

# **Nations and Nationalism Exchange on the Quantitative Measurement of Ethnic and National Identity**

Daniel Bochsler,\*/† Elliott Green,\*\* Erin Jenne,\* Harris Mylonas\*\*\* and Andreas Wimmer\*\*\*\*

\* Central European University, Vienna, Austria

\*\* London School of Economics, London, UK

\*\*\* George Washington University, Washington, DC, USA

\*\*\*\* Columbia University, New York City, USA

† University of Belgrade, Belgrade, Serbia

## **Abstract**

In recent decades there has been a noticeable increase in the use of quantitative techniques in the study of ethnicity and nationalism, without, however, a sufficient amount of reflection on how these techniques have contributed to our understanding of ethnic and national identities. As such in this exchange three sets of authors answer questions about the degree to which it is possible to use quantitative data to measure ethnic and national identities, which types of methods are most suitable in measuring these identities, and what the major research findings of this quantitative research are that were not possible using qualitative approaches.

**Key words: Census, ELF, Ethnic Identity, National Identity, Quantitative Methods, Surveys**

## **Introduction – Elliott Green**

There has been a well-documented phenomenon by which the social sciences have taken a quantitative turn in the past half-century. Due to the rise of new quantitative techniques, better computing power and new datasets, there has been a sharp rise in the use of statistical analysis within political science since the 1960s – initially “lower-level” methods such as percentages, means and correlations, but then followed by “higher-level” use of multivariate regressions, multi-level modelling, instrumental variables, difference-in-difference and regression discontinuity techniques, among others (Sigelman 2006). This trend has been pronounced in sociology as well and has similarly become important in the study of economic history, where it has been known as cliometrics (Greif 1997). Indeed, rather than detail the history by which each discipline in the social sciences has become increasingly quantitative in focus, it might be easier to note the exceptions to this trend, such as in British anthropology and sociology as well as in (non-economic) history (Platt 2012). It is also important to note the degree to which mixed-methods research has become important in applied social science (Alise and Teddlie 2010), with a turn away from a pure quantitative focus evident in the past towards one that integrates both qualitative and quantitative techniques, albeit not in every discipline (such as economics).

This rise in the use of quantitative techniques is evident as well in the social scientific study of ethnicity and nationalism, inasmuch as there has been a recent rise in the proportion of publications that use regression analysis in particular to study variation in ethnic and national identities. Indeed, whereas an older generation of books on nationalism were entirely qualitative in their methods – such as Anderson (1991 [1983]), Gellner (2006 [1983]) and Smith (1991) – more recent work such as Cederman (1997), Laitin (2007) and Wimmer (2018) has drawn hugely upon quantitative evidence and computer modelling as a means to answer questions about national identity formation. This trend has been evident as well in empirical work on individual countries or regions, as with work on Africa (Ahlerup et al. 2017, Depetris-Chauvin et al. 2020, Green 2020, Robinson 2014), the Balkans (Mylonas 2012) and India (Singh 2015), for instance.

In particular this literature has drawn from the increased availability of surveys that ask questions about national identification and national pride, as well as datasets that collate national census data together for the use of scholars. In the former case the World Values Survey, which has been conducted since the 1980s, as well as regional surveys such as the Afrobarometer (since the late 1990s), Latinobarometer (1990s) and Asiabarometer (2000s) have asked questions about ethnic identity, national pride and/or national identification, thereby generating large amounts of repeated cross-sectional data. These surveys have changed the kinds of questions they ask in different rounds – in the case of the Afrobarometer, a question asking respondents to identify their most salient form of identity (with choices including ethnic and national identities as well as region, class and gender) was replaced in its third round by one that only compared ethnic and national identities – but the sheer volume of observations contained within these datasets has been a gold mine for scholars. Since the late 1980s there has similarly been a huge rise in the availability of Demographic and Health Survey (DHS) data, where individuals (mostly women age 15-49) answer survey questions about a variety of topics, one of which is their ethnic identity. For a number of years scholars have used DHS data extensively in the literature on ethnic favouritism, while more recently Green (2021) used DHS data from 14 African countries to show how a change in the ethnic identity of the President caused individuals to switch ethnic identities to that of the President, most likely in anticipation of receiving resources targeted to the President's kin.

In the case of censuses, the United Nations issues regular Demographic Yearbooks since the 1950s that contain individual country census data on ethnic and racial identities, with data listed online from 1995 by the United Nations Statistics Division. Census data is, of course, more subject to political manipulation than surveys, in that it determines the allocation of public goods; as such a number of countries no longer ask questions about ethnicity or race, have radically changed the nature of such questions over time or have seen censuses delayed or even dismissed as inaccurate due to controversies around how questions on ethnicity and race are asked (Morning 2008). Nonetheless, census data has been a fruitful source for scholars asking questions about the dynamics of ethnic and racial identification in a number of contexts, as in Africa (Posner 2005), India (Cassan 2015) and the United States (Nix and Qian 2015).

This increasing trend in the use of quantitative data to study ethnicity and nationalism has not, however, been without controversy or debate. For instance, the way data on ethnic diversity has been compiled into a single number via the ELF (ethno-linguistic fractionalization) index – which is computed like a Herfindahl (firm concentration) Index, such that it ranges from 0 (totally homogenous) to 1 (totally heterogeneous) – has been criticized by a number of scholars as far too simplistic in understanding both the causes and consequences of ethnic diversity (Brown and Langer 2010, Platteau 2009). The result has been an ongoing debate about how to measure ethnicity and nationalism and how to design census and survey

questions that accurately capture how individuals identify (Aspinall 2012, Burton et al. 2010, Chandra and Wilkinson 2008).

This exchange thus attempts to grapple with how, when, why and whether to measure ethnicity and nationalism using quantitative techniques. The following three responses address three specific questions on this topic:

1. *Is it possible to capture multiple/dynamic ethnic and national identities with quantitative data?*
2. *What is the best method(s) to measure ethnic and/or national identities?*
3. *What are the major research findings of quantitative approaches to the study of ethnicity and nationalism that were not possible using qualitative approaches?*

In the first response, the political scientists Erin Jenne and Daniel Bochsler, who are both based at the Central European University, tackle each question individually. They provide an overview of the types of quantitative data on ethnicity and nationalism available to researchers, and highlight the problems encountered in quantification, illustrated by the exemplary case of Belgium. Jenne and Bochsler conclude by making a case for the continued use of quantitative data on ethnicity and nationalism, especially as it continues to evolve and capture more detailed, multidimensional and dynamic understandings of identity.

The second response comes from the political scientist Harris Mylonas, who, like Jenne and Bochsler, has written extensively on ethnic and nationalist politics in Europe, as well as a variety of other contexts. Mylonas discusses the myriad ways in which quantitative techniques capture ethnic and national identities, including the use of ethnography, indices and censuses. He then focusses on how scholars can produce higher quality data, such as not measuring identities during a period of conflict and relying upon older census data. Finally, Mylonas shows that statistical analysis has greatly added to our knowledge of the relationship between ethnicity and civil wars, especially as scholars have continued to use multiple techniques to understand this relationship.

In the third and final essay, the sociologist Andreas Wimmer answers the three questions together, and focusses on three types of data, namely naturally-occurring observations, large-scale surveys, and expert codings. In the first case he highlights the increasing attention to data from Google, Twitter and Facebook, which have collectively generated a variety of findings in recent years; in the second he discusses the use of multiple survey instruments with different types of design; and in the third he discusses how the use of expert codings of ethnic identification have evolved from the Soviet *Atlas Narodov Mira* in the 1960s to the Ethnic Power Relations Dataset more recently.

These essays arguably collectively agree on a number of conclusions. First, the use of quantitative techniques has come very far in the past few decades, with many different types of data upon which scholars can now draw. Second, quantitative data, as with qualitative data, has numerous problems and pitfalls, and thus researchers will need to continue to be careful about how and when they use such data. Third, there remains a huge scope for future researchers to refine older datasets and create new ones, all of which should add greatly to our understanding of ethnicity and nationalism.

We now turn to each essay and their answers to the questions posed above.

### **Erin Jenne and Daniel Bochsler**

1) *Is it possible to capture multiple/dynamic ethnic and national identities with quantitative data?*

Scholars of nationalism and ethnicity have long criticized quantitative researchers for measuring multi-layered, fluid ethnic and national identities with simple and static variables. We recognise that social categorisation is based on several, sometimes nested or cross-cutting layers of identities (e.g. religion, race, language) (see, for example, Posner 2004). Some individuals might combine several identities of the same category (e.g., bilingualism), and identify with different groups in different contexts. Yet the widely-used static and categorical measures capture at best “provisional stabilities” that mask the extent to which identity is contested, fragmented, unstable and multiple (Suny 2000). In this short essay, we first explain the function of static measures of identity in large-N analysis and then suggest how more fluid measures might be devised for use, provided that we have the appropriate data for doing so.

Quantitative studies using static, categorical variables are designed for a specific function, namely testing whether identities and the politicisation of identities can help account for variation in conflict behavior across individuals, groups and even states. To answer these questions, quantitative researchers must bracket questions about the origins, functions or boundaries of group identities in order to create large-N databases that can be used to answer why certain individuals or groups mobilise along certain identity-based lines, but not others. Other researchers take states, regions, or municipalities as the units of analysis to account for variation in the political use of ethnicity at the subnational level as well as the effects of ethnic diversity or polarisation across societies.

Today, most large-N conflict research relies on a few global databases on ethnic identities and politics (Bruk and Apenčenko 1964; Fearon 2003; Cederman et al. 2010; Birnir et al. 2015; Alesina et al. 2003; Gurr 2005; Vogt et al. 2015). All of these datasets utilize limited, categorical concepts of ethnicity that facilitate comparisons of identity and behavior, either at the level of individuals, or aggregated to groups or states. Some sources are based on primordial concepts of ethnicity, such as language as a mother tongue (Easterly and Levine 1997). Others report several dimensions of identities, such as language, race or religion (see Lambert 2005 for an overview). Studies at the individual level rely on data gathered by researchers through surveys with representative samples, which are often replicated across many countries. Cross-national surveys such as Eurobarometer, the European Social Survey (ESS), the World Values Survey (WVS) and the DHS ask survey respondents either to rank their national identification or rate their strength of identification (Sinnott 2006). The set of categories and dimensions of identity is usually limited to some nationally salient categories.

However, studies at the individual level have done more than simply “freeze-frame” ethnic identities. Innovative survey questions have been created to capture the multi-layered, nested and fluid nature of identities. Such measures have been around for decades. Linz and Moreno asked respondents to weight dual (regional-state) identities, asking whether they identify as a) only regional, b) more regional than national, c) equally national as regional, d) more national than regional or e) only national (Linz et al. 1981). Their question has been widely used, either for regional vs. state, or – in the Eurobarometer – for (nation-)state versus supranational (European) identities. It is designed to capture the trade-off between supra-national, national and ethnic identities within a single society over time. However, the Linz-Moreno scale has been subject to criticism because it deals with these identities as if they were competing, yet they are often not mutually exclusive (Burton et al. 2010; Levy 2014). It also fails to offer respondents the option to express low salience of both identities (Guinjoan and Rodon 2016: 129; Sinnott 2006). This conflation of multi-dimensionality and identity

salience can be solved by splitting the questions into two separate scales that measure the salience of each layer of identity (Guinjoan and Rodon 2016). Pollsters can separately measure multiple dimensions of identity, such as race, ancestry, religion or language (Burton et al. 2010).

While traditional survey research is based on static measures of identity, the analysis of identity change calls for individual-level panel data across different situations in life. The well-established experimental research tradition in this field (e.g. Sherif et al. 1961) is likely to be complemented by the growing availability of big data through which individuals can be tracked over time. In the future, researchers will need to establish new methods and procedures to aggregate information from big data on changing identities at the level of municipalities, regions and countries.

Despite these important innovations, a lot of quantitative research still relies on population censuses or other state-collected data, and the lack of a common standard for the collection of data on ethnic identities is an ongoing challenge (Aspinall 2007). Hence, the power to define identities is not in the hands of scholars, but rather states. Some state censuses are restricted to a set of constitutionally recognised ethnic categories. What is more, one out of three countries worldwide do not even record the ethnic identities of their citizens, although they often gather data on related aspects such as native languages (Kertzer and Arel 2001; Bochsler and Schläpfer 2016; Simon 2012).

The UN Statistics Division (2008: 139-40) recommends recognising the multidimensional nature of ethnicity in census questions. However, we do not know of any state that applies census questions that address fluid, multi-level and nested identities. Some countries are moving incrementally in this direction. For example, the US census counted only three ethno-racial categories in 1860, but hundreds in 2000. Today, the US allows respondents now to (self-)identify with multiple categories (Lee 2009: 114-7). In order to capture multiple layers of identities in censuses, Lee (2009) tested a new “identity point allocation” system whereby respondents allocate 10 identity points to multiple ethnic and racial categories. It must be said, however, that this point-allocation-question—similar to the Linz-Moreno question—does not fully capture fluid and multi-layered identities: the numeric implementation is based on the idea that multiple layers of identities are always competing, and never complementing, each other.

The range of innovative methods of measurement shows that the quantitative researchers have recognised the importance of capturing fluid and multi-layered identities.

## 2) *What is the best method(s) to measure ethnic and/or national identities?*

Ethnic and national identities are not individual but collective identities. Quantitative researchers seek to collect and analyze individual-level data in order to map the shape of these collective identities across societies or groups. The question for researchers is whether we can generate population-wide measures that allow us to aggregate and compare the ethnic structure across countries, regions, or capture social change over time without the need to assign individuals to a single category.

We believe that we can. And we can even use these data to measure important concepts in the field such as ethnic polarization and fractionalization. However, researchers must do more to develop appropriate, context-sensitive indicators that can be used for this purpose. There is no single “best” method for achieving this; much will depend on the research question. There are indicators that can be used to measure ethnic closure (or permeability), the stability/change and salience of ethnic identities (a theoretical suggestion: Wimmer 2008;

exemplified: Eschbach et al. 1998), but we need innovative, comparative datasets to measure these indicators across countries for the purpose of large-N analysis.

Often we are interested in determining whether ethnic diversity or polarisation can explain differing levels of conflict across countries. To test for this, we can choose from among a well-established family of measures to explain nationalist radicalisation (e.g. Key 1949; Glazer et al. 1998; Bochsler 2013) or political violence (Montalvo and Reynal-Querol 2005). Polarisation is defined as the (latent) potential for rivalries between two large groups, which is a function of the ethnic structure of the population. In highly polarised societies, the majority or plurality group is challenged by an almost equally large minority. Conventional indices of polarisation, such as the Reynal-Querol-Index (2002), are based on single-dimensional, categorical notions of ethnicity, wherein identities are stable over time. But how do we understand ethnic polarisation in the context of multi-layered, nested, and fluid identities?

We are convinced that polarization, fragmentation and most other aggregate measures of national or ethnic identities are still relevant in a multi-layered and fluid world. Therefore, there is a large potential for innovative families of indicators of identities that move beyond categorical and static identities (Lieberman and Singh 2012 offer a review of this field). The case of Belgium illustrates a possible application. The country is divided between Catholics, Liberals (seculars), French-, Flemish- and German-speakers. Most Belgian citizens belong to two of these groups, a religious and a linguistic community (some even to three or more, if they are raised multilingually). Identities are multi-layered and allow for changing significance. Indeed, different cleavages have been mobilised in Belgium over time (De Winter and Dumont 1999).

In order to measure ethnic polarisation in an environment of fluid and multi-layered identities, we need to consider multiple potential lines of conflict. An index can be used to identify and score the conflict line with the largest potential to mobilise two large groups against each other (Esteban and Ray 1994). In Belgium, the two largest groups are the Catholics (57%) and the Flemish-speakers (55%). However, political entrepreneurs would fail if they tried to create a Catholic-Flemish conflict because most Catholics speak Flemish, and most Flemish-speakers are Catholic. A more likely line of rivalry is between the Flemish- and French-speaking communities: both are large, but the overlap between the two groups is small. Indeed, this identity-based divide is currently the most salient in the political arena: the Belgian party system is deeply divided between the two main linguistic groups. A multi-dimensional version of the polarisation index would need to take into account not only the two largest possible groups (along with multiple lines of conflict), but all possible pairings of the groups. Recognising that different layers of identities may be politicised, a novel index of ethnic polarisation could be devised that would identify and score each line according to its conflict. This new index would recognise that the fiercest rivalries can emerge between groups with a low membership overlap.

Other relevant indicators that capture ethnic identities across countries while recognising the multi-layered nature of identities include cross-cuttingness (the degree of overlaps between identities) (Moerman 1965; Selway 2011a), ethnic diversity, and ethnic homogeneity (Laitin and Posner 2001). All of these concepts can, with some innovation, easily be expressed in numbers. However, they require systematic data on multiple layers of identities, which informs us about overlaps, for a large number of countries. Again, recent moves in data collections go in this direction.

3) *What are the major research findings of quantitative approaches to the study of ethnicity and nationalism that were not possible using qualitative approaches?*

Much of what we know about nationalism and ethnicity has come via the contributions of quantitative research. Researchers have used data analysis to show that ethnic identities play a vital role in electoral politics (Birbir 2007), citizenship and immigration policies (Manatschal 2012; Just and Anderson 2012), power-sharing (Cederman et al. 2010) minority claim-making (Jenne 2007; Jenne et al. 2007; Cunningham 2014; Saideman and Ayres 2000; Siroky and Hale 2017), political violence (Wilkinson 2004; Jenne 2007; Kalyvas 2008 Costalli and Ruggeri 2015; Balcells 2017), and right-wing extremism (Koopmans and Olzak 2004; Boutcher et al. 2017).

The breadth and depth of the state of the art belies the popular perception that quantitative methods are poorly equipped to shed new light on the politics of ethnicity or nationalism. We believe this popular misconception is based on at least three misunderstandings about quantitative work. The first is that all quantitative research relies on aggregate measures of identification. In fact, social psychologists have used experimental techniques to shed light on the individual bases of ethnic identities going as far back as the 1960s and 1970s (see Turner et al. 1979) Today, this research program is so well established that we can only pick a few illustrative sketches of notable experimental work on nationalism and ethnic identity.

Since the experimentalist turn in the social sciences some twenty years ago, randomized controlled trials (RCTs) have become a mainstream technique for showing how threat priming, for example, alters ethnic identification or ethnic policy preferences. For example, researchers have used survey experiments to show that exposing respondents to a nationalist rather than a neutral text makes respondents more likely to favour restrictive immigration policies (Sniderman et al. 2004) or affects how they identify on ethnic grounds (Hierro and Gallego 2018). Other research suggests that ethnopopulist foreign policy statements lead respondents to favour more hawkish military stances than nationalist statements alone (Zelman 2019). Survey-based research has also shed light on how a person's cognitive style might influence their identity formation (Rosenberg and Beattie 2019).

The second misconception is that standard quantitative techniques are incapable of capturing the complexity of ethnic and national identification. In point of fact, survey researchers have long engaged in investigating the components of national and ethnic identities as well as the relationships between them. Nuanced survey questions, such as the Linz-Moreno question, are now the methodological standard. Meanwhile, social scientists have used surveys to identify the correlates and components of national and ethnic identity. For instance, Shayo (2009) used WVS and International Social Survey Program (ISSP) to show that strength of national identity correlates with inegalitarian welfare preferences across societies. Others have isolated different components of national identity, demonstrating, for example, that national pride is associated with national attachment (Theiss-Morse 2009; McCrone and Surridge 1998). Psychologists, meanwhile, have used the Multigroup Ethnic Identity Measure (MEIM) developed by Phinney (1992) to measure ethnicity among adolescents. Multiple studies using MEIM have validated the bifactor structure of ethnic identity—identity exploration and group commitment—demonstrating the experiential basis of ethnic identification across a diverse range of ethnic groups (Roberts et al. 1999; Yap et al. 2014).

The third misconception is that quantitative research on nationalism and ethnicity is necessarily ahistorical and de-contextualized. In fact, a considerable body of quantitative research has demonstrated the social constructedness of ethnic and national identities by examining the intersection of indirect measures of identity with household and cadastral data, educational and linguistic data and other demographic characteristics. For example, Howe et

al. (2019) used historical sub-national data from the late Habsburg empire to show how the variable intersection of class and ethnic identities throughout the empire led political competitors to adopt variable electoral strategies. Other historically-minded quantitative researchers have processed household, electoral and other types of data to re-examine what we know about the interrelation between identity and war. Kopstein and Wittenberg (2011), for example, matched voting records with census data in interwar Poland to show that anti-Jewish pogroms were more likely to occur in politically polarized local communities. Likewise, Charnysh and Finkel (2017) analysed communities close to the Treblinka concentration camp in Poland, revealing that localities with higher levels of Jewish property transfers were more likely to support an anti-Semitic party in 2001. In another **macro**-level study, Bormann et al. (2017) assessed whether different layers of identity – language or religion – can be used in similar ways to produce spirals of ethnic hatred and civil war. Much can be learned about the constructed yet embedded nature of identities through longitudinal analysis of databases based on cadastral and other types of quantitative data.

This cursory review suggests that there is a significant body of quantitative work that successfully deals with multi-layered, fluid and nested identities—far more than critics appear to believe. Moreover, traditional quantitative measures of ethnic and national identity very often suffice for answering many pressing questions in the field. At the same time, we should bear in mind that there is no one-fits-all approach: while identities might be multi-dimensional and fluid in some societies, they are sharp and stable in others. For the latter, simple, categorical measures may not only be adequate, but actually the most suitable for analysis.

## **Harris Mylonas**

1) *Is it possible to capture multiple/dynamic ethnic and national identities with quantitative data?*

It depends on what we mean by “capture,” on whether we are studying historical or contemporary developments, and on the purpose the quantitative data collection is gathered for. We can certainly quantify many aspects of social life and we can categorize individuals on the basis of “ethnic” or “national” identity categories. And, in theory, certain methodologies—survey research and/or experimental designs; ethnography; participatory observation—can also “capture” multiple or dynamic elements of such identities. However, there are several problems that are close to impossible to overcome without attention to conceptualization, research design and causal identification.

To begin with, scholars disagree over the conceptualization of ethnicity. Horowitz in his *Ethnic Groups in Conflict* used the term as an umbrella concept that “easily embraces groups differentiated by color, language, and religion; it covers ‘tribes,’ ‘races,’ ‘nationalities,’ and castes” (Horowitz 1985: 53). Chandra more recently proposed an analytically more useful definition: “ethnic identities are a subset of identity categories in which eligibility for membership is determined by attributes associated with, or believed to be associated with, descent.” (2006: 398). In her influential work ethnic identities do change but this pattern is constrained by an, often visible, set of attributes. The main way ethnic and national identities are different seems to be that ethnic identities do not necessarily involve a self-determination claim and that national identities do not necessarily involve qualifying features that are genetically transmitted.



Experimental designs have long been the standard methodology for social psychologists and it has been making inroads within political science as well. Certain survey research instruments can capture multiple identities co-existing, including ethnic, national, supra-national/regional, even cosmopolitan (Stelzl and Seligman 2009). As Burton et al. put it “while it is not possible to measure a complex, multidimensional, fluid concept using a single question at one point in time [...] it is possible for survey research to provide a much more sensitive measure of ethnicity, where variation over time and between parents and children in elements of a multi-dimensional measure are seen as assets rather than as problems for consistency.” (2010: 1344).

Ethnography and/or participatory observation is definitely another way to capture multiple, dynamic, and often contradictory ethnic, regional, religious, gender, class, and national identities. The episodic and often contextual character of ethnic and national identities is a key feature of works such as Lisa Wedeen’s *Peripheral Visions* or Roger Brubaker et al.’s *Nationalist politics and everyday ethnicity in a Transylvanian town*.

Experimental designs involve some degree of intervention, the “treatment” variables, which may distort respondents’ responses or perceptions, but the information collected is likely to isolate the effect of the “treatment” on self-identification/categorization. On the other hand, participant observation is less invasive, but it involves a significant amount of subjective interpretation.

An important aspect, which changes the degree of the challenge, is whether we are studying historical cases or if our interests are contemporary. If one is interested in studying historical cases, then the actors involved are long gone and depending on the literacy levels at the time, there may hardly be any available written records. Deciding which “groups” should be included in an analysis is a daunting problem in the study of nationalism and nation-building. I have called this the politics of “counting people”—referring to the difficulties and politics involved in three interrelated choices: identifying a group as an “ethnic group” or a “minority,” deciding on an estimate of its population, and studying it as a relatively unitary and homogeneous entity.” (Mylonas 2012 and 2015). Existing historical research cannot provide many of the answers we need and this is a politicized field thus we have a systematically biased picture of the situation on the ground. The most common methodological problems resulting from this practice are selection bias and over aggregated actors (for examples, see Mylonas 2015). The politics surrounding census statistics on “national minorities” or “ethnic groups” is an extremely problematic aspect in the study of the emergence and change of ethnic and national identities.

Cross-national statistical work in Political Science, Sociology, and Economics has relied on various indices, such as the Ethnolinguistic Fractionalization Index (Bruk and Apenčenko 1964), the Politically Relevant Ethnic Groups (Posner, 2004), Fearon (2003), Alesina et al. (2003), and so forth, to capture ethnic diversity. Most of these studies rely either on census categories or on material produced by the elites of “oppressed” groups.

As I have argued elsewhere (Mylonas 2015), whoever is responsible for the census often predetermines the results. The main objective of censuses is to “manage” the diversity by taxonomizing people into categories that facilitate governance. In most cases there is no way to know if the categories used in a census were the most salient ones for the people on the ground. In fact, state census committees can be seen as the very institutions that “flatten” the multiple ethnic and national identities of a population by choosing to highlight a particular dimension—religious affiliation, mother tongue, regional identity, racial categories, or ethnic origin—and thus producing an “ethnic structure” that fits the government’s goals. Census data on “ethnicity” should be understood as components of a state’s nation-building strategy. Certain groups are not represented in census categories in order to mute them. Moreover, we are likely to get deflated population estimates of certain non-core groups (in my work this term refers to any aggregation of individuals that is perceived as unassimilated by the ruling political

elite of a country) and no estimates at all for the ones that are targeted with assimilationist policies.

Scholars sometimes attempt to get a more complete picture of the ethnic makeup of a country, or a region, by relying on reports or memoirs produced by elites of non-core groups. This is definitely an important source, but we need to be careful since such actors have incentives to inflate the numbers of the group they represent and to favour particular characteristics over others. Of course, it is next to impossible to capture any of the non-core groups whose elites have been successfully co-opted, or those that have no literate elites. Selection bias is unavoidable in the study of ethnic diversity.

We can now turn to the *purpose* of the quantitative data collection. Most of these data collection projects try to capture “ethnic diversity” in a country or its “ethnic makeup” rather than “multiple/dynamic ethnic and national identities”. Thus, they probably would not count as successful examples of data collection even without the problems I have outlined above. However, more thoughtful data collection efforts have focused on coding “politically relevant” ethnic groups, using expert surveys, and including as many recorded groups as possible by lowering their inclusion threshold. Cross-national time-series datasets on “multiple/dynamic ethnic and national identities” may be especially hard--if not impossible--to be developed. But, some of the existing ones like the Ethnic Power Relations data set (Cederman, Wimmer, and Min, 2010) and the revamped Minorities at Risk project (Birnie et al., 2015) are adequate in order to answer particular research questions. In terms of research trying to capture individual behavior and/or attachments important advancements have taken place but we are still far from constructing instruments that can travel across countries and be used for multiple research purposes.

## 2) *What is the best method(s) to measure ethnic and/or national identities?*

Once again, I would argue that the best way to measure ethnic and/or national identities depends on *who* you are and *what* you want to capture, and at *what level of analysis*. It also depends on the nature of the research question.

For government officials, for instance, the goal may be to bring about a particular ethnic structure in their country or to silence the existence of certain ethnic or national identities. Thus, government officials can come up with census questions that achieve this goal as well as repress the expression of any other identity category or cleavage dimension in the public sphere. For elites of oppressed, under-represented or silenced ethnic or national groups, the goal is to get noticed, counted, and gain access to political power—ranging from representation to self-determination. Thus, they are likely to challenge governmental policies, reach out to international actors and seek help to gain visibility, and attempt to produce their own statistics. Clearly for these actors the best method to measure ethnic and/or national identities is the one that serves their goals. Scholars should always be vigilant and aware of these political struggles in the areas they study.

Turning to the ways that academic researchers have dealt with this, we can divide them on the basis of what they are studying, at what level of analysis, and which methods they are using. For instance, scholars who focus on cross-national statistical study of civil wars, public goods provision, regime type, party systems, or economic development tend to use one (or more) of the datasets on ethnolinguistic fractionalization I discussed above (e.g. Cox 1997; Fearon and Laitin 2003; Toft 2003; Alesina and Ferrara 2005; Benhabib and Przeworski 2006; Goldstone et al. 2010). These datasets to a great extent reproduce the biases that exist in censuses around the world. But there is a further problem, when it comes to cross-national

datasets trying to capture the ethnic makeup of a country. Most data-collection projects do not include small groupings of people since the cut-off point for inclusion in a dataset is often 1 percent of a country's population. The seriousness of this problem depends on the research question of interest. A straightforward way in which ignoring small groups is problematic is to think of the possibility that a country has 30 groups smaller than 1 percent, which cumulatively total 10 or 15 percent of the country's population, as compared with one that has just a couple groups below the 1 percent threshold.

The seriousness of such a coding strategy becomes obvious when we consider the possibility that this very fragmentation is often the goal of state policy as I describe in *The Politics of Nation-Building* (2012), rather than a mere "fact" of the ethnic landscape. The Minorities at Risk project is working on improving its dataset in this direction by significantly lowering the population threshold needed for inclusion (Birnie et al., 2015). The Ethnic Power Relations (EPR) dataset (Cederman Wimmer, and Min, 2010) has also moved the discussion further by looking at changes in power relations and constructing a typology of access to state power by explicitly relying on expert surveys. This was an important innovation in this part of the literature. The Uppsala Conflict Data Project (UCDP) has also innovated in various ways. Finally, Joel Selway (2011a) has recently produced a dataset which innovates over previous efforts by going beyond the coding of a single-cleavage characteristic of social structure, such as ethnic or religious fractionalization. He developed a new index of ethno-religious cross-cuttingness, derived from national public opinion surveys. This way it becomes possible to capture the relationship *between* cleavages.

Partially in response to these problems of over-aggregation scholars turned to the micro-level trying to unearth the microfoundations of political violence (Kalyvas 2006). Scholars that study the *internal* dynamics of civil wars took a more nuanced approach to identifying warring groups and often see ethnicity or national identity as more fluid (Staniland 2014) even epiphenomenal to civil war dynamics (Christia 2012).

Other scholars are interested in constitutive stories of particular nation-states so they often take a state perspective, thus focusing on how the ruling elites or organic intellectuals have defined a particular national identity, and the place of various ethnic groups may fit in it, as well as under what conditions these change (Mylonas 2012, Akturk 2012, Brand 2014).

But cross-national datasets capturing ethnic diversity face further challenges. Comparability across countries is complicated by the fact that statistical bureaus collect data on identity categories that are arrayed around different cleavage dimensions in each country and over time. And, as we know from the empirical record, once the decision concerning the cleavage dimension and the relevant identity categories has been taken, a particular incentive structure is created for individuals on the ground. The result is usually overreporting in favour of the core group. The same inhabitants of a particular territory could be aggregated with respect to language or religion or a notion of national consciousness, each time producing a different ethnic makeup.

There are some practices that can mitigate these problems. To begin with, one should avoid relying on data produced during conflict. Another approach involves looking at census data that was produced *prior* to significant political changes in a region. In the case of my study of nation-building in the Balkans, I turned to Ottoman censuses from the 19th century—which did not use national categories—as well as confidential diplomatic reports written prior to the Balkan Wars or World War I. Looking at confidential reports housed in archives provides a window to state policymakers' thinking and often "reveals" the existence of groups of people that rarely come up in censuses or newspapers. These public omissions are purposeful. Again, if the goal is to capture ethnic diversity in a country, this approach is better than relying

on the census, but one needs to keep in mind that state officials have their own agendas and mainly report on politically relevant groups. Conducting intercoder reliability tests by consulting academics, human rights activists, and investigative journalists is another way to overcome this problem.

Finally, there is a wide range of scholarship trying to study individual level attachments and their impact on behavior or attitudes. This type of scholarship utilizes surveys or survey experiments to answer its research questions. As Burton et al. put it "Subjective measures need to recognize both instabilities in affiliation and multiple affiliations." (2010: 1345). This field, primarily involving social psychologists, is trying to find the best instruments to allow individuals to "fit" in multiple identity categories based on different answers they provide and attributes they qualify for. However, coming up with a "valid measure of ethnicity" remains challenging.

### *3) What are the major research findings of quantitative approaches to the study of ethnicity and nationalism that were not possible using qualitative approaches?*

Quantitative political science necessarily focuses on more parsimonious arguments that can be represented in a statistical model specification. The arguments need to be falsifiable and generate several testable observations. On the other hand, qualitative approaches are often linked to particular observations of individuals, villages, regions, countries or global patterns. Various methods could be used by a qualitative researcher to collect their data including fieldwork, participatory observation, elite interviews, focus groups, archival research, or even the synthetic approach to an existing body of literature. Both quantitative and qualitative researchers utilise data collected through similar methods. For instance, World Bank indicators or household survey data are collected in the field. The primary difference between qualitative and quantitative approaches has to do with how they analyse these data.

Certain types of conclusions and insights cannot come from ethnographic or medium-N qualitative work. Also, some qualitative work suffers from selection bias, i.e. drawing general conclusions from a non-representative sample of cases. For example, Fearon and Laitin (1996) have used quantitative methods to show that peaceful and even cooperative relations between ethnic groups are far more common than is large-scale violence as suggested from journalists and much of the academic literature on ethnic conflict. In the 1990s, many research programs became more critical of value laden definitions of concepts, case selection followed standard criteria that avoided selection bias (Geddes 1990), data collection processes became more transparent, the hypotheses proposed were crafted so that they could be falsifiable, and the interpretation of phenomena no longer predated the analyses of data. As a result, a new comparative approach emerged studying ethnic conflict, political violence, and nationalism (Mylonas and Tudor 2021).

Given the limited space I have, I will focus on developments in the study of civil wars as illustrative examples. There has been a lot of theorising and testing with regard to the role that ethnic (and religious) fragmentation plays in the onset of armed conflict. Identity categories, ethnic or religious, have been hypothesised to have direct causal effects on the outbreak and duration of civil wars (Cederman et al. 2013; Cunningham and Weidmann 2010; Goddard 2006). In an influential article, Fearon and Laitin (2003) did not find any effect of ethnic or religious fractionalization on the onset of civil war after controlling for per capita income. For ethnicity to factor into such a scenario, the simple existence of many ethnic groups in a country would not suffice. The politics surrounding the various ethnic groups, the access to state and non-state resources, and the hierarchies of these groups were more likely to be linked to

conflict. Collier & Hoeffler found that only ethnic dominance—one ethnic group being the majority—increases the probability of civil war onset (2004: 588). Andreas Wimmer (2012), using the EPR dataset, convincingly demonstrated that the institutional setting within which ethnic *practices* take place is more important than the heterogeneity *index*.

Another claim that has been supported by statistical analyses is that ethnic and non-ethnic civil wars have different causes. Sambanis called for the disaggregation of ‘civil war’ as a category into identity (ethnic/religious) and nonidentity wars (2001). The assumptions driving the logic of this distinction is that identifiability is easier in ethnic wars and that the security dilemma is operative in such cases (Sambanis 2000: 455). Relatedly, Balch-Lindsay and Enterline (2000) found that ‘ethnicized’ civil wars were protracted conflicts. Cunningham et al. (2009) later on suggested that rebel group characteristics mediated the effect of ethnicity on war duration. What all of these approaches have in common is that they treat ethnic identities as exogenous to the conflict and assume their salience.

More recently, Selway (2011b) utilized a new dataset (which I discussed above) and looked at how social structure affects the onset of civil war. Going beyond the existing studies that focused only on single-cleavage characteristics of social structure, such as ethnic or religious fractionalization, Selway argued that models that do not take into account the relationship between cleavages (or cleavage structure) are biased. He developed a new index of ethno-religious cross-cuttingness, derived from national public opinion surveys, and found that ethno-religious cross-cuttingness accounts for variation in civil war onset when interacted with ethnic fractionalization. Relatedly, Kuo and Mylonas (2020) have recently proposed that once we consider the socio-political orders established by the respective nation-building experiences of countries experiencing civil wars, we can make better sense of the patterns of variation in the role of identity in civil war fractionalization and alliance formation.

All in all, ethnic identifiability and cleavage salience should both be treated as variables. As Kalyvas (2008) has shown, ethnic identities are not stable or unitary during civil conflicts, nor do warring groups and their leaders necessarily have congruent motivations. Statistical analyses unearthed interesting patterns and generated important puzzles that qualitative approaches could not have come up with; but, they ultimately required the help of ethnographic accounts, archival research, and fieldwork to discern the microfoundations of these processes.

## **Andreas Wimmer**

One of the more exciting developments in the social sciences is that issues long considered the exclusive domains of qualitative, interview or fieldwork based research methods have become accessible to quantitative investigation. This is true for the study of collective identities such as those related to ethnicity, race, or nationhood. As I see it, there are three approaches, not mutually exclusive, to study these forms of classification and identification in quantitative ways: on the basis of naturally occurring data (such as Twitter, google search terms, or Facebook pages), through large-scale surveys, and through codings generated by experts. In this essay, I will give some examples of these three approaches and discuss their advantages and limitations. Before I begin, I should note here that these three approaches are usually connected to different aspects of ethno-racial and national identities: surveys are mostly investigating the individual level, psychological relationship to these identities (the subjective identification with them); naturally occurring data often relate to the behavioral or discursive consequences of these identities; and expert-coded data are mostly exploring the macro-structural features of these identities (such as their relationship to political power or economic

wealth). In other words, the three different approaches are complementary from a substantial point of view.

#### 1) Naturally occurring data: Twitter, Facebook, et al.

With the advent of the internet and social media, many new sources of data have appeared that are relevant for the comparative study of ethnicity, race, and nationalism. For example, one can gain new insights into the study of racial resentment by analyzing google search terms (which are supposedly not subject to any social desirability bias) on a fine-grained geographic scale (Stephens-Davidowitz 2014). Another source is Twitter, which can be explored to study the effects of laws on anti-immigrant sentiment (Flores 2017), among other topics. Wimmer and Lewis (2010) use the Facebook pages of a cohort of American college students to study at which levels of segmental differentiation (to illustrate: a Japanese American, is an Asian American, is a person of color) friendship ties are clustering.

As this last example illustrates, naturally occurring data can take into account the complexity of multiple levels of ethnic (and other) forms of collective identification, if the data sources are rich enough to allow such differentiation. As the first example demonstrates, naturally occurring data have certain advantages over interview based methods as they are not necessarily as influenced by what individuals think others (including the interviewer) would consider a legitimate, good answer. The downside of these data is that it is not generated in view of the kinds of questions that social scientists are interested in. For example, how do we know that individuals googling the n-word after Obama's election are indeed motivated by racial resentment—or by an interest in learning about the resentment of others? How do we know that individuals who are making friends with others of the same background are actually motivated by ethnic homophily? In other words, many of these sources don't allow to measure ethnic, racial, or national identifications directly, but only its behavioral or discursive consequences. New advances in natural language processing, in the meantime, are likely to change this in the future. Using word embedding techniques, for example (Kozlowski et al. 2019), will allow researchers to extract relevant forms of identification and its meaning from large corpi of texts (such as political speeches or newspaper articles or emails).

#### 2) Surveys: From ISS to TeO

The second approach is based on surveys, which allows us to answer at least some of the questions that are difficult to access via naturally occurring data. Most famously, the International Social Survey Programme (ISSP) contained a module, run several times over the past two decades, about individual identities. It included a battery of questions about national pride, about whether individuals identify more with their ethnic community than with the nation (including a "Moreno version" of this question), and so forth. This data has been used very often, not the least to identify different variations of national identities, going beyond the well-established dichotomy opposing ethnic to civic forms of nationalism (Bonikowski 2016). While the ISSP data is available for "only" 32 or so countries, many other multi-country surveys contain questions relevant to the study of ethnicity, race, and nationalism. For example, Wimmer (2017) combined dozens of such surveys, including the various continental Barometer surveys as well as all available rounds of the World Values Survey, to explore the determinants of national pride.

Needless to say, many more surveys were conducted at the national level to explore questions of national and ethnic identity. This allows us to include more contextual questions (such as

those referring to specific everyday situations or political issues relevant only to the country at hand). In some surveys, researchers have even ventured beyond the typical questionnaire form, which consists in asking respondents how relevant identity x or y is for them, or how relevant identity x is vis-a-vis identity y (as in the Moreno question). This is obviously interesting but also very limiting, as it prompts individuals to think about a particular form of categorisation that they might otherwise not bother about much. The survey I have in mind is *Trajectoire et Origine* (Beauchemin et al. 2016), conducted with a representative French sample and an oversampling of immigrants. It contains a completely open identity question, allowing individuals to write in that they mostly identify with their trade (e.g. carpenter, sociologist), age or cohort (retired, generation X), region (Niçois), etc.

Even such extraordinarily rich survey data are obviously haunted by the limitations of the methodology: It captures attitudes and abstract identities, rather than observable behavior that is linked to such attitudes and identities; it suffers from social desirability bias; it is decontextualized (individuals think of their identity differently when entering the emergency room of a hospital than when they take a stroll in their neighbourhood); it prompts respondents to think about certain kind of identities, and so forth.

### 3) Expert codings: From ELF to EPR and beyond

The final type of data comes from administrative or expert sources. In the field of study we are concerned with here, the prime example is the ethno-linguistic fractionalization index. The original version is based on the linguistic maps assembled by Soviet ethnographers and linguists in the fifties and sixties (Bruk and Apenčenko 1964). In the meantime, other similar indices have been produced by economists (Alesina and La Ferrara 2005) and political scientists (Roeder 2007). With the exception of the latter, these indices are not time-varying, but represent snapshots at a particular point in time. On the basis of this data, other kind of indices can easily be constructed, including the often used index of polarization (which reaches the maximum value when two groups make up half of the population each) (Montalvo and Reynal-Querol 2005). If used to test mechanisms that are actually related to linguistic diversity, these measures are certainly useful (see e.g. Wimmer 2018). More often than not, however, they were used to test arguments related to ethnic inequality and grievances, to which the actual measures are obviously only tangentially related (Fearon and Laitin 2003): Tanzania has dozens and dozens of spoken languages, but none of them (with the exceptions of the Masai) are related to politically relevant groupings or claims.

This situation has recently changed with the advent of data that are tracing political processes of ethnic mobilization and contestation more directly. The Ethnic Power Relations Dataset (Wimmer et al. 2009) is a widely used dataset that codes all minimally politically relevant ethnic categories (including those marked by linguistic, religious, and other boundaries) and notes their level of political representation in national level government. Unlike the static fractionalization indices, EPR is fully compatible with constructivist principles: the relevant categories can (and do) change over time, groups can split and fuse, and so on. Still, the dataset does not allow for multiple, overlapping forms of identification, such that each citizen at any point in time is assigned to only one politically relevant ethnic group. The same is true for the older Minorities at Risk dataset, which like EPR is based on expert codings. It contains many more variables than EPR, but was originally limited to a subset of ethnic groups “at risk” of political persecution and exclusion, a classic case of coding on a dependent variable. This has recently changed and the dataset is now supposedly including all “socially relevant” ethnic groups around the world, irrespective of their political status or levels of politicization (Birnie et al. 2018). Both EPR and MAR leave it up to country experts to decide how individuals identify,

which categories are relevant for them, and how strongly these categories are related to behaviour. This obviously leaves lots of space for misunderstandings and misrepresentations, for all kinds of coder biases, and so on. On the other hand, however, such data is uniquely suited to represent the macro-structural correlates of ethno-racial diversity, including levels of political and economic inequality.

#### 4) Evaluation and outlook

As the short evaluative comments above make clear, it is obvious that the three quantitative approaches each have their specific advantages and disadvantages: naturally occurring data rarely allow us to measure identification, such that observers have to deduce these from behaviour; survey data allow to measure such identification directly, but it is unclear how far these are relevant and for which kind of behaviour; expert-coded datasets are best to explore the dynamics of ethnic politics, but at the prize of a simplified and one-dimensional understanding of identification. Ideally, therefore, one would combine multiple data sources together.

Such an ideal dataset would be coded at the individual level, record the traces of their behavior (as in tweets, Facebook posts, email content, networking behavior, voting, etc.), combine it with interview based statements about how they identify in which situations (using open, write-in questions wherever feasible), and relate all of this with other, “objective” or expert-coded data sources organized at the level of the ethnic and national groups mentioned in these interviews. This data should be organized as a panel, that is, observe the same individuals over—ideally—very long stretches of time and include the behaviours, identities, and associated traits of their children as well. Obviously, we are quite far away from such an integrated, multi-sourced and multi-generational system of data collection. If current advances in big data and online interview techniques continue into the near future, however, we should arrive there before too long.

It is now time to directly address the questions that the organisers of this symposium have posed to its participants. First, they ask if it is possible to capture multiple and dynamic ethnic and national identities with quantitative data? Obviously, quantitative data are at a disadvantage compared to ethnographies (see Brubaker et al. 2007) when it comes to capturing the context-dependent and dynamic nature of ethnic and national identifications. As always, there is a trade-off: Quantitative data are necessarily more static and less sensitive to context, but they come with the advantage of a bird-eye comparative horizon and the promise of generalisation. Some datasets related to ethnic and national identities are dynamic (e.g. EPR), some allow for multiple identities (TeO in an exemplary way, some advanced surveys such as ISSP), but many others are not. Most importantly, we lack a true panel that would observe individuals’ forms of identification over their life course and over generations (comparable to the Panel Study of Income Dynamics in the USA, for example). Maybe it is time to join forces and start such a panel.

The convenors also ask what the best methods are to measure ethnic and/or national identities. I am not in a position to answer this question from a more technical, survey construction or measurement theoretical point of view (see for example Worrell et al. 2019). In my limited view, it is astonishing how little survey designers have been informed by the theoretically more advanced discussions in the social sciences. It seems that psychologists and their understandings of identity (as “in-group” identification) are dominating the survey production industry. In other words, I don’t see a “best” method to measure ethnic and national identities so far, they all seem to be rather limited. To improve, they would have to take into



account a) the nested character of ethno-national systems of categorization (again, a Japanese American is an Asian American, is a Person of Color, is an American), b) the multiple dimensions along which these categories can vary in the eyes of individuals (including the cultural content or degrees of consensus, see Abdelal et al. 2006; Wimmer 2008), c) the overlap with or relationship to other forms of identification. Beyond surveys, as noted above, natural language processing techniques will further enhance our ability to detect and interpret identity discourses and their meaning in big corpi of texts. It is too early to tell how these techniques will evolve, but they are certainly superior to classical discourse analysis, which is marred by problems of interpretative bias and selectivity.

Finally, the organisers would like us to comment on the major research findings of quantitative approaches to the study of ethnicity and nationalism that were not possible using qualitative approaches. Here, I allow myself to take a rather parochial point of view and focus on one particular question that my own research has been concerned with: The study of ethno-political mobilization and conflict. The advent of new quantitative datasets such as EPR, AMAR, and others has led to a quantum leap in our understanding of these processes, specifically allowing us to understand the crucial role of ethno-political inequality—or exclusion from national level government—as one of the major drivers of such mobilization and conflict (see Wimmer et al. 2009; for other important factors, see Lindemann and Wimmer 2018). Most excitingly, these datasets are now linked to other datasets, including those on armed conflict, on the rebels participating in these conflicts, on the outside sponsors of these rebels, on the geographic areas in which they operate and the individual incidents of violence that occur, and so forth. The level of specificity and precision that these new sources of quantitative data allow is quite extraordinary—and represent a quantum leap compared to an older, foundational generation of scholarship that pioneered the study of ethnic politics (such as the magisterial Horowitz 1985). Much room for improvement remains, including in terms of measurement, modelling processual dynamics in statistically appropriate ways (going beyond the dominant comparative statics), theoretical precision, and addressing concerns of endogeneity in more state-of-the-art ways.

Beyond the field of ethno-racial politics and conflict, the study of identities would certainly profit from a rigorous multi-method research program, in which qualitative comparative scholarship, which has also tremendously improved recently with more careful attention to case selection and scope conditions, would be closely integrated with the different types of quantitative research outlined in this essay. The demands for mastering multiple methods are considerable, but the younger generation of researchers is also better equipped than ever to meet them.

## Bibliography

- Abdelal, R., Herrera, Y. M., Johnston, A. I. and McDermott, R. (2006). Identity as a variable. *Perspectives on Politics*, **4**(4), 695-711.
- Ahlerup, P., Baskaran, T. and Bigsten, A. (2017). Regional Development and National Identity in Sub-Saharan Africa. *Journal of Comparative Economics*, **45**(3), 622-643.
- Aktürk, Ş. (2012). *Regimes of ethnicity and nationhood in Germany, Russia, and Turkey*. Cambridge University Press.
- Alesina, A. and Ferrara, E. L. (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, **43**(3), 762-800.
- Alesina, A., Devleeschauwer, A., Easterly, W., Kurat, S. and Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, **8**(2), 155-194.
- Alise, M. A., and Teddlie, C. (2010). A Continuation of the Paradigm Wars? Prevalence Rates of Methodological Approaches Across the Social/Behavioral Sciences. *Journal of Mixed Methods Research*, **4**(2), 103-126.
- Anderson, B. (1991). *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. London: Verso.
- Aspinall, P. J. (2007). Approaches to Developing an Improved Cross-National Understanding of Concepts and Terms Relating to Ethnicity and Race. *International Sociology*, **22**(1), 41-70.
- Aspinall, P. J. (2012). Answer Formats in British Census and Survey Ethnicity Questions: Does Open Response Better Capture 'Superdiversity'? *Sociology*, **46**(2), 354-364.
- Balcells, L. (2017). *Rivalry and Revenge*. Cambridge: Cambridge University Press.
- Balch-Lindsay, D., and Enterline, A. J. (2000). Killing time: The world politics of civil war duration, 1820–1992). *International Studies Quarterly*, **44**(4), 615–642.
- Beauchemin, C., Hamel, C. and Simon, P. (2016). *Trajectoires et origines. Enquête sur la diversité des populations en France*. Paris: INED éditions.
- Benhabib, J. and Przeworski, A. (2006). The political economy of redistribution under democracy. *Economic Theory*, **29**(2), 271-290.
- Birnir, J. K. (2007). Divergence in Diversity? The Dissimilar Effects of Cleavages on Electoral Politics in New Democracies. *American Journal of Political Science*, **51**(3), 602-19.
- Birnir, J. K., Laitin, D., Wilkenfeld, J., Waguespack, D. M., Hultquist, A., and Gurr, T. R. (2018). Introducing the AMAR (All Minorities At Risk) Data. *Journal of Conflict Resolution*, **62**(1): 203-226.
- Birnir, J. K., Wilkenfeld, J., Fearon, J. D., Laitin, D. D., Gurr, T. R., Brancati, D., Saideman, S. M., Pate, A. and Hultquist, A. S. (2015). Socially relevant ethnic groups, ethnic structure, and AMAR. *Journal of Peace Research*, **52**(1), 110-115.
- Bochsler, D. (2013). Radicalizing Electoral System Effects on Support for Nationalist Hardliners. *Nations and Nationalism*, **19**(1), 21-43.
- Bochsler, D., and Schläpfer, B. (2016). An Indirect Approach to Map Ethnic Identities in Post-conflict Societies. *Ethnopolitics*, **15**(5), 467-486.

- Bonikowski, B. (2016). *Schemas of the Nation in Modern Democracies*. Cambridge: Department of Sociology, Harvard University.
- Bormann, N.-C., Cederman, L.-E., and Vogt, M. (2017). Language, Religion and Ethnic Civil War. *Journal of Conflict Resolution*, **61**(4), 744-771.
- Boutcher, S. A., Jenkins, J. C. and Dyke, N. V. (2017). Strain, Ethnic Competition, and Power Devaluation: White Supremacist Protest in the US, 1948-1997. *Social Movement Studies*, **16**(6), 686-703.
- Brand, L. A. (2014). *Official stories: Politics and national narratives in Egypt and Algeria*. Stanford University Press.
- Brown, G. K., and Langer, A. (2010). Conceptualizing and Measuring Ethnicity. *Oxford Developmental Studies*, **38**(4), 411-436.
- Brubaker, R., Feischmidt, M., Fox, J. and Grancea, L. (2018). *Nationalist politics and everyday ethnicity in a Transylvanian town*. Princeton, NJ: Princeton University Press.
- Bruk, Solomon Il'ič, and V. S. Apenčenko. (1964). *Atlas narodov mira*. Glavnoe upravlenie geodezii i kartografii Gosudarstvennogo geologičeskogo komiteta SSSR. Institut étnografii im. NN Miklucho-Maklaja Akademii nauk SSSR.
- Burton, J., Nandi, A. and Platt, L. (2010). Measuring ethnicity: challenges and opportunities for survey research. *Ethnic and Racial Studies*, **33**(8), 1332-1349.
- Cassan, G. (2015). Identity Based Policies and Identity Manipulation: Evidence from Colonial Punjab. *American Economic Journal: Economic Policy*, **7**(4), 103-31.
- Cederman, L.-E. (1997). *Emergent Actors in World Politics: How State and Nations Develop and Dissolve*. Princeton, NJ: Princeton University Press.
- Cederman, L.-E., Wimmer, A. and Min, B. (2010). Why do ethnic groups rebel? New data and analysis. *World Politics*, **62**(1), 87-119.
- Cederman, L.-E., Gleditsch, K. S. and Buhaug, H. (2013). *Inequality, grievances, and civil war*. New York, NY: Cambridge University Press.
- Chandra, K. (2006). What is Ethnic Identity and Does it Matter? *Annual Review of Political Science* **9**, 397-424.
- Chandra, K. (2009). A Constructivist Dataset on Ethnicity and Institutions. In R. Abdelal, Y. M. Herrera, A. I. Johnston, and R. McDermott (eds.), *Measuring Identity: A Guide for Social Scientists* (pp. 250–276). Cambridge: Cambridge University Press.
- Chandra, K. (ed.) (2012). *Constructivist Theories of Ethnic Politics*. Oxford: Oxford University Press.
- Chandra, K. and Laitin, D. (2002). A Framework for Thinking About Identity Change. Presentation at LICEP 5: 2-16.
- Chandra, K. and Wilkinson, S. (2008). Measuring the Effect of "Ethnicity." *Comparative Political Studies*, **41**(4-5), 515–563.
- Charnysh, V. and Finkel, E. (2017). The Death Camp Eldorado: Political and Economic Effects of Mass Violence. *American Political Science Review*, **111**(4), 801-818.
- Christia, F. (2012). *Alliance formation in civil wars*. Cambridge: Cambridge University Press.

- Collier, P., and Hoeffler, A. (2004). Greed and Grievance in Civil War. *Oxford Economic Papers*, **56**(4), 563-595.
- Costalli, S., and Ruggeri, A. (2015). Indignation, Ideologies, and Armed Mobilization: Civil War in Italy, 1943-45. *International Security*, **40**(2), 119-157.
- Cox, G. W. (1997). *Making votes count: strategic coordination in the world's electoral systems*. Cambridge: Cambridge University Press.
- Cunningham, D. E., Gleditsch, K. S. and Salehyan, I. (2009). It takes two: A dyadic analysis of civil war duration and outcome. *Journal of Conflict Resolution*, **53**(4), 570–597.
- Cunningham, K. G. (2014). *Inside the Politics of Self-Determination*. Oxford: Oxford University Press.
- Cunningham, K. G. and Weidmann, N. B. (2010). Shared space: Ethnic groups, state accommodation, and localized conflict. *International Studies Quarterly*, **54**(4), 1035–1054.
- De Winter, L. and Dumont, P. (1999). Belgium: Party System(s) on the Eve of Disintegration? in D. Broughton, M. Donovan (eds.), *Changing Party Systems in Western Europe* (pp. 183-206). London: Pinter.
- Depetris-Chauvin, E., Durante, R., and Campante, F. (2020). Building Nations Through Shared Experiences: Evidence from African Football. *American Economic Review*, **110**(5), 1572-1602.
- Easterly, W. and Levine, R. (1997). Africa's Growth Tragedy: Policies and Ethnic Divisions. *Quarterly Journal of Economics*, **112**(4), 1203-50.
- Eschbach, K., Supple, K. and Snipp, C. M. (1998). Changes in Racial Identification and the Educational Attainment of American Indians, 1970-1990. *Demography*, **35**(1), 35-43.
- Esteban, J.-M. and Ray, D. (1994). On the measurement of polarization. *Econometrica*, **62**(4), 819-851.
- Fearon, J. D. (2003). Ethnic and Cultural Diversity by Country. *Journal of Economic Growth*, **8** (2), 195-222.
- Fearon, J. D. (2004). Why Do Some Civil Wars Last So Much Longer Than Others? *Journal of Peace Research*, **41**(3), 275-301.
- Fearon, J. D. and Laitin, D. D. (1996). Explaining Interethnic Cooperation. *American Political Science Review*, **90**(4), 715-735.
- Fearon, J. D. and Laitin, D. D. (2003). Ethnicity, insurgency, and civil war. *American Political Science Review*, **97**(1), 75-90.
- Flores, R. (2017). Do anti-Immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology*, **123**(2), 333-384.
- Geddes, B. (1990). How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics. *Political Analysis*, **2**, 131-150.
- Gellner, E. (2006 [1983]). *Nations and Nationalism*. Oxford: Blackwell.
- Glazer, A., Grofman, B. and Owen, G. (1998). A neo-Downsian model of group-oriented voting and racial backlash. *Public Choice*, **97**, 23-34.

- Goddard, S. E. (2006). Uncommon ground: Indivisible territory and the politics of legitimacy. *International Organization*, **60**(1), 35–68.
- Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J. and Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science*, **54**(1), 190-208.
- Green, E. D. (2020). Ethnicity, Nationalism and the State: Evidence from Sub-Saharan Africa. *British Journal of Political Science*, **50**(2), 757-779.
- Green, E. D. (2021). The Politics of Ethnic Identity in Sub-Saharan Africa. *Comparative Political Studies*, **forthcoming**.
- Green, D. P, Kim, S. Y. and Yoon, D. H. (2001). Dirty Pool. *International Organization*, **55**(2), 441-468.
- Greif, A. (1997). Cliometrics After 40 Years. *American Economic Review*, **87**(2), 400-403.
- Guinjoan, M., and Rodon, T. (2016). A Scrutiny of the Linz-Moreno Question. *Publius: The Journal of Federalism*, **46**(1), 128-42.
- Gurr, T. R. Minorities at Risk Project 2005 [cited 6 June 2008]. Available from <http://www.cidcm.umd.edu/inscr/mar/>.
- Hierro, M. J. and Gallego, A. (2018). Identities in between: Political Conflict and Ethnonational Identities in Multicultural States. *Journal of Conflict Resolution*, **62**(6), 1314-1339.
- Horowitz, Donald L. (1985). *Ethnic groups in conflict*. Berkeley, CA: University of California Press.
- Howe, P. J., Szöcsik, E. and Zuber, C. (2019). The Promise of Status Improvement. How Nationalists Attract Working-Class Votes. In IPSA colloquium “Diversity and Democratic Governance”. Sarajevo.
- Jenne, E. K. (2007). *Ethnic Bargaining. The Paradox of Minority Empowerment*. Ithaca (NY): Cornell University Press.
- Jenne, E. K., Saideman, S. S. and Lowe, W. (2007). Separatism as a Bargaining Posture: The Role of Leverage in Minority Radicalization. *Journal of Peace Research*, **44**(5), 539-558.
- Just, A. and Anderson, C. J. (2012). Immigrants, Citizenship and Political Action in Europe. *British Journal of Political Science*, **42**(3), 481-509.
- Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.
- Kalyvas, S. N. (2008). Ethnic defection in civil wars. *Comparative Political Studies*, **41**(8), 1043–1068.
- Kertzer, D. I. and Arel, D. (eds). (2001). *Census and Identity. The Politics of Race, Ethnicity and Language in National Censuses*. Cambridge: Cambridge University Press.
- Key, V. O. (1949). *Southern Politics in State and Nation*. New York: Vintage.
- Koopmans, R. and Olzak, S. (2004). Discursive Opportunities and the Evolution of Right-Wing Violence in Germany. *American Journal of Sociology*, **110**(1), 198-230.
- Kopstein, J. S. and Wittenberg, J. (2011). Deadly Communities: Local Political Milieus and the Persecution of Jews in Occupied Poland. *Comparative Political Studies*, **44**(3), 259-283.

- Kozlowski, A. C., Taddy, M. and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, **84**(5), 905-949.
- Kuo, K. and Mylonas, H. (2020). Nation-Building and the Role of Identity in Civil Wars. *Ethnopolitics*, forthcoming.
- Laitin, D. D. (2007). *Nations, States and Violence*. Oxford: Oxford University Press.
- Laitin, D. D., and Posner, D. N. (2001). 'The Implications of Constructivism for Constructing Ethnic Fractionalization Indices. APSA-CP 12.
- Lambert, P. S. (2005). Ethnicity and the Comparative Analysis of Contemporary Survey Data. In J. H. P. Hoffmeyer-Zlotnik and J. Harkness (eds.), *Methodological aspects in cross-national research* (pp. 259-278). Mannheim: Gesis-Zuma.
- Lee, T. (2009). Between Social Theory and Social Science Practice. Toward a New Approach to the Survey Measurement of 'Race'. In R. Abdelal, Y. M. Herrera, A. I. Johnston and R. McDermott (eds.), *Measuring Identity* (pp. 113-144). Cambridge: Cambridge University Press.
- Levy, Naomi. (2014). Measuring multiple identities: what is lost with a zero-sum approach. *Politics, Groups, and Identities*, **2**(4), 624-642.
- Lieberman, E. S. and Singh, P. (2012). Conceptualizing and Measuring Ethnic Politics: An Institutional Complement to Demographic, Behavioral, and Cognitive Approaches. *Studies in Comparative International Development*, **47**, 255-86.
- Lindemann, S. and Wimmer, A. (2018). Repression and refuge: Why only some politically excluded ethnic groups rebel. *Journal of Peace Research*, **55**(3), 305-319.
- Linz, J. J., Gómez-Reino y Carnota, M., Orizo, F. A. and Carro, D. V. (1981). *Informe Sociológico sobre el cambio político en España 1975-1981: IV*. Madrid: Euroamérica.
- Manatschal, A. (2012). Path-dependent or dynamic? Cantonal integration policies between regional citizenship traditions and right populist party politics. *Ethnic and Racial Studies*, **35**(2), 281-297.
- McCrone, D, and Surrige, P. (1998). National identity and national pride. In R. Jowell (ed.), *British and European – Social Attitudes – The 15th Report: How Britain Differs*. Aldershot: Ashgate.
- Moerman, M. (1965). Ethnic Identification in a Complex Civilization: Who Are the Lue? *American Anthropologist*, **67**(5), 1215-1230.
- Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic Polarization, Potential Conflict, and Civil Wars. *American Economic Review*, **95**(3), 796-816.
- Morning, A. (2008). Ethnic Classification in Global Perspective: A Cross-National Survey of the 2000 Census Round. *Population Research and Policy Review*, **27**(2), 239-272.
- Mylonas, H. (2012). *The Politics of Nation-Building: Making Co-Nationals, Refugees, and Minorities*. New York: Cambridge University Press.
- Mylonas, H. (2015). Methodological Problems in the Study of Nation-Building: Behaviorism and Historicist Solutions in Political Science. *Social Science Quarterly*, **96** (3), 740–758.

- Mylonas, H. and Tudor, M. (Forthcoming). Nationalism-What we know and what we still need to Know. *American Review of Political Science*.
- Nix, E., and Qian, N. (2015). The Fluidity of Race: Racial Passing in the United States, 1880-1940. NBER Working Paper #20828.
- Phinney, J. S. (1992). The Multigroup Ethnic Identity Measure: A New Scale for Use with Diverse Groups. *Journal of Adolescent Research*, **7**(2), 156-176.
- Platt, J. (2012). Making them count: How effective has official encouragement of quantitative methods been in British sociology? *Current Sociology*, **60**(5), 690-704.
- Platteau, J.-P. (2009). Institutional Obstacles to African Economic Development: State, Ethnicity and Custom. *Journal of Economic Behavior and Organization*, **71**(3), 669-689.
- Posner, D. N. (2004). Measuring Ethnic Fractionalization in Africa. *American Journal of Political Science*, **48**(4), 849-863.
- Posner, D. N. (2005). *Institutions and Ethnic Politics in Africa*. Cambridge University Press.
- Reynal-Querol, M. (2002). Ethnicity, Political Systems, and Civil Wars. *Journal of Conflict Resolution*, **46**(1), 29-54.
- Roberts, R. E., Phinney, J. S., Masse, L. C., Chen, Y. R., Roberts, C. R. and Romero, A. (1999). The Structure of Ethnic Identity of Young Adolescents from Diverse Ethnocultural Groups. *Journal of Early Adolescence*, **19**(3), 301-322.
- Robinson, A. L. (2014). National versus Ethnic Identity in Africa: Modernization, Colonial Legacy and the Origins of Territorial Nationalism. *World Politics*, **66**(4), 709-746.
- Roeder, Philip G. (2007). *Where Nation-States Come From. Institutional Change in the Age of Nationalism*. Princeton: Princeton University Press.
- Rosenberg, S. and Beattie, P. (2019). The Cognitive Structuring of National Identity: Individual Differences in Identifying as American. *Nations and Nationalism*, **25**(1), 361-384.
- Saideman, S. M. and Ayres, R. W. (2000). Determining the Causes of Irredentism: Logit Analyses of Minorities at Risk Data from the 1980s and 1990s. *Journal of Politics*, **62**(4), 1126-1144.
- Sambanis, N. 2000. Partition as a solution to ethnic war: An empirical critique of the theoretical literature. *World Politics*, **52**(4), 437-483.
- Sambanis, N. (2001). Do ethnic and non-ethnic civil wars have the same causes? *Journal of Conflict Resolution*, **45**(3), 259-282.
- Selway, J. S. (2011a). The Measurement of Cross-cutting Cleavages and Other Multidimensional Cleavage Structures. *Political Analysis*, **19**(1), 48-65.
- Selway, J. S. (2011b). Cross-Cuttingness, Cleavage Structures and Civil War Onset. *British Journal of Political Science*, **41**(1), 111-138.
- Shayo, M. (2009). A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution. *American Political Science Review*, **103**(2), 147-174.
- Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., and Sherif, C. W. (1961). *The Robbers Cave Experiment: Intergroup Conflict and Cooperation*. Norman, OK: University of Oklahoma.

- Sigelman, L. (2006). The Coevolution of American Political Science and the "American Political Science Review." *American Political Science Review*, **100**(4), 463-478.
- Simon, P. (2012). Collecting ethnic statistics in Europe: a review. *Ethnic and Racial Studies*, **35**(8), 1366-1391.
- Singh, P. (2015). *How Solidarity Works for Welfare: Subnationalism and Social Development in India*. Cambridge: Cambridge University Press.
- Sinnott, R. (2006). An Evaluation of the Measurement of National, Subnational and Supranational Identity in Crossnational Surveys. *International Journal of Public Opinion Research*, **18**(2), 211-223.
- Siroky, D. S. and Hale, C. W. (2017). Inside Irredentism: A Global Empirical Analysis. *American Journal of Political Science*, **61**(1), 117-128.
- Smith, A. D. (1991). *National Identity*. London: Penguin.
- Sniderman, P. M., Hagendorn, L. and Prior, M. (2004). Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities. *American Political Science Review*, **98**(1), 35-49.
- Staniland, P. (2014). *Networks of rebellion: Explaining insurgent cohesion and collapse*. Ithaca, NY: Cornell University Press.
- Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: evidence using Google search data. *Journal of Public Economics*, **118**, 26-40.
- Stelzl, M. and Seligman, C. (2009). Multiplicity Across Cultures: Multiple National Identities and Multiple Value Systems. *Organization Studies*, **30**(9), 959-973.
- Theiss-Morse, E. (2009). *Who Counts as an American? The Boundaries of National Identity*. Cambridge: Cambridge University Press.
- Toft, M. D. (2003). *The geography of ethnic violence: Identity, interests, and the indivisibility of territory*. Princeton, NJ: Princeton University Press.
- Turner, J. C., Brown, R. J. and Tajfel, H. (1979). Social Comparison and Group Interest in Ingroup Favouritism. *European Journal of Social Psychology*, **9**(2), 187-204.
- United Nations Statistics Division. (2008). Principles and Recommendations for Population and Housing Censuses. New York: United Nations.
- Uppsala Conflict Data Project (UCDP), from <https://ucdp.uu.se/>
- Vogt, M., Bormann, N.-C., Rügger, S., Cederman, L.-E., Hunziker, P. and Girardin, L. (2015). Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Dataset Family. *Journal of Conflict Resolution*, **59**(7), 1327-1342.
- Wedeen, L. (2009). *Peripheral visions: Publics, power, and performance in Yemen*. Chicago, IL: University of Chicago Press.
- Wilkinson, S. (2004). *Votes and Violence. Electoral Competition and Ethnic Riots in India*. Cambridge: Cambridge University Press.
- Wimmer, A. (2008). The Making and Unmaking of Ethnic Boundaries: A Multilevel Process Theory. *American Journal of Sociology*, **113**(4), 970-1022.



- Wimmer, A. (2012). *Waves of war: Nationalism, state formation and ethnic exclusion in the modern world*. New York, NY: Cambridge University Press.
- Wimmer, A. (2017). Power and pride. National identity and ethnopolitical inequality around the world. *World Politics*, **69**(4), 605-639.
- Wimmer, A. (2018). *Nation Building. Why Some Countries Come Together While Others Fall Apart*. Princeton, NJ: Princeton University Press.
- Wimmer, A, Cederman, L.-E. and Min, B. (2009). Ethnic politics and armed conflict. A configurational analysis of a new global dataset. *American Sociological Review*, **74**(2), 316-337.
- Wimmer, A. and Lewis, K. (2010). Beyond and below racial homophily. ERG models of a friendship network documented on Facebook. *American Journal of Sociology*, **116**(1), 583-642.
- Worrell, F. C., Mendoza-Denton, R. and Wang, A. (2019). Introducing a new Assessment tool for measuring ethnic-racial identity: The Cross Ethnic-Racial Identity Scale–Adult (CERIS-A). *Assessment*, **26**(3), 404-418.
- Yap, S. C. Y., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., Weisskirch, R. S., Lee, R. M., Park, I. J. K., and Whitbourne, S. K. (2014). Investigating the Structure and Measurement Invariance of the Multigroup Ethnic Identity Measure in a Multiethnic Sample of College Students. *Journal of Counseling Psychology*, **61**(3), 437-446.
- Zellman, A. (2019). ‘Hawking’ Territorial Conflict: Ethnopolitism and Nationalist Framing Strategies. *East European Politics*, **35**(4), 474-495.