions (measured using EVOTE) or the politicization of ethnic divisions generally (measured by combining the vote share for parties that engage in the implicit or explicit mobilization of ethnic identity) on some outcome of interest, including war, riots, economic growth, public policy, welfare spending and so on? There is already a voluminous body of work in political science and economics that examines the effect of the ELF index, on these outcomes. But, as I argued earlier, we do not know what ELF measures. Replacing ELF with EVOTE in this body of work would be a meaningful test whether one specific concept - the degree to which ethnic identities are explicitly politicized in the party system at a particular point in time – matters in explaining any of these outcomes, while leaving open the possibility that “ethnicity” might matter in ways not captured by EVOTE.

Other questions which can be immediately answered using CDEI data include: Is the politicization of particular types of ethnic divisions associated (e.g. region or religion or language or tribe) associated with particular types of outcomes? This would entail using the percentage of votes won by regional or religious or linguistic or tribal parties across countries as measures of the political salience of these particular types of divisions across countries? What determines the size of the coalition that an ethnic party is likely to mobilize? This would entail using the party as the unit of analysis, and taking the proportion of the ethnic party’s target category as the dependent variable. Are we more likely to see the ethnification of politics in new democracies? This would entail treating EVOTE (or one of its substitutes as the dependent variable) and using the age of the democracy (measured in the Przeworski dataset) or the presence of founding elections (measured by CDEI) as an independent variable. Is there a link between colonial history and the degree of ethnic politicization? This would entail regressing EVOTE (or its substitutes) on the range of variables on colonial history being collected by Wilkinson and others. In these cases, the researcher simply needs to combine CDEI variables with other variables already collected in other datasets.

Questions that can be addressed through a minor modification of existing variables include the following: How rapidly do politically activated ethnic identities change over time and what explains such change? (This would entail constructing a “volatility” index comparing changes in mobilized identities across election periods). When might political parties activate complex categories rather than simple ones? (This would entail constructing a variable measuring the degree of complexity in an ethnic coalition). Do ethnic parties mobilize minimum winning coalitions? (This would entail creating a new variable subtracting the size of the coalition that a party mobilized from the winning threshold imposed by an electoral system). Is there a link between ethnic “majoritarianism” and conflict (This would entail coding a dummy variable based on the whether the size of the ethnic group mobilized by a political party was a majority or minority). These variables could be constructed by using the recode commands in Stata or another statistical package.

Finally, the archives supporting the dataset can also serve as the basis for constructing “spin-off” variables which allow researchers to ask a broader range of questions. The archives can be used, for instance, to produce a more fine tuned classification of parties, based on the particular issues they emphasise, than the blunt typology of ethnic, multi-ethnic and non-ethnic parties that CDEI has used so far. Rachel Gisselquist, one of the co-authors of the dataset, for instance, has created a new variable on economic parties using our source materials. This is an important variable that

disaggregates the residual category of “non-ethnic” party according to the issues and groups that non-ethnic parties activate. Gisselquist uses this variable, combined with the dataset as a whole, in a dissertation that asks – Under what conditions do economic rather than ethnic cleavages become salient (Gisselquist 2005)? Another student, who wrote a paper investigating the theories linking economic inequality and democratic transitions and consolidation (DelaO 2003), has explored the possibility of using the source materials on election campaigns to operationalize and test one of the basic assumptions of these theories – that elections are about the redistribution of economic resources. She developed a coding protocol for coding redistributive versus non-redistributive election campaigns that can be used to test theories of the relationship between economic inequality and democratic breakdown (Acemoglu and Robinson 2004, Boix 2003). CDEI could also be used to test predictions about whether the nature of political rhetoric (promising public rather than private goods, for example) has an effect on the nature of political governance. These data may be more useful for this purpose than existing party data on party programs, which rely on the written manifestos of political parties rather than an analysis of their actual rhetoric. These variables can be constructed fairly efficiently, since construction requires only that the researcher can use the archives that support CDEI rather than having to construct them from scratch.

6. Frequently asked questions about CDEI

This section raises some important questions about the quality and use of the data in CDEI for an analysis of democratic breakdown and other dependent variables.

(i) How much fluidity do we really expect there to be in ethnic structure or ethnic practice in the short term? Does it justify the rejection of previous data and a substantial investment in CDEI? There is a catch-22 in this question: we will only know if there is enough fluidity to justify the effort once we have made the effort and collected data on fluidity. But there is a substantial body of work that indicates that the effort is worth undertaking. Consider some examples: In Puerto Rico, the majority of the population changed from “negro” or “mulatto” to “white” over fifty years, despite the fact that there was no significant change in immigration and fertility rates. This shift in the aggregate pattern of ethnic diversity was the result of a reclassification of individual identities. (Dominguez 1997, 267). In Brazil, the opposite happened -- many of those who identified themselves as “white” or “black” in 1960 switched to calling themselves “brown” by 1990. The result was the transformation of Brazil in thirty years from a white to a non-white majority nation. (Nobles 2000, 105). In Zambia, many of those who identified with their tribe under a one-party regime switched to identifying with their linguistic group under a multi-party regime. The result was the transformation of an ethnic landscape initially composed of over seventy small ethnic groups to one composed of four large ones each time the nature of the regime changed. (Posner 2005) In Sri Lanka, many of those who had hitherto called themselves “Kandyan” and “Low Country” abandoned these regional identities in the course of twenty years to unite in a cohesive “Sinhala” identity. The result was the transformation of Sri Lanka’s multipolar ethnic demography into a bipolar one in less than two decades. (Tambiah 1986, Rajasingham 1999). As these examples, all of which indicate large-scale identity change in the short term, proliferate in the constructivist literature, we must take them seriously in our data collection efforts in political science. CDEI is one effort to do that, but there are several others which I discuss below.

(ii) What are some of the biases in CDEI and how do they affect analyses using these data? One bias in the data has to do with the selection of countries. The countries initially included in CDEI were those which received a high score on Keefer’s scale of legislative competitiveness for

1996. I employed this selection rule based on the reasoning that since I was studying the breakdown of democracies, I needed only to track the pattern of party behaviour in countries which met the minimal criterion for democracy (i.e. relatively free and fair electoral competition) in the first place. But this selection rule means that I cannot examine whether the degree of competitiveness in a democracy is itself a variable which affected the effect of ethnic parties. Further, it introduces selection bias by eliminating from the dataset countries in which ethnic parties may have been responsible for democratic breakdown in the past. To correct for this bias, I have now begun to expand the country coverage of CDEI to include all countries for which we can obtain vote share data regardless of the degree of competitiveness (in addition to the 100 countries already coded, several additional countries are in various stages of completion) and I expect to add more as we expand the dataset across time. The time-series expansion is especially important in allowing us to identify whether ethnic parties were the cause of the breakdown of democracy in countries that are dictatorships at any given point in time.

There is also a selection bias in the parties included in CDEI. CDEI has data on individual parties at a greater level of disaggregation than any other cross-national database that I am aware of. Nevertheless, even when we are able to obtain perfectly disaggregated vote shares, we exclude tiny parties which get less than .01% of the vote. And when we are not able to obtain perfectly disaggregated vote shares, and there are large residual categories for “other” parties, we can exclude parties which obtain more than .01% of the vote. Further, highly disaggregated data on vote shares is available in some regions (e.g. Europe). Thus, the selection bias is likely to vary systematically by region. This bias may not matter much for the current project, which measures the effect of ethnic parties on democratic stability. But I have found this bias to be important in a different project, on explaining the variation in the success of ethnic parties across countries (Chandra, Gisselquist, Metz, Wendt and Ziegfeld 2005) because it results in the systematic underestimation of the failure of ethnic parties. We address this so far in two ways – first, by coding all parties that we know competed in an election campaign, even if we do not have vote share on them. This way, although we cannot enter these parties in a measure of EVOTE, we can tell if the excluded parties have any ethnic parties among them. Second, we calculate an upper bound on the magnitude of the bias, based on the percentage of vote share that is excluded, and see if the results change if adjust the estimates of EVOTE accordingly. The documentation for the dataset identifies other possible biases and errors for particular types of analyses and provides suggestions on how to compensate for them.

(iii) Isn't the definition of an ethnic party -- as a party that makes an explicit, exclusive and central ethnic appeal -- used here too narrow? Why exclude implicit ethnic appeals, and a party's support base, from the definition? In principle, an ethnic party could be defined by its explicit and/or implicit message, and/or its support base and/or its candidates and/or its policies. There are no absolute criteria by which one such definition would be more “correct” than another. But my approach to definition here is driven by what I want to explain. The more capacious the definition of a concept, the less it allows us to explain, since anything included in the definition of a concept must be excluded from the dependent variable explained by that concept. Thus, I define and collect data on an ethnic party and every other concept in CDEI in narrow terms. By defining an ethnic party narrowly based on its explicit message, I am able to test both theories of whether explicit messages have an effect on support base(which I could not do if I included the support base in the definition) and whether explicit messages have a different effect on support from implicit messages (something that I could not do if I put explicit and implicit messages in the same

conceptual category). I am also able to ask whether the explicit mobilization of ethnic identities has a different effect on democratic stability than implicit mobilization.

(iv) Isn't the definition of an ethnic group too broad? Are regional identities and racial identities the same thing as religious identities? And how is it operationalized? The definition emphasizes "inherited" attributes, but categories based on region and religion are often not inherited. I employ a broad definition of ethnic identity in order to be consistent with the theoretical and empirical literature that I follow, where it has become conventional to employ an umbrella definition of ethnic identity (see for instance Horowitz 1985). Unless I use the term in the same way, I will not be able to evaluate claims made by this previous body of literature. However, CDEI further codes each ethnic category according to whether it belongs to the dimension of race, ethnicity, nationality, tribe, language, caste, religion, region, and so on, or to some combination of these dimensions. Thus, the researcher can investigate whether identities classified in the same family as "ethnic" are actually associated with the same outcomes by conducting separate analyses on each type.

We operationalize this definition in our coding criteria in two steps: as a first cut, we treat all categories based on race, ethnicity, nationality, tribe, language, caste, religion, region, and so on as ethnic categories, and use this to make judgments about whether the party in question is making an explicit or implicit ethnic appeal or has an ethnically identified support base and so on. (Our data sources routinely name such categories but are typically insufficient to determine the rules for membership in that category.) For those categories where we are in doubt about whether the rules for membership require inherited attributes or not, we then refer to country-specific information, trying as far as possible to rely on primary documents produced by political parties. In this second step, we eliminate categories that appear not to meet the descent based criterion.

(v) Isn't the range of variables collected here too narrow? What about other demographic variables germane to the performance of ethnic parties such as the regional concentration of ethnic groups, and institutional variables such as party registration rules, electoral laws, federalism, consociationalism and so on? I try to focus in CDEI only on those variables which cannot be obtained from other sources. Thus, we collect data on government laws that restrict the political activation of ethnic identities, which we cannot otherwise obtain. We also expect to create a variable on the regional concentration of groups, by using existing data from MAR and Fearon and Laitin, and filling in the gaps based on our own data sources. There are also a range of spin-off variables that can be created from our existing archives if they become important to the analysis. But we also expect to take advantage of the explosion of high quality data collection projects in comparative politics in the last decade on several of the institutional variables germane to the performance of ethnic parties can now be imported from other datasets – e.g. electoral rules (Golder 2005), district magnitudes (Golder 2005), federalism (Rodden 2002, 2004), institutional structure (Przeworski et al, Keefer 2002), economic development (Banks) – or constructed by combining variables from these other datasets– e.g. consociationalism.

There are several other questions that I do not address here because of the constraints of space: Is there a connection between the conditions that give rise to ethnic parties and their effect on democracy? Relatedly, is there endogeneity in the relationship between the rise of ethnic parties and democratic stability? What are the consequences of focusing on national rather than regional elections across countries? What is the pattern of missing data and how should it be interpreted?

These are all more fully addressed in the list of FAQs accompanying the dataset, which I would be glad to supply on request.

7. Conclusion: Other Routes to “Constructivist” Data Collection

Although EVOTE and other variables generated by CDEI can be put to broad use, CDEI is not the only way in which a constructivist approach to ethnic identity might be employed in data collection. As I have tried to emphasize in this chapter, it is a very specific route, focused on the behaviour of political parties, that produces several very precisely defined variables that measure some aspect of ethnicity. But progress in the field of ethnic politics depends upon diversified attempts by several scholars to collect data on these and other variables.

Several other datasets informed by a constructivist perspective collect data on variables located elsewhere on the conceptual tree breaking down the many concepts related to “ethnicity.” Posner and Scarritt and Mozaffar have independently compiled data on politically significant ethnic categories in Africa (Posner 2004, Scarritt and Mozaffar 1999). The Scarritt and Mozaffar data is unique among cross-national datasets in coding categories at multiple levels. The MAR (Minorities at Risk Dataset), initially time-insensitive, now updates its data periodically, tracking changes in the composition of ethnic groups included in the dataset. Wittenberg and Kopstein use Gary King’s Ecological Inference (EI) method to track voting patterns among ethnic groups in Eastern Europe. EI is especially useful from a constructivist perspective because it allows the researcher to impose different ethnic categories on a population, and investigate which category predicts voter behaviour, rather than imposing ethnic categories on the analysis *ex ante*. Taeku Lee is exploring variation in the self-identification of voters in the U.S. using an innovative survey design that allows voters to distribute “identity points” across a range of ethnic identity categories (Lee 2004). Steven Wilkinson is opening up avenues for collecting constructivist inspired data on the activation of identities in non-institutionalized contexts, by constructing a time-series dataset on all incidents instances of non-institutionalized collective action in post-colonial India, including riots, strikes, demonstrations, and coding each event according to all the identities that are relevant to describing it. (Wilkinson 2005). Each of these datasets measures some distinct aspect of ethnicity in some context not included in CDEI. Further, while CDEI is cross-national in its coverage, each of these datasets, with the exception of MAR is focused on a single country (U.S., India) or region (Eastern Europe, Africa). Studies based on data in CDEI would be complemented and deepened by further analyses based on these data. In particular, the data on voting patterns in individual countries should allow us to test the implications of the patterns of party behaviour revealed through CDEI.

Constructivist approaches to data collection are currently in their infancy, and can only improve and increase over time. Ultimately, we will know when we have made progress when the all-purpose concepts, variables and datasets that currently dominate the field have been replaced by several narrow concepts, variables and datasets that compete with and complement each other.

REFERENCES

- Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg (2003) "Fractionalization" Journal of Economic Growth. 8.2 : 55-94.
- Atlas Narodov Mira (1064). Moscow: Miklukho-Maklai Ethnological Institute of the Department of Geodesy and Cartography of the State Geological Committee of the Soviet Union.
- Birbir, Johanna (2004). The Ethnic Effect. Forthcoming, Cambridge University Press.
- Brass, Paul (1974). Language, Religion and Politics in North India. Cambridge: Cambridge University Press.
- Chandra, Kanchan (2001) "Cumulative Findings in the Study of Ethnic Politics." APSA-CP 12.1: 7-11.
- Chandra, Kanchan (2004a). Why Ethnic Parties Succeed: Patronage and Ethnic Headcounts in India. New York: Cambridge University Press, 2004.
- Chandra, Kanchan. (2004b) "What is Ethnicity?" Paper presented at Harvard Conference on Identity, December.
- Chandra, Kanchan (2005a). "Ethnic Parties and Democratic Stability." Perspectives on Politics.3.2: 235-252.
- Chandra, Kanchan (2005b), "Ethnic Diversity and Democratic Stability," Manuscript in Progress.
- Chandra, Kanchan (2005c). "A Constructivist Dataset on Ethnicity and Institutions." In Rawi Abdelal, Yoshiko Herrera, Alastair Ian Johnston, and Rose McDermott eds, Identity as a Variable. Manuscript in Progress.
- Chandra, Kanchan, Rachel Gisselquist, Daniel Metz, Chris Wendt and Adam Ziegfeld. (2005) "The Weakness of Explicit Ethnic Appeals." Paper prepared for presentation at the 2005 meeting of the American Political Science Association.
- Collier, Paul. (n.d.) "Implications of Ethnic Diversity." Working Paper.
- Collier, Paul and Anke Hoeffler (2001). "Greed and Grievance in Civil War." World Bank. Typescript.
- Dahl, Robert. (1971) Polyarchy. New Haven: Yale University Press.
- Dahl, Robert. (1956). A Preface to Democratic Theory. Chicago: University of Chicago Press.
- Dominguez, Virginia. (1997). White by Definition. New Brunswick: Rutgers University Press.

Elbadawi, Ibrahim A. and Nicholas Sambanis. (2000). "Why are There So Many Civil Wars in Africa? Understanding and Preventing Violent Conflict." Journal of African Economies 9.3: 244-269.

Fearon, James (2003). "Ethnic Structure and Cultural Diversity by Country" Journal of Economic Growth 8.2: 195-222.

Fearon, James and David Laitin (2003). "Ethnicity, Insurgency and Civil War." American Political Science Review 97.1: 75-90.

Geertz, Clifford. (1973) "The Integrative Revolution: Primordial Sentiments and Civil Politics in the New States." The Interpretation of Cultures. New York: Basic Books, 255-311.

Gilroy, Paul. (1996). "One Nation under a Groove: The Cultural Politics of "Race" and Racism in Britain," In Geoff Eley and Ronald Grigor Suny, eds. Becoming National. New York: Oxford University Press. pp. 352-369.

Golder, Mathew. (2005). "Democratic Electoral Systems Around the World 1946-2000." Electoral Studies 24: 103-121.

Horowitz, Donald. (1985) Ethnic Groups in Conflict. Berkeley: University of California Press.

Kasfir, Nelson. (1976). The Shrinking Political Arena. Berkeley: University of California Press.

Kasfir, Nelson (1979). "Explaining Ethnic Political Participation." World Politics 1.3: 365-388.

Keefer, Philip. (2002). "DPI Database of Political Institutions: Changes and Variable Institutions." Typescript, World Bank.

Laitin, David and Daniel Posner. (2001). "The Implications of Constructivism for Constructing Ethnic Fractionalization Indices." APSA-CP 12.1:

Lee, Taeku (2004). "Between Social Theory and Social Science Practice: Towards A New Approach to the Survey Measurement of Race." Paper presented at Conference on Identity, Harvard University.

Lijphart, Arend (1977). Democracy in Plural Societies. New Haven: Yale University Press.

Mendelberg, Tali. (2001). The Race Card: Campaign Strategy, Implicit Messages and the Norm of Equality. Princeton NJ: Princeton University Press.

Mill, John Stuart. (1991). Considerations on Representative Government. New York: Prometheus Books. Originally Published 1861.

Minorities at Risk Database (<http://www.cidcm.umd.edu/inscr/mar/>)

Mozaffar, Shaheen and James Scarritt. (2002). "Electoral Institutions, Ethnopolitical Cleavages, and Party Systems in Africa's Emerging Democracies." American Political Science Review. 97.3: 379-390.

Nobles, Melissa (2000). Shades of Citizenship: Race and the Census in Modern Politics. Stanford, CA: Stanford University Press, 2000.

Posner, Daniel (2004). "Measuring Ethnic Fractionalization in Africa" American Journal of Political Science, 48.4: 849-863.

Posner, Daniel. (2005). The Institutional Origins of Ethnic Politics in Africa. Cambridge: Cambridge University Press.

Przeworski, Adam et al. (2000). Democracy and Development. Cambridge: Cambridge University Press.

Rabushka, Alvin and Kenneth Shepsle. (1972). Politics in Plural Societies. Columbus, Ohio: Charles E. Merrill.

Rajasingham-Senanayake, Darini. (1999). "Democracy and the Problem of Representation: The Making of Bi-polar Ethnic Identity in Post/Colonial Sri Lanka." Ethnic Futures, Joanna Pffaff-Czarnecka et al., Eds. New Delhi: Sage Publications, 99-134.

Rodden, Jonathan and Eric Wibbels. (2002). "Beyond the Fiction of Federalism." World Politics 54.3.

Rodden, Jonathan. (2004). "Comparative Federalism: Meaning and Measurement." Comparative Politics, 36.4: 481-500.

Rustow, Dankwart. (1970). "Transitions to Democracy: Towards a Dynamic Model." Comparative Politics 3.2: 337-364.

Sambanis, Nicholas. (2001). "A Review of Recent Advances and Future Directions in the Quantitative Literature on Civil War." Working paper.

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.

Tambiah, Stanley. (1986). Sri Lanka: Ethnic Fratricide and the Dismantling of Democracy. Chicago: University of Chicago Press.

Wilkinson, Steven. (2005). "Which group identities lead to conflict? Evidence from India," in Stathis Kalyvas and Ian Shapiro eds. Order Conflict and Violence, Under Review

Wittenberg, Jason (2004). "Ethnic Diversity and Electoral Extremism in Interwar Eastern Europe." Paper Presented at the Laboratory in Comparative Ethnic Processes, Columbia University.