


ARTICLE

Measuring Ethnic Inequality: An Assessment of Extant Cross-National Indices

Lasse Egendal Leipziger 

Department of Political Science, Aarhus University, Aarhus, Denmark
Email: lel@ps.au.dk

(Received 15 April 2021; revised 13 December 2021; accepted 1 April 2022; first published online 1 June 2022)

Abstract

This article offers an evaluation of cross-national measures of ethnic socio-economic inequality. It demonstrates that the measures differ in important ways regarding empirical scope, conceptualization, measurement and aggregation. Despite significant advances in the measurement of ethnic inequality, all measures have shortcomings, such as limited and biased coverage, as well as measurement error from the underlying data sources. Moreover, the empirical convergence between conceptually similar measures is strikingly low: some of the measures show no or even negative covariation. Four replication studies also indicate that extant measures of ethnic inequality are generally not interchangeable. Scholars should therefore take the various features highlighted in this evaluation into account before employing any of them. Based on this conclusion, the article offers multiple suggestions for improving existing measures and developing new ones.

Keywords: ethnic inequality; horizontal inequality; conceptualization; measurement validity; dataset assessment

Economic and social inequalities between ethnic groups – also known as horizontal inequalities (Stewart 2008) – have received increased attention in academia and policy circles in recent years. This growing interest is clear from the dramatic increase in the number of related academic publications.¹ A considerable body of political science research suggests that within-country inequalities between ethnic groups have major negative implications for peace, economic and political development, public goods provision, and individual well-being (see, for example, Alesina, Michalopoulos and Papaioannou 2016; Baldwin and Huber 2010; Canelas and Gisselquist 2019; Cederman, Weidmann and Bormann 2015; Houle 2015; Houle and Bodea 2017; Stewart 2008; Wang and Kolev 2019; Ye and Han 2019). Furthermore, the reduction of group-level inequalities is included in United Nations Sustainable Development Goal 10 (UN 2020), and the issue was emphasized in a recent Organisation for Economic Co-operation and Development (OECD) report (Deere, Kanbur and Stewart 2018).

Much of the comparative research that has flourished in the past decade is premised on a series of relatively new datasets. These are valuable tools that can help monitor variation across space and time, as well as analyse causes and consequences. A few methodological studies have addressed measurement challenges related to survey and census data (Canelas and Gisselquist 2019), suggested good measurement practices (Stewart, Brown and Mancini 2010), and discussed data sources (Baghat et al. 2017; Tetteh-Baah 2019). However, problems of causal inference have largely overshadowed important problems of conceptualization and measurement, and there are currently no systematic comparative evaluations of how extant cross-national measures relate to each other

¹A keyword search on dimensions.ai for ‘ethnic inequality’ in social sciences subjects returns eighty-seven publications in 2000 against 419 in 2020. The corresponding numbers for ‘horizontal inequality’ are nine in 2000 and 263 in 2020.

conceptually and empirically. This also means that we have limited knowledge of the strengths and weaknesses of the various measures, including whether they can be considered interchangeable.

Against this background, this article contributes to the emerging literature on ethnic inequalities by discussing and comparing six different cross-national measures offered by Alesina, Michalopoulos and Papaioannou (2016), Cederman, Gleditsch and Buhaug (2013), Houle (2015), Baldwin and Huber (2010), Omoeva, Moussa and Hatch (2018) and Varieties of Democracy (V-Dem) (Coppedge et al. 2021a).² Scholars have used the evaluated measures in empirical studies to operationalize social or economic ethnic inequality cross-nationally, and they cover the majority of contemporary countries – or at least include countries from several world regions. Even though not all of these measures were created for broad purposes, they are increasingly being used for different empirical research (see, for example, Fleming et al. 2020; Ye and Han 2019), which underlines the need for systematic comparison.

The examination of the six indices is inspired by the steps in the integrated assessment framework suggested by Munck and Verkuilen (2002), which provides a comprehensive checklist to evaluate data. However, my examination also goes beyond their framework by providing a series of new data visualizations, as well as four replication studies. In the assessment, I find clear differences in conceptualization, measurement, aggregation and empirical scope. Dramatic differences in coverage influence their relevance for research questions about the causes or consequences of ethnic inequality, which rely on cross-national and, especially, cross-temporal variation. The majority of measures are also afflicted by important biases, such as mainly covering developing countries or focusing exclusively on democracies. Moreover, a comparison of the data sources – including mass surveys, expert surveys, administrative data and satellite data on night lights – reveals likely sources of measurement error. A number of correlation analyses show that the empirical convergence between the measures is surprisingly low, even when taking into account the differences in conceptualization and aggregation procedures. Notably, two measures based on similar definitions exhibit no significant correlation at all. Moreover, the replication studies suggest that the results of a number of prominent studies are sensitive to measurement choice. The article thus aims to raise awareness about extant measures of ethnic inequality so that their respective strengths and weaknesses can be taken into account in the assessment of previous studies and the design of new ones. Based on these findings, I discuss potential avenues forward, including more disaggregated analyses and combining various data sources.

Conceptualization

At the most general level, inequality is about ‘the ability of households to maintain economically a certain standard of living and lifestyle’ (Jensen and van Kersbergen 2016, 36). If individuals or families have very different options in terms of how to live their lives, we intuitively consider them as living in an unequal society. Conceptually, we may distinguish between inequality on the individual and group levels. Interpersonal (or ‘vertical’) inequality is about differences between individuals or households, typically referring to disparities in post-tax-transfer disposable household income in a given year (Jensen and van Kersbergen 2016, 36). The empirics are typically summarized into comparable measures using Gini coefficients, ratios between income percentiles or income shares going to the top percentiles (Jensen and van Kersbergen 2016, 36–47; Piketty 2014).

Intergroup (or ‘horizontal’) inequality concerns between-group differences, which are defined according to the type of group identification one is interested in studying, such as ethnicity (Stewart 2002, 13). Ethnic inequalities can be measured both at the aggregate, country level (providing a single figure that represents the entire distribution in a country) and at the group level

²Houle as well as Omoeva, Moussa and Hatch generously shared their data. The remaining datasets were downloaded from online databases.

(providing figures for each group relative to the country mean or another group). This article focuses exclusively on aggregate, cross-national measures, which have been employed by most comparative studies so far (Baghat et al. 2017, 67). They use the average differences in outcomes, such as income or education, between ethnic groups in a society, aggregating them for comparisons across countries and over time.³ In the surveyed works, ethnicity is generally understood in an encompassing manner consistent with the recent literature on ethnic politics (Canelas and Gisselquist 2018, 306; Chandra 2006, 398; Horowitz 2000). Following the tradition of Max Weber, ethnicity may be defined as a subjectively experienced sense of commonality based on a belief in common ancestry and shared culture (Weber 1976 [1922], 389). Ethnic identity markers indicating a shared ancestry and culture include language (for example, in Belgium), religion (for example, in Bosnia and Herzegovina), tribe (for example, in Kenya), caste (for example, in India), phenotypical features (for example, in the United States) or some combination thereof. In other words, ethnic categories are social constructs linked to descent-based attributes.

Are the surveyed measures of ethnic inequality based on similar conceptual foundations? The examined datasets variously refer to ‘economic horizontal inequality’ (Cederman, Gleditsch and Buhaug 2013, 93), ‘differences in the economic well-being of groups’ (Baldwin and Huber 2010, 645), ‘between-ethnic-group inequality (BGI)’ (Houle 2015, 470), ‘within country differences in well-being across ethnic groups’ (Alesina, Michalopoulos and Papaioannou 2016, 429), ‘inequalities in education ... between ethnic groups’ (Omoeva, Moussa and Hatch 2018, 3) and ‘inequalities in access to public services ... between particular social groups’ (Coppedge et al. 2021a, 218). To invoke a useful distinction by Adcock and Collier (2001), they are not ‘systematized concepts’, but seem to agree on the ‘background concept’. That is to say, despite different terminologies, all of the surveyed measures share a common conceptual core, as they all reflect asymmetries in socio-economic conditions between ethnic groups. Importantly, all datasets are explicit about which dimension of ethnic inequality they are capturing (see Stewart 2002): the economic dimension concerns the distribution of income and wealth between ethnic groups (as used by Alesina, Michalopoulos and Papaioannou, by Baldwin and Huber, by Cederman, Gleditsch and Buhaug, and by Houle); while the social dimension concerns the uneven access of groups to public services, such as healthcare and education (as used by Coppedge et al. and by Omoeva, Moussa and Hatch). These dimensions not only reflect a common core – socio-economic ethnic inequality – but are also likely to be highly correlated due to common determinants and reciprocal relationships: inequality in access to public services may translate into, and be highly associated with, economic ethnic inequality and vice versa (see Stewart, Brown and Mancini 2010). The conceptual structure is illustrated in Figure 1.

Measurement

The various dimensions and sub-dimensions of socio-economic ethnic inequality can be operationalized in various ways, using a range of indicators. Since the data providers implicitly agree on a background concept (that is, ethnic inequality concerns differences in standards of living between ethnic groups), a comparison of these measures seems meaningful.

Overview of Extant Measures

Alesina, Michalopoulos and Papaioannou’s (2016) ethnic Gini indices are based on satellite images of night-time luminosity, combined with the homelands of ethnolinguistic groups. This measure reflects differences in ‘mean income’ – as reflected by luminosity per capita across ethnic

³Economic and social ethnic inequality is closely related to and may be subsumed under the broader concept of *horizontal inequalities*, which deserves brief clarification due to its prominence in the literature. Horizontal inequalities refer to ‘inequalities in economic, social, or political dimensions or cultural status between culturally defined groups’ (Stewart 2008, 3). Stewart coined this term to distinguish it from interpersonal inequality – also referred to as ‘vertical inequality’. Horizontal inequalities are multidimensional (including political and cultural inequalities) and potentially refer to any relevant group inequality, such as gender (Deere et al. 2018).

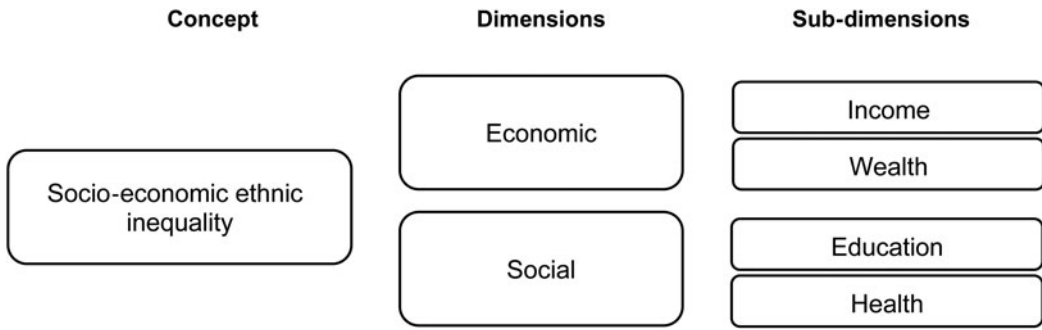


Fig. 1. Illustration of conceptual structure.

homelands – between groups within 173 countries (in 1992, 2000 and 2012). Ethnic groups are located using two datasets/maps: first, the Geo-Referencing of Ethnic Groups (GREG), which is the digitized version of the Soviet Atlas Narodov Mira from the 1960s (Weidmann, Rød and Cederman 2010); and, second, the fifteenth edition of the *Ethnologue* (Gordon 2005), which mapped 7,581 language-country groups worldwide in the mid- to late 1990s. The GREG attempts to map major immigrant groups, whereas *Ethnologue* generally does not. Hence, the two ethnolinguistic mappings capture different ethnic groups, which is particularly important for countries in the Americas (Alesina, Michalopoulos and Papaioannou 2016, 433).

Cederman, Gleditsch and Buhaug (2013) geographically match subnational economic data (the G-Econ data by Nordhaus et al. [2006]) with data on the geographical boundaries of ethnic settlements from the geocoded extension (GeoEPR) of the Ethnic Power Relations (EPR) dataset (Wucherpfennig et al. 2011). While their analytical focus is on investigating group-level data and civil war onset, they also conduct cross-national analyses (see also Buhaug, Cederman and Gleditsch 2014). Strictly speaking, the temporal scope is limited to a single year because the G-Econ data only reflect 1990 values and only the GeoEPR is dynamic, taking into account major changes in ethnic settlement patterns over time (Cederman, Gleditsch and Buhaug 2013, 101, 106).

Houle (2015) uses information from a range of surveys, including the Demographic and Health Survey (DHS), World Values Survey (WVS) and various regional barometers, to construct an asset-based wealth indicator for within-group, between-group and cross-national ethnic inequality. Since the data were originally gathered to study democratic breakdowns, the measure covers 89 countries from 1960 to 2007 that have been democratic for at least one year and are ethnically heterogeneous. The panel is unbalanced and exhibits limited variation over time (Houle 2015, 500).

Baldwin and Huber (2010) construct a between-group inequality measure (BGI), similar to a group Gini coefficient, for 46 democracies based on income variables from a series of surveys. The sample includes democracies from all regions of the world, though Asia and especially Latin America are under-represented in so far as these regions have a higher proportion of democracies than the dataset suggests (Baldwin and Huber 2010, 648). Each country is measured in one year between 1996 and 2006, effectively making the data cross-sectional. The data only include democracies because they were originally collected for the purpose of studying public goods provision in heterogeneous democracies. As pioneers in the field, Baldwin and Huber are careful to validate their measure empirically, including comparison of their measure to a handful of countries, where the nature of inequality between groups is widely acknowledged. Moreover, they turn to a number of more fine-grained household surveys that identify income by ethnic group (Baldwin and Huber 2010, 649–50).⁴

⁴Wang and Kolev's (2019) dataset explicitly builds on Baldwin and Huber's and has thus not been included in the main discussion.

Finally, two measures capture unequal access to public services rather than economic outcomes. In the newest data release (v11.1), V-Dem provides an expert-coded indicator of inequality in access to basic public services (for example, primary education, clean water and healthcare) distributed by ‘social group’. The group definition corresponds to a broad conception of ethnicity, covering, among other things, language, race and religion (Coppedge et al. 2021a, 209). The dataset covers all sovereign states in the world since 1900, with the exception of a number of micro-states.

As part of the Education Inequality and Conflict Project (EIC 2015), Omoeva, Moussa and Hatch (2018, 16) have created measures of inequality in educational attainment between ethnic/religious groups by constructing a group Gini coefficient (as well as Theil Index, coefficient of variation and parity ratio). They draw on educational attainment data from three public household survey datasets (Omoeva, Moussa and Hatch 2018, 15) and fill in missing country-year observations using a logical backward projection technique. The unbalanced dataset covers a set of 86 predominantly developing countries between 1946 and 2013 (Omoeva, Moussa and Hatch 2018, 50).

Despite the measures all sharing a common focus on ethnic inequality, they are marked by dramatic differences in scope, ranging from a cross-section of 46 countries to a measure covering most polities since 1900 (see Table 1 and Figure 2).⁵ There are also large differences in terms of how time varying the data are. This is illustrated in Figure 3, which plots the values of the different measures over time for Bolivia, where a high level of ethnic inequality is widely acknowledged (Houle 2015, 485). The V-Dem, the Omoeva, Moussa and Hatch, and the Alesina, Michalopoulos and Papaioannou measures exhibit significant variation over time. In contrast, save for a minor change in the Houle measure, the Cederman, Gleditsch and Buhaug and the Houle measures are time invariant.⁶

The creators of the first cross-national measures, such as Baldwin and Huber (2010), deserve much credit for paving the way with their work to conceptualize and create the first ethnic inequality measures. For the purpose of future empirical studies, however, the restricted empirical scope of most measures limits their value for particular research questions. In particular, the ability to track the developments over time with respect to ethnic inequality is severely restricted.

Finally, there are strong non-random patterns in the data. Most clearly, not all measures support direct comparisons between poor countries and the experiences of rich, long-enduring democracies. In particular, the Omoeva, Moussa and Hatch (2018) data only include a limited number of high-income countries, whereas the Baldwin and Huber (2010) and Houle (2015) datasets only include democracies. I further explore this issue in the Online Appendix (see Table A1) with a simple test of non-random missingness (see Rios-Figueroa and Staton 2012, 125). These findings show that most measures provide samples that are not representative regarding gross domestic product per capita (GDP/cap), democracy and state capacity. Looking across the tests, the V-Dem and the Alesina, Michalopoulos and Papaioannou measures appear to be least afflicted by non-random missingness. These non-random patterns in the data reduce the ability to infer from the sample to the general population of all countries, and they mean we should avoid being overly confident about any robustness analysis using alternative measures.

Data Sources

In addition to well-known measurement constraints for interpersonal (or vertical) inequality, the measurement of ethnic inequality depends on comparable group classifications. This represents a significant challenge, as ethnic identities are not static, people hold multiple identities and data

⁵Figures A1–A6 in the Online Appendix present maps showing the countries covered by each dataset.

⁶Table A1 in the Online Appendix also reveals important differences in the temporal granularity of the different measures.

Table 1. Scope, sources and operationalization of extant measures

Data provider and index	Countries	Years	Country-year observations	Sources	Operationalization
Alesina, Michalopoulos and Papaioannou (2016): ethnic Gini	173	1992, 2000, 2012	519	Night lights/ethnic homelands	Group Gini (0–1)
Cederman, Gleditsch and Buhaug (2013): G-Econ/ethnic homeland	163	1990	163	Local economic data/ethnic homelands	Ratio (poorest/richest group relative to country mean)
Houle (2015): BGI	75	1960–2007 (unbalanced)	1,641	Mass survey	BGI indicator (0–6)
Baldwin and Huber (2010): BGI	46	1996–2006 (unbalanced)	46	Mass survey	BGI indicator (0–1; standardized scores available)
V-Dem (Coppedge et al. 2021a; Coppedge et al. 2021b): access to public services by group	179	1900–2020	18,157	Expert survey	Point estimate and confidence bounds based on a Bayesian item response theory model (original scale: 0–4)
Omoeva, Moussa and Hatch (2018): educational group Gini	86	1946–2013 (unbalanced)	4,254	Mass survey	Group Gini (+ Theil, coefficient of variation and Parity Index)

Notes: It should be noted that night lights may also proxy for access to public services, meaning the distinction between economic and social dimensions is not clear-cut.

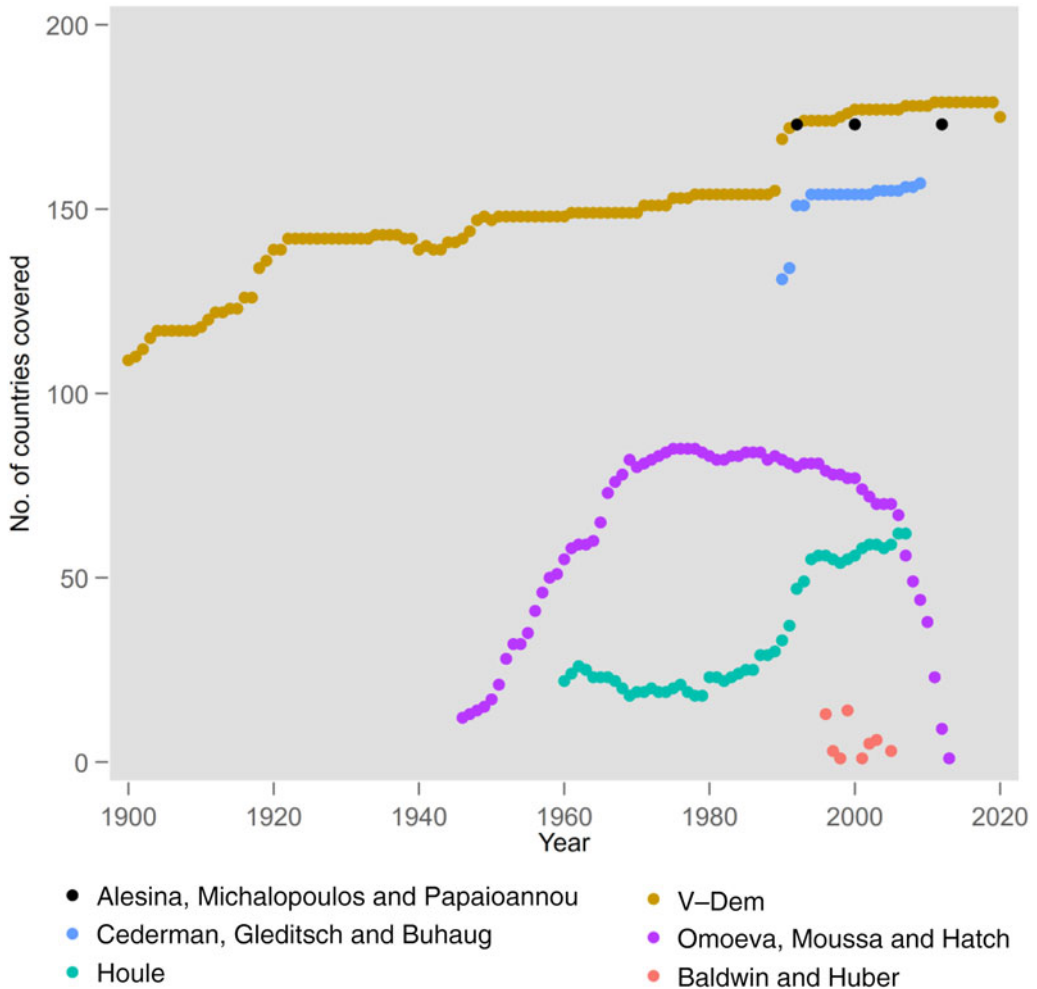


Fig. 2. Country coverage over time, by measure.

Notes: Each point indicates how many countries are covered in a given year by the dataset in question. V-Dem covers many colonies, which explains its high country-coverage prior to decolonization.

are often unavailable or incomplete (Bochsler et al. 2021; Canelas and Gisselquist 2019, 161; Stewart, Brown and Mancini 2010, 10). Dataset creators have creatively addressed these challenges and collected data in three general ways: (1) surveys, which include information on both socio-economic well-being and ethnic group affiliations; (2) spatial datasets, which geographically match economic data with data on the geographical boundaries of ethnic settlements; and (3) expert coding.

More specifically, the challenge of identifying comparable ethnic categories has been addressed in three main ways. One strand, which includes Baldwin and Huber (2010), Alesina, Michalopoulos and Papaioannou (2016), Cederman, Gleditsch and Buhaug (2013) and Houle (2015), adheres to the ethnic group classification as coded by either Fearon (2003) or *Ethnologue* (Gordon 2005), or by the EPR dataset or its geocoded extension (GeoEPR) (Vogt et al. 2015; Weidmann, Rød and Cederman 2010). Another strand, represented by Omoeva, Moussa and Hatch (2018), uses the ethnic categories that have been predefined by the teams

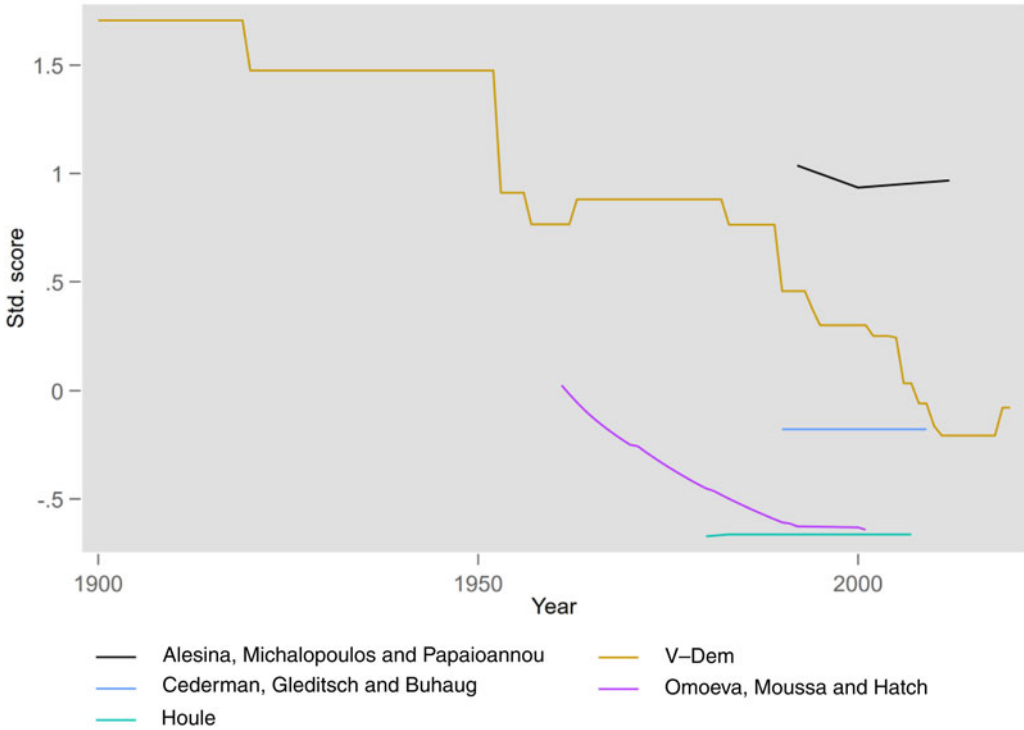


Fig. 3. Temporal variation of measures – Bolivia as example.

Notes: The cross-sectional Baldwin and Huber measure is not included. To ensure comparability, the variables have been standardized (mean of 0; standard deviation of 1).

that develop surveys. Finally, V-Dem (Coppedge et al. 2021a) uses experts' local knowledge to assess ethnic groups based on a prior group definition.⁷

In terms of socio-economic data sources, Baldwin and Huber (2010), Houle (2015) and Omoeva, Moussa and Hatch (2018) all use national household surveys, which include information on both socio-economic well-being and ethnic group affiliations. On the one hand, biased information is unlikely when data are generated from surveys like the DHS, as the original intention was not to assess socio-economic inequalities between ethnic groups.⁸ On the other hand, survey and census data on ethnic issues may entail (intentionally or not) incomplete and biased responses: minority groups may not be accurately represented in national surveys; answers could be significantly affected by the sometimes politically sensitive nature of ethnic identities (Canelas and Gisselquist 2019, 165); and more politically stable countries are more often surveyed in the DHS programme. In the African context, for instance,

⁷This represents an alternative approach to measuring identity-based socio-economic inequality. Instead of first identifying the ethnic groups in a state and subsequently determining their mean socio-economic status (as a basis for calculating, e.g., a Gini measure), a scholar may determine the overall degree to which ethnic identities are associated with socio-economic inequalities. The choice of comparable ethnic categories clearly matters, as exemplified by the fact that V-Dem categorizes Qatar as highly unequal, whereas it receives a score of 0 in the Alesina, Michalopoulos and Papaioannou data, which suggests that V-Dem coders also take the large non-citizen populations into account, whereas the *Ethnologue* does not.

⁸Surveys include the DHS, Multiple Indicator Cluster Surveys (MISCs) and more opinion-focused surveys, such as the regional barometers and WVS (for an overview, see Baghat et al. 2017, 75).

Table 2. Strengths and weaknesses of different data types

	Surveys	Spatial data	Expert coding
Strengths	<ul style="list-style-type: none"> • Most direct measure of relative well-being 	<ul style="list-style-type: none"> • Country coverage • Absence of political biases (for night light) 	<ul style="list-style-type: none"> • Country coverage • Expertise to capture latent phenomena
Weaknesses	<ul style="list-style-type: none"> • Unrepresentative of ethnic composition • Answers affected by politically sensitive nature • Unstable countries/regions under-sampled 	<ul style="list-style-type: none"> • Indirect measure • Data quality of official sources (G-Econ) • Inability to account for overlapping settlement patterns 	<ul style="list-style-type: none"> • Indirect measure • Risk of personal biases • Limited access to relevant information • Inconsistent application of coding criteria • Inability to revisit data sources

Libya, Eritrea, Somalia, Sudan and the Central African Republic are not included (Tetteh-Baah 2019, 31).⁹

In light of the gaps and weaknesses in survey- and census-based data on ethnic inequality, Cederman, Gleditsch and Buhaug (2013) combine data on ethnic groups' settlement areas with the Nordhaus et al. (2006) G-Econ dataset on local economic activity to measure economic ethnic inequalities. Similarly, Alesina, Michalopoulos and Papaioannou (2016) have worked with various proxy measures to combine geocoded night-light data with historical maps of ethnic territories or homelands. While these spatial measures provide higher coverage, they also suffer from numerous drawbacks. Measures of local economic activity hinge on the quality of the underlying sources, and data quality is particularly poor for countries with unreliable official statistics and substantial informal economies (Baghat et al. 2017, 82; Chen and Nordhaus 2011). Night-light emissions from satellite data are an alternative that is independent of governmental bias or the limited quality of official statistical sources. However, like the other measures, this data source is also afflicted by weaknesses, such as constituting a relatively indirect proxy for economic development (Chen and Nordhaus 2011), and official data sources are likely to be more accurate in developed countries (Mellander et al. 2015). Moreover, both spatial methods may lead to measurement error in cases where the ethnic group settlement areas largely overlap. Consequently, spatial approaches cannot accurately estimate the economic inequalities between, for example, the Tutsi and Hutu in Rwanda and Burundi (Alesina, Michalopoulos and Papaioannou 2016, 449; Cederman, Weidmann and Bormann 2015, 807). Returning to the issue of scope and temporal variation, it is worth noting that surveys may or may not be available in a regular time-series format, whereas satellite-based measures – as well as updated ethnic homelands data (for example, GeoEPR) – are available in time series from the 1990s onward. Consequently, satellite-based measures may help track trends across and within countries with improved temporal granularity in the future.

The V-Dem measure is based on coding by multiple country experts of the question as to whether 'basic public services, such as order and security, primary education, clean water and healthcare, [are] distributed equally across social groups' (Coppedge et al. 2021a, 218). The advantage of this approach is the ability to capture latent phenomena based on experts' country-specific knowledge. Given the difficulty of obtaining comparable observable data for public service provision by ethnic group, the assessments made by country experts can become useful when measuring social ethnic inequalities (see Munck, Møller and Skaaning 2020, 341). As with any judgement-based data, however, this approach also has its challenges, including the risk

⁹For further discussion of survey and census-data challenges, see Canelas and Gisselquist (2019, 164-8) and Cederman, Weidmann and Bormann (2015, 808). For general considerations of measuring ethnic identities quantitatively, including census and survey-based measures, see also Bochsler et al. (2021).

of personal biases, limited or biased background information, and reliability issues stemming from inconsistently applied coding criteria (Skaaning 2018, 111–13). As elaborated later, the V-Dem approach increases comparability and reduces the biases inherent in expert codings, which alleviates some of these concerns. However, compared to the other measures, it is much more difficult for us to revisit the data sources, which is relatively easy with the public surveys, G-Econ or night-lights data. As such, it is impossible to verify, for instance, which ethnic groups form the basis of the expert coding or how much relevant information the expert actually has about ethnic inequality regarding a particular year. Both regarding concept and empirics, we simply cannot know exactly what the coders had in mind when arriving at their assessments, as coders are not required to justify their decisions.

The discussed strengths and weaknesses of the data sources are reported in Table 2. As should be apparent, there are no fundamentally superior data sources with the current data availability. In this sense, data choices should be governed by the research question at hand: when studying a specific region, survey measures may prove superior to spatial or expert-coded data, whereas spatial or expert-coded data are more likely to be relevant for global patterns. In that sense, there is a certain trade-off between the geographical and temporal coverage of the data versus its quality (see also Baghat et al. 2017, 82). I discuss the option of combining various data sources at the end of the article.

Aggregation

All of the measures are based on different items of information that must be combined to develop the overall measure. Stewart, Brown and Mancini (2010) consider principles of good measures and make the case for three ways to measure aggregate group inequality: the GGini, GTheil and GCOV, which correspond to the classical Gini coefficient, the Theil index and the coefficient of variation. Instead of calculating inequality based on each individual's income, it assigns each group's mean income to every member of that group (Baldwin and Huber 2010, 646–8; Stewart, Brown and Mancini 2010, 15). The most established measure – the group Gini index – captures the normalized mean difference between all group incomes in a country, weighted by the population size of each group. Like the Gini coefficient, it ranges from 0 to 1 and offers an interpretation related to the Lorenz curve, as described in detail by Baldwin and Huber (2010, 646). The measure takes on its minimum value when the average incomes of all groups in society are the same, and it takes on 1 when one infinitely small group controls all income (Baldwin and Huber 2010, 646).

The group Gini index is adequate in terms of capturing the general level of inequality across countries over time. Alesina, Michalopoulos and Papaioannou (2016) follow this procedure and construct two 'ethnic Ginis'. Similarly, Omoeva, Moussa and Hatch (2018) calculate a group Gini coefficient (as well as a group Theil and a coefficient of variation) for educational attainment across ethnic groups. Although differing in terminology, the Baldwin and Huber (2010) BGI measure is calculated in the same way as the group Gini (Baldwin and Huber 2010, 646). Although similar, the aggregate measure by Houle (2015) departs slightly from the Baldwin and Huber GGini or 'BGI' formula.¹⁰

Another group of measures is employed by scholars who empirically investigate theoretical arguments that only require one group to mobilize. Cederman, Gleditsch and Buhaug's (2013, 143–67) cross-national measure captures the difference between the national average per capita income level and the per capita income of the most (dis)advantaged ethnic group in the country.

¹⁰Houle calculates a population-weighted average of a given country's group-level inequalities. Group-level inequalities, in turn, are calculated following Cederman, Gleditsch and Buhaug's (2013) as the logarithmized ratio between the average income of members of an ethnic group and the average per capita income of the entire country (Houle 2015, 482). The justification for this aggregation is unclear.

The authors are explicit that such a ‘weakest link logic’ is more theoretically relevant when studying civil war onset (Cederman, Gleditsch and Buhaug 2013, 145) because measures based on averages or summed features would discount small, atypical groups, especially in large countries, but such groups might also be the most conflict-prone. While diverging from Stewart, Brown and Mancini’s (2010) suggested approach, the data providers have based their aggregations on explicit theory. Overall, this aggregation procedure means that the measure should differ substantially from the others, as it is not intended to measure overall inequality.

Finally, the V-Dem measure is aggregated using the standard V-Dem methodology. Expert-assigned scores are aggregated through a Bayesian item response theory (IRT) measurement model, which also uses information about coder agreement, self-assigned uncertainty estimates, personal coder characteristics, links between countries based on experts assessing more than one country and responses to vignettes related to the survey questions in order to align the experts’ thresholds and calculate uncertainty estimates (Coppedge et al. 2021b, 16–25; Pemstein et al. 2019). This procedure supposedly reduces potential biases, but it cannot eliminate them altogether. For the purpose of comparison, the measure has been recoded to go from 0 to 1, with higher values indicating greater inequality.

In sum, the datasets aggregate their data in three different ways: first, measures that reflect the entire distribution of resources or access to public services in a society through measures such as the GGini (Alesina, Michalopoulos and Papaioannou 2016; Baldwin and Huber 2010; Houle 2015; Omoeva, Moussa and Hatch 2018); second, ratio measures focusing explicitly on the poorest (or wealthiest) groups in society relative to the country mean (Cederman, Gleditsch and Buhaug 2013); and, third, indices summarizing different experts’ codings, providing an easy-to-interpret number (Coppedge et al. 2021a). As discussed, most measures are aggregated based on existing best practice or explicit theory. In this sense, each measure is appropriate for different research questions. Most clearly, researchers interested in cross-national differences that take into account the entire group distribution should opt for the first or third categories, whereas the second category may be relevant when studying particular ethnic mobilization patterns.

Empirical Comparison

The many differences and similarities in the conceptualizations and measurements of ethnic inequality render it relevant to explore the statistical association between the indices. Comparisons of competitive measures linked to similar background concepts are often assessed by simple correlation tests to clarify whether they tend to tap into the same phenomenon. Following this tradition, Table 3 presents the bivariate correlations between the ethnic inequality indicators. For the purposes of this exercise, I only include the Alesina, Michalopoulos and Papaioannou (2016) *Ethnologue*-based measure, which draws on more recent spatial data (the two measures are correlated at 0.73). Moreover, from Cederman, Gleditsch and Buhaug (2013), I only include the ratio of the poorest group relative to the mean (for a full correlation analysis covering all measures, see Table A8 in the Online Appendix).

Since all of the measures were argued to reflect the same background concept, share causal determinants and affect each other, we would expect them all to be at least moderately correlated. Moreover, measures supposed to capture the same dimension (that is, public services or income/wealth) should show a high level of covariation. In Table 3, the topmost measures (Alesina, Michalopoulos and Papaioannou, Baldwin and Huber, Cederman, Gleditsch and Buhaug, and Houle) reflect the economic dimensions, whereas the lower two (Coppedge et al. and Omoeva, Moussa and Hatch) reflect the social dimension. In addition, since Cederman, Gleditsch and Buhaug use a distinct aggregation procedure, we expect this measure to exhibit lower correlations with the other measures.

The most striking observations from Table 3 are the many weak correlations. Only three out of fourteen are higher than 0.4. To provide a point of comparison, measures of democracy – which

Table 3. Correlations between measures

	Alesina, Michalopoulos and Papaioannou	Cederman, Gleditsch and Buhaug	Houle	Baldwin and Huber	Coppedge et al.	Omoeva, Moussa and Hatch
Cederman, Gleditsch and Buhaug (2013)	0.16 (295)					
Houle (2015)	0.01 (102)	0.12 (970)				
Baldwin and Huber (2010)	n/a	0.04 (46)	0.01 (30)			
V-Dem (2021)	0.55 (484)	0.05 (3042)	0.31 (1641)	0.64 (46)		
Omoeva, Moussa and Hatch (2018)	0.40 (160)	−0.07 (1380)	0.30 (750)	0.05 (21)	0.17 (3983)	
Factor loadings (factor 1)	0.82	0.53	0.67	n/a	0.58	0.54
Factor loadings (factor 2)	−0.33	0.68	0.50	n/a	−0.63	−0.12

Notes: Results refer to bivariate Pearson's *r* correlations (*n* in parentheses), with values over 0.4 in bold. 'n/a' indicates no country-year overlap. The topmost three measures reflect the economic dimension, whereas the lower two reflect the social dimension. Principal component factor analysis (unrotated).

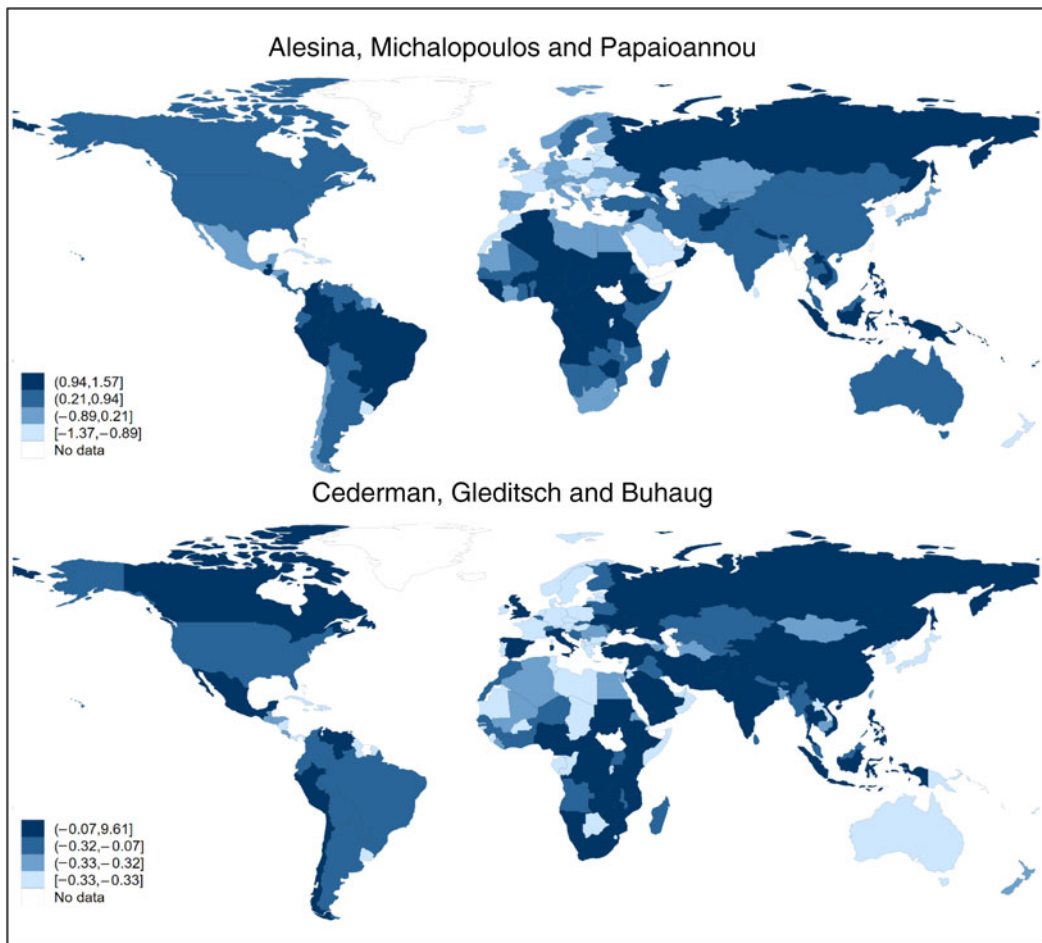


Fig. 4. Standardized values for Alesina, Michalopoulos and Papaioannou and for Cederman, Gleditsch and Buhaug (2000). *Notes:* The figure displays the standardized country scores for the data by Alesina, Michalopoulos and Papaioannou (top panel) and Cederman, Gleditsch and Buhaug (lower panel) in the year 2000. The values are grouped by quartiles.

also vary substantially in terms of their exact conceptualization and measurement – tend to be highly correlated, typically at 0.8 or higher (Marquez 2016, 11–16). In the same vein, despite varying definitions and data sources, conventional measures of socio-economic inequality also tend to be highly correlated (in the range of 0.44–0.90).¹¹ The Alesina, Michalopoulos and Papaioannou measure shows a moderate correlation with the two measures of equal access to social services by V-Dem and Omoeva, Moussa and Hatch, yet it is virtually uncorrelated with the other measures.¹² Moreover, the V-Dem measure is relatively highly correlated with the Baldwin and Huber measure (0.62). Most surprisingly, the Houle measure shows virtually no correlation with the Alesina, Michalopoulos and Papaioannou measure. Moreover, it is only weakly correlated with the Cederman et al. measure, while showing a slightly stronger covariation with the two measures capturing equal access to public services (about 0.3). Perhaps equally surprising, the Cederman et al. G-Econ measure is negatively correlated with the Omoeva, Moussa and

¹¹Referring to disposable income Gini by The Standardized World Income Inequality Database (SWIID), World Development Indicators, a wage share measure and V-Dem’s measure of inequality in ‘access to public services by socio-economic status’ (see Table A8 in the Online Appendix).

¹²This may partly be a result of the fact that night lights may also capture access to public services.

Hatch measure. Contrary to expectations, the two measures of social ethnic inequality (V-Dem and Omoeva, Moussa and Hatch) are only weakly correlated with each other (0.17).¹³ To ensure that these results are not simply an artefact of differences in samples, I conduct a series of additional correlation analyses in the Online Appendix, including overlapping time periods and a core set of countries (see Tables A3–A6). This exercise corroborates the overall pattern of surprisingly low correlations between most measures.

Figure 4 maps the standardized values (mean of 0; standard deviation of 1) for the Alesina, Michalopoulos and Papaioannou and the Cederman et al. data to provide a better sense of the empirical patterns in each dataset and show how individual countries are scored relative to each other. This also provides country or regional experts with an opportunity to assess the face validity of these scores (maps for the other measures are provided in the Online Appendix).

While there is rough agreement on a number of cases, such as Peru, the Democratic Republic of Congo and Ethiopia, several important exceptions also stand out. For instance, there is large disagreement with regard to South Africa: the Alesina, Michalopoulos and Papaioannou measure scores it as surprisingly equal (close to the global mean), whereas it is considered as highly unequal in the Cederman, Gleditsch and Buhaug data. In this case, the Cederman, Gleditsch and Buhaug data are probably closer to the widespread perception that socio-economic group differences remain high in post-apartheid South Africa. To take another example, Saudi Arabia emerges as highly equal in the Alesina, Michalopoulos and Papaioannou data, whereas it scores as highly unequal in the Cederman, Gleditsch and Buhaug data. Finally, Sweden is scored as relatively unequal in the Alesina, Michalopoulos and Papaioannou data, whereas the Cederman, Gleditsch and Buhaug measure scores it as highly equal.¹⁴ Overall, such large disagreements between country scores help to explain the low correlation between these measures (0.16).

In the Online Appendix, I graphically explore the non-correlated measures of Alesina, Michalopoulos and Papaioannou and Houle (see Figure A13). In addition, in Table A7 in the Online Appendix, I conduct a systematic comparison of the measures for a number of countries where the nature of ethnic inequality is well established (South Africa, Guatemala, Peru, Brazil, Nigeria and Switzerland). The takeaway from this exercise is that most measures agree only very roughly on the relative order of a country, with significant variation and hard-to-explain exceptions.

Returning to the question of possible clustering, a principal component factor analysis¹⁵ reveals two principal factors with eigenvalues above 1 (see Table 3). The first factor shows moderate to high loadings by all measures, suggesting that they tap into a common, latent phenomenon. This corresponds to the previously discussed conceptual logic, in which all dimensions reflect socio-economic ethnic inequality. The second factor exhibits moderate loadings by the Cederman, Gleditsch and Buhaug and by the Houle measures, to which there is no straightforward interpretation.¹⁶ In line with the bivariate correlation analysis, the factor analysis reveals no clustering around an economic and social dimension, respectively.

To further probe my interpretations, I follow Adcock and Collier's (2001, 540) recommendation to assess correlations between the measures and those of neighbouring concepts

¹³Moreover, there is no strong covariation between measures based on overlapping data sources: the Cederman et al. and Alesina, Michalopoulos and Papaioannou spatial measures versus the Houle, Baldwin and Huber, and Omoeva, Moussa and Hatch survey-based measures.

¹⁴According to the EPR data underlying Cederman, Gleditsch and Buhaug, there is only one politically relevant group in Sweden.

¹⁵I exclude Baldwin and Huber due to the low $N = 46$, which would exclude most observations. Moreover, I use an interpolated measure of Alesina, Michalopoulos and Papaioannou that fills the years between 1992, 2000 and 2012 to increase the number of observations from 49 to 415, thereby ensuring sufficient overlap to conduct the factor analysis.

¹⁶A partial explanation for the clustering is that both Cederman, Gleditsch and Buhaug (2013) and Houle (2015) use EPR data to classify ethnic groups and calculate their underlying group-level measures in the same fashion. However, they aggregate them differently to the country level.

(discriminant validation). This allows me to check whether the measures diverge from established measures of different yet related concepts. I have thus correlated the various measures with the interpersonal income Gini, interpersonal educational Gini and two measures of ethnic fractionalization. The full analysis is provided in the Online Appendix (see Table A3). Most measures behave largely as we would expect, being moderately correlated with the different neighbouring concepts.

Meanwhile, the Houle measure demonstrates relatively low correlations with the neighbouring concepts (mostly around 0.15–0.20). Strikingly, the Cederman, Gleditsch and Buhaug measure has very low and even negative correlations with the neighbouring concepts. The low correlations of this measure with neighbouring concepts could partly be explained by the ratio aggregation approach, which reflects the poorest (or richest) group in society relative to the mean, whereas the selected neighbouring concepts capture aggregate distributions. As the status of the poorest (or richest) groups in society does not necessarily correspond to the level of ethnic inequality based on the entire distribution of groups, we may see low correlations. In short, these findings further underscore how the choice between ratio-based and aggregate measures (which represent the entire distribution) has important consequences.¹⁷

Do the Differences Matter?

To see whether the reported dissimilarities in conceptualization and measurement affect the findings of empirical analyses, I conduct replication analyses of four prominent studies published in highly recognized journals or book series (Alesina, Michalopoulos and Papaioannou 2016; Baldwin and Huber 2010; Cederman, Gleditsch and Buhaug 2013; Houle 2015). In each replication analysis, I have used the original datasets and Stata code, only substituting the measures of ethnic inequality, which have been standardized to ensure comparability.¹⁸ To save space, I only report the main coefficients in the following, whereas the full regression tables, including controls, are available in the Online Appendix.¹⁹

Alesina, Michalopoulos and Papaioannou (2016, 454) find a negative and statistically significant cross-country association between ethnic inequality and economic development – measured as the log of per capita GDP in 2000. In Figure 5, I report the ordinary least squares (OLS) regressions, relating logged GDP per capita and the different measures of ethnic inequality. In these analyses, only the coefficients for the Alesina, Michalopoulos and Papaioannou and V-Dem measures are negative and statistically significant, whereas Omoeva, Moussa and Hatch have the expected sign yet fail to reach statistical significance. Contrary to expectations, the coefficients for the Houle and the Cederman, Gleditsch and Buhaug measures are positive, and the coefficient for Houle's measure is statistically significant. However, the results could partly be a product of sample differences in country and temporal coverage. I have thus run regressions based on the exact same sample of countries and years (reported in Table A11 in the Online Appendix). Overall, these results show that the differences in Figure 5 are not only a product of the different samples; they also reflect measurement differences. While the number of observations drops dramatically, all coefficients remain signed in the same direction, with the exception of the coefficient for Omoeva, Moussa and Hatch, which turns positive.

Houle (2015) finds that ethnic inequality (BGI) is associated at the country level with an increased risk of democratic breakdown, but only when levels of within-group inequality (WGI) are low. In Figure 6, I report the results from the probit estimations of ethnic inequality's

¹⁷In the Online Appendix, I further discuss how the observed empirical divergence is the product of choices at each of the following levels: (1) ethnic categories; (2) socio-economic data; and (3) aggregation procedures.

¹⁸The samples are thus bounded by the empirical scope of the original analyses.

¹⁹I only include the measure by Baldwin in Huber in Figure 8 because it is only available for one year per country between 1996 and 2005, yielding too few observations for the other replication analyses. Regressions underlying Figures 6–8 are based on an interpolated version of the Alesina, Michalopoulos and Papaioannou measure to provide sufficient observations.

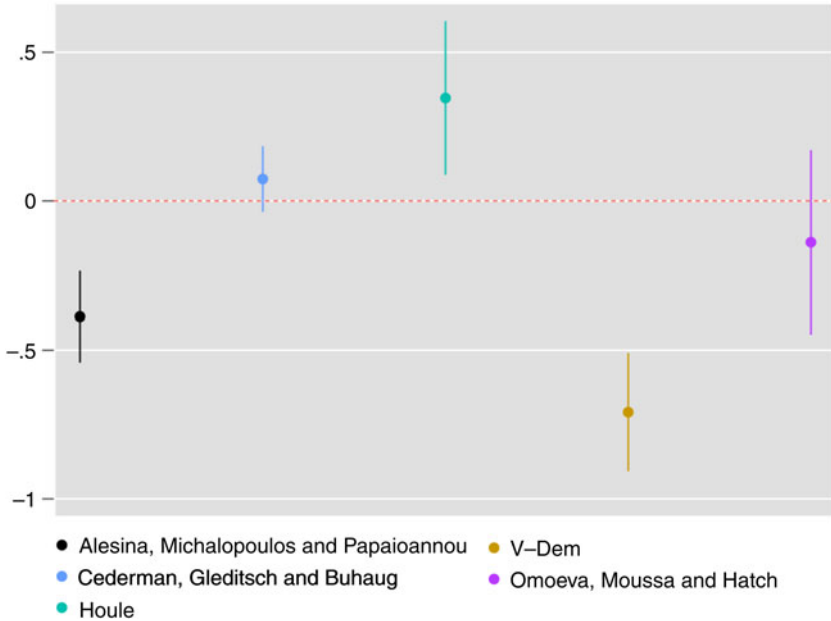


Fig. 5. Replication of Alesina, Michalopoulos and Papaioannou (2016): ethnic inequality and GDP per capita.

Notes: Coefficients for ethnic inequality measures from OLS regressions relating the different measures of ethnic inequality and logged GDP per capita in 2000 (for the underlying full regression table, including controls, see Table A10 in the Online Appendix). Bars indicate 95 per cent confidence intervals.

Source: Based on Alesina, Michalopoulos and Papaioannou (Table 2, Model 1).

association with democratic breakdown. Houle's hypothesis is supported if the coefficient of ethnic inequality is positive. This means that ethnic inequality increases the likelihood of democratic reversals when WGI is zero (Houle 2015, 491). The results from Figure 6 suggest that – in addition to Houle's own measure – the measures by V-Dem, Omoeva, Moussa and Hatch, and Alesina, Michalopoulos and Papaioannou show positive associations as expected, though the latter two are not statistically significant. In contrast, the measure by Cederman, Gleditsch and Buhaug is signed negatively and is very imprecisely estimated. Again, the result may be influenced by differences in country and temporal coverage. Rerunning the analysis with a perfectly overlapping but smaller sample in Table A13 in the Online Appendix, yields similar results, with all variables being signed in the same direction as before.

Cederman, Gleditsch and Buhaug (2013) present country-level evidence that ethnic economic inequality is associated with the risk of civil war onset. Figure 7 shows a replication of the association between the examined ethnic inequality measures and civil conflict. Although all measures are signed in the expected direction, there are important differences. The Cederman, Gleditsch and Buhaug measure is estimated precisely, whereas the others are either very close to zero (Alesina, Michalopoulos and Papaioannou; V-Dem) or have very large confidence intervals (Houle; Omoeva, Moussa and Hatch). Restricting the analysis to a smaller sample for which all measures have coverage yields somewhat similar results, with all coefficients keeping their original signs (see Table A15 in the Online Appendix).

Finally, Baldwin and Huber (2010) find that economic differences between groups are negatively associated with public goods provision. In Figure 8, I show that, with the exception of Cederman, Gleditsch and Buhaug, all measures are negatively associated with public goods provision, though Houle's measure has very large confidence intervals. Since there are only 13 observations for which all measures overlap, checking this replication analysis for sample influence is more difficult.

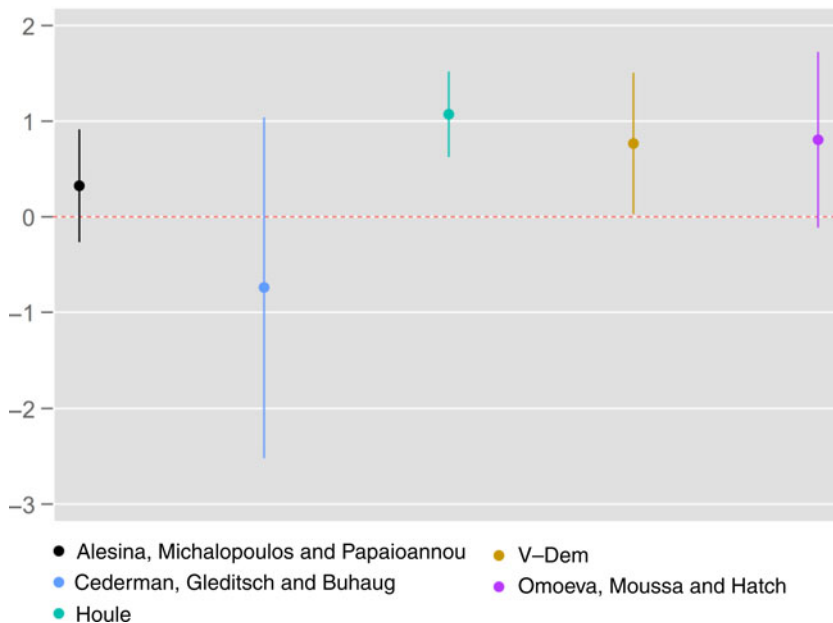


Fig. 6. Replication of Houle (2015): ethnic inequality and democratic breakdown.

Notes: Coefficients for ethnic inequality measures from probit regressions of relationship between various ethnic inequality measures and democratic breakdown (for the underlying full regression table, including controls, see Table A12 in the Online Appendix). Bars indicate 95 per cent confidence intervals.

Source: Based on Houle (Table 2, Model 1).

The findings suggest that the choice of measure has important implications for empirical analysis. The results were generally sensitive to the employed measure, indicating that the examined measures are not interchangeable.

Discussion

An overview of the most important strengths and weaknesses in the different datasets indicates that no measure offers a fully satisfactory response to all of the challenges of coverage, conceptualization, measurement and aggregation (see Table 4). The array of options confronts researchers with a dilemma: which measure is the most valid and reliable measure of ethnic inequality? First, the answer to this question should rest on theoretical foundations regarding a particular research question. If one is interested in mobilization patterns among severely deprived groups, the theoretical arguments would point towards measures like that of Cederman, Gleditsch and Buhaug (2013), which capture this type of socio-economic disparity. If one is interested in the causes and consequences of the entire distribution of resources between ethnic groups, the other examined measures are likely to be more appropriate. Secondly, considering the examined strengths and weaknesses in Table 4, the Alesina, Michalopoulos and Papaioannou (2016) and Coppedge et al. (2021a) measures appear superior in terms of capturing the overall distribution of resources while providing high empirical coverage. That said, researchers considering using one of the measures should still closely study the precise concept and measurement techniques in order to be conscious of biases and errors.

Although this article has focused on highly aggregated country-level measures, more disaggregated research designs are possible and have indeed been applied to some of the examined datasets. Cederman, Gleditsch and Buhaug (2013) and Houle (2015) present their country-level analyses together with group-level analyses, finding that groups with wealth levels far from the

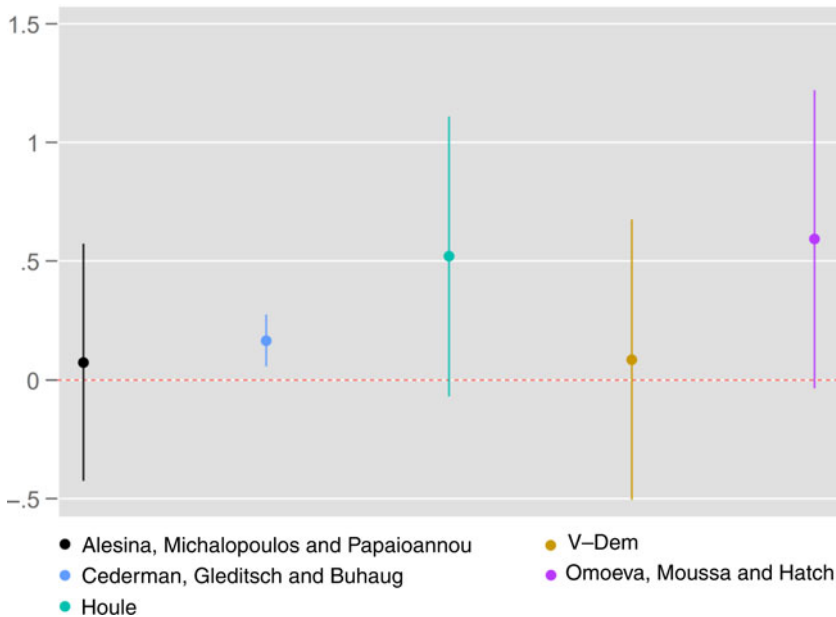


Fig. 7. Replication of Cederman, Gleditsch and Buhaug (2013): ethnic inequality and civil war.

Notes: Coefficients for ethnic inequality measures from logit regressions of relationship between various ethnic inequality measures and civil war onset (for the underlying full regression table, including controls, see Table A14 in the Online Appendix). Bars indicate 95 per cent confidence intervals.

Source: Based on Cederman, Gleditsch and Buhaug (2013, chapter 7, Model 7.1).

country mean are more likely to experience civil war or initiate democratic breakdown, respectively. In the same vein, group-level measures may also help track country-level developments, as illustrated by Bormann et al. (2021), who use night-time luminosity data from 1992 to 2012 and a global sample of ethnic groups to show how the gap between politically marginalized groups and their included counterparts has narrowed over time.²⁰

To the extent that researchers are only interested in two groups – or clusters of groups (for example, politically included/excluded) – ratios of the average achievement of relevant groups constitute a straightforward and intuitive measure of inequality. That said, more aggregate measures are clearly needed if there are larger numbers of groups and we are interested in a single figure representing the entire distribution (Stewart, Brown and Mancini 2010, 16). Beyond the benefits of including an additional level of analysis, more fine-grained group-level data also hold the promise of more transparency, as it becomes possible to validate the scores for individual groups (see, for example, Houle 2015, 488–9). Even when presenting highly aggregate country-level measures, data providers should ideally also make public the underlying group-level values that were used to calculate the aggregate measures. This was found to be a clear limitation with the V-Dem data. Since questions involving ethnic inequality usually have clear group-level implications, it is often advisable to supplement country-level with group-level analyses. Not least given the discussed measurement and aggregation challenges, additional disaggregated analyses constitute one way to increase our confidence in any findings involving the examined ethnic inequality data.

Another encouraging development led by Cederman, Weidmann and Bormann (2015) is the introduction of a group-level composite indicator combining the strengths of three different

²⁰Other examples of group-level measures of socio-economic ethnic inequality include the All Minorities at Risk (AMAR) dataset, which includes the level of socio-economic discrimination of a group (Birnie et al. 2017). Moreover, in addition to Houle (2015), Huber and Mayoral (2019) and Kuhn and Weidmann (2015) offer *within-group* inequality measures.

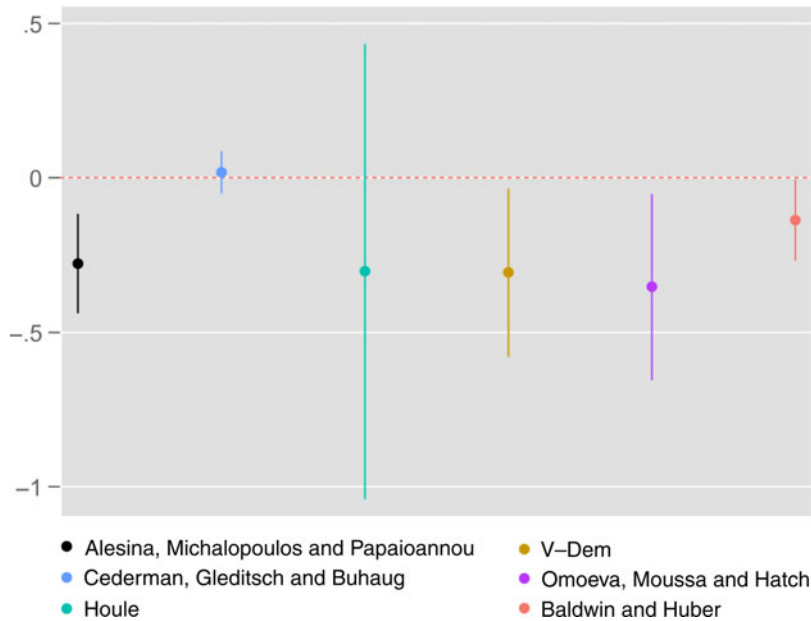


Fig. 8. Replication of Baldwin and Huber (2010): ethnic inequality and public goods provision.

Notes: Coefficients for ethnic inequality measures from logit regressions of relationship between various ethnic inequality measures and public goods provision (for the underlying full regression table, including controls, see Table A16). Bars indicate 95 per cent confidence intervals.

Source: Based on Baldwin and Huber (2010, Table 5, Model 4).

sources of data on local wealth: the G-Econ data; survey data; and night-light emissions combined with geographical data on the settlement of ethnic groups. They weigh economic data more heavily in countries where official statistics are more trustworthy and weight night-light data more heavily where government statistics are poor or lacking. This triangulated measure has not been included in the main discussion and analysis, as it is not publicly available at the country level.²¹ It nevertheless deserves mentioning because such efforts to overcome the respective weaknesses in the different data sources provide a promising avenue towards more valid and reliable measures of ethnic inequality. This avenue is particularly promising if such measures could be made available for longer time periods. Although this is likely to entail further data collection and to be resource intensive, it would allow researchers to investigate a range of new and important questions. Finally, providing triangulated measures with different aggregation procedures is crucial if such measures are to be used for broader research purposes.

An additional way forward is to combine various existing cross-national measures into a composite index. This approach relies on the reasonable assumption that socio-economic ethnic inequality is imperfectly but more or less accurately observed by the compilers of various existing datasets, and that each of them taps into a common dimension. This allows researchers to leverage the enormous effort that scholars have invested in creating an ethnic inequality measure, and it provides a way of dealing with considerable measurement error. Combining measures – based on explicit conceptual foundations – should thus help improve measurement accuracy and

²¹To see how this measure empirically relates to others, I have transformed it into a cross-national indicator following Cederman, Gleditsch and Buhaug's (2013, 150) suggested approach. While the measure exhibits slightly stronger correlations with all of the other evaluated measures than their G-Econ measure, the correlations remain relatively low, i.e., 0.11–0.32 (see Table A18 in the Online Appendix).

Table 4. Summary of strengths and weaknesses of extant datasets

Data provider and index	Strengths	Weaknesses
Alesina, Michalopoulos and Papaioannou: ethnic Gini	<ul style="list-style-type: none"> • Comprehensive spatial scope • Clear, detailed description of measurement and aggregation 	<ul style="list-style-type: none"> • Somewhat restricted temporal scope • Builds on indirect economic proxy
Cederman, Gleditsch and Buhaug: G-Econ/ethnic homeland	<ul style="list-style-type: none"> • Comprehensive spatial scope • Detailed conceptual discussion • Clear, detailed description of measurement and aggregation 	<ul style="list-style-type: none"> • Restricted temporal scope (time invariant) • Builds on crude economic measure
Houle: between-group income inequality	<ul style="list-style-type: none"> • Clear, detailed description of measurement • Face validation of measure 	<ul style="list-style-type: none"> • Restricted and biased empirical scope: only covers democracies • Restricted temporal variation • Aggregation procedure is not justified • Potential survey biases
Baldwin and Huber: Between-group income inequality	<ul style="list-style-type: none"> • Thorough validation procedures; face validity of scores • Clear, detailed measurement discussion 	<ul style="list-style-type: none"> • Severely restricted and biased spatial and temporal scope: only covers 46 democracies • Potential survey biases
Omoeva, Moussa and Hatch: educational group Gini	<ul style="list-style-type: none"> • Relatively comprehensive empirical scope • Multiple, plausible aggregation techniques 	<ul style="list-style-type: none"> • Under-representation of developed countries • Exclusive focus on education • Limited conceptual discussion
V-Dem (Coppedge et al.): access to public services by social group	<ul style="list-style-type: none"> • Comprehensive empirical scope • Sophisticated aggregation procedure, including reliability test • Uncertainty estimates 	<ul style="list-style-type: none"> • Difficult to assess basis of coding decisions • Potential biases in expert coding • Limited conceptual discussion

minimize the impact of idiosyncratic error associated with particular estimates (see Munck, Møller and Skaaning 2020, 345).

In the Online Appendix, I demonstrate this approach with an illustrative example. The resulting index provides plausible values for most countries, and when running the replication analyses with the index, it yields results in line with the original studies in three out of four cases. Since this index is only a relatively crude illustration, the approach should be further exploited using more sophisticated methods, such as latent variable models and IRT, which have been employed for other concepts that are impossible or difficult to observe directly (see, for example, Fariss 2014; Pemstein, Meserve and Melton 2010; Solis and Waggoner 2021).

Conclusion

The literature on ethnic (or horizontal) inequalities has made a series of important contributions to political science. This research has relied on new datasets compiled by scholars creatively exploiting a range of different data sources. This article has compared extant measures, which have been used to operationalize economic and social inequality between ethnic groups at the country level. The assessment has found that measures differ in important ways. Differences in conceptualization and measurement are clearly reflected in the fact that several of the indicators do not correlate highly with each other. Indeed, many of the correlations were surprisingly weak (or even negative). Four replication analyses suggested that the choice of indicator seriously affects our empirical analyses and that the results may depend strongly on the employed indicator. As such, extant measures of ethnic inequality are generally not interchangeable.

Future research can benefit in three ways from the clarifications and critical points put forward in this assessment, which offers helpful information to data users. First, systematic information about the different strengths and weaknesses of various measures of ethnic inequality can help future data users to make conscious choices regarding what measures to use and how.

Secondly, the results suggest that it might be worthwhile to re-examine many of the previous studies using the evaluated measures. Thirdly, the findings can inform the development of new measures that either rely on novel data collection or combine existing indicators in new ways.

Supplementary Material. Online appendices are available at: <https://doi.org/10.1017/S000712342200014X>

Data Availability Statement. Replication data for this article is available in Harvard Dataverse at: <https://doi.org/10.7910/DVN/6LJEGI>

Acknowledgements. I am grateful to Svend-Erik Skaaning, Kees van Kersbergen, Gerardo L. Munck, Kristian Skrede Gleditsch, Christian Houle, Kyle L. Marquardt, Jonathan Doucette, Nicholas Haas, Jacob Nyrup and Lars Johannsen, as well as the three anonymous reviewers and the editor, for highly constructive comments and suggestions.

Financial Support. None.

Competing Interests. None.

References

- Adcock R and Collier D** (2001) Measurement validity: a shared standard for qualitative and quantitative research. *The American Political Science Review* **95**, 529–546.
- Alesina A, Michalopoulos S and Papaioannou E** (2016) Ethnic inequality. *Journal of Political Economy* **124**, 428–488.
- Baghat K et al.** (2017) Inequality and Armed Conflict: Evidence and Data. Background report for the UN and World Bank Flagship Study on Development and Conflict Prevention, Peace Research Institute Oslo (PRIO).
- Baldwin K and Huber JD** (2010) Economic versus cultural differences: forms of ethnic diversity and public goods provision. *American Political Science Review* **104**, 644–662.
- Birnir JK et al.** (2017) Introducing the AMAR (All Minorities at Risk) data. *Journal of Conflict Resolution* **62**, 203–226.
- Bochsler D et al.** (2021) Exchange on the quantitative measurement of ethnic and national identity. *Nations and Nationalism* **27**, 22–40.
- Bormann N-C et al.** (2021) Globalization, institutions, and ethnic inequality. *International Organization* **75**, 665–697.
- Buhaug H, Cederman LE and Gleditsch KS** (2014) Square pegs in round holes: inequalities, grievances, and civil war. *International Studies Quarterly* **58**, 418–431.
- Canelas C and Gisselquist RM** (2018) Horizontal inequality as an outcome. *Oxford Development Studies* **46**, 305–324.
- Canelas C and Gisselquist RM** (2019) Horizontal inequality and data challenges. *Social Indicators Research* **143**, 157–172.
- Cederman L-E, Gleditsch KS and Buhaug H** (2013) *Inequality, Grievances, and Civil War*. Cambridge: Cambridge University Press.
- Cederman L-E, Weidmann NB and Bormann N-C** (2015) Triangulating horizontal inequality: toward improved conflict analysis. *Journal of Peace Research* **52**, 806–821.
- Chandra K** (2006) What is ethnic identity and does it matter? *Annual Review of Political Science* **9**, 397–424.
- Chen X and Nordhaus WD** (2011) Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences – PNAS* **108**, 8589–8594.
- Coppedge M et al.** (2021a) V-Dem Codebook V11.1. Varieties of Democracy (V-Dem) project.
- Coppedge M et al.** (2021b) V-Dem Methodology V11.1. Varieties of Democracy (V-Dem) project.
- Deere CD, Kanbur R and Stewart F** (2018) Horizontal inequalities. In Stiglitz JE, Fitoussi J-P and Durand M (eds), *For Good Measure: Advancing Research on Well-Being Metrics Beyond GDP*. Paris: OECD., pp. 85–100.
- EIC (Education Inequalities and Conflict Project)** (2015) *Education Inequalities and Conflict Database*. Washington, DC: Education Policy and Data Center.
- Fariss CJ** (2014) Respect for human rights has improved over time: modeling the changing standard of accountability. *The American Political Science Review* **108**, 297–318.
- Fearon JD** (2003) Ethnic and cultural diversity by country. *Journal of Economic Growth* **8**, 195–222.
- Fleming CM et al.** (2020) Ethnic economic inequality and fatalities from terrorism. *Journal of Interpersonal Violence*, online ahead of print. doi:10.1177/0886260520976226
- Gordon RG** (2005) *Ethnologue: Languages of the World*, 15th edn. Dallas: SIL Internat.
- Horowitz DL** (2000) *Ethnic Groups in Conflict*. Berkeley, CA: University of California Press.
- Houle C** (2015) Ethnic inequality and the dismantling of democracy: a global analysis. *World Politics* **67**, 469–505.
- Houle C and Bodea C** (2017) Ethnic inequality and coups in sub-Saharan Africa. *Journal of Peace Research* **54**, 382–396.
- Huber JD and Mayoral L** (2019) Group inequality and the severity of civil conflict. *Journal of Economic Growth* **24**, 1–41.
- Jensen C and van Kersbergen K** (2016) *The Politics of Inequality*. London: Palgrave Macmillan Education.
- Kuhn PM and Weidmann NB** (2015) Unequal we fight: between- and within-group inequality and ethnic civil war. *Political Science Research and Methods* **3**, 543–568.

- Leipzig LE** (2022) Replication Data For: 'Measuring Ethnic Inequality: An Assessment of Extant Cross-National Indices', <https://doi.org/10.7910/DVN/6LJEGI>, Harvard Dataverse, V1, UNF:6:IwftBbyc0ulyg3MhZ40mdg = = [fileUNF]
- Marquez X** (2016) A Quick Method for Extending the Unified Democracy Scores. Social Science Research Network Working Paper Series.
- Mellander C et al.** (2015) Night-time light data: a good proxy measure for economic activity? *PloS One* **10**, e0139779, 1–18.
- Munck GL and Verkuilen J** (2002) Conceptualizing and measuring democracy: evaluating alternative indices. *Comparative Political Studies* **35**, 5–34.
- Munck GL, Møller J and Skaaning S-E** (2020) Conceptualization and measurement: basic distinctions and guidelines. In Curini L and Franzese R (eds), *The Sage Handbook of Research Methods in Political Science and IR*. London: Sage Publications, pp. 331–352.
- Nordhaus W et al.** (2006) The G-Econ Database on Gridded Output: Methods and Data. 1–30. Working paper downloaded from: https://gecon.yale.edu/sites/default/files/files/gecon_data_20051206_1.pdf
- Omoeva C, Moussa W and Hatch R** (2018) The Effects of Armed Conflict on Educational Attainment and Inequality. EPDC Research Paper No. 18-03.
- Pemstein D, Meserve S and Melton J** (2010) Democratic compromise: a latent variable analysis of ten measures of regime type. *Political Analysis* **18**, 426–449.
- Pemstein D et al.** (2019) The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data. *V-Dem Working Paper Series* **2019**(21), 1–37.
- Piketty T** (2014) *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.
- Rios-Figueroa J and Staton JK** (2012) An evaluation of cross-national measures of judicial independence. *Journal of Law, Economics, & Organization* **30**, 104–137.
- Skaaning S-E** (2018) Different types of data and the validity of democracy measures. *Politics and Governance* **6**, 105–116.
- Solis J and Waggoner P** (2021) Measuring media freedom: An item response theory analysis of existing indicators. *British Journal of Political Science* **51**, 1685–1704.
- Stewart F** (2002) Horizontal Inequalities: A Neglected Dimension of Development. Queen Elizabeth House Working Paper Series, Working Paper Number 81.
- Stewart F** (2008) *Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies*. Basingstoke: Palgrave Macmillan.
- Stewart F, Brown GK and Mancini L** (2010) Monitoring and Measuring Horizontal Inequalities. *Centre for Research on Inequality, Human Security and Ethnicity* **4**, 1–43.
- Tetteh-Baah SK** (2019) *Measurement and Impact of Horizontal Inequality*. PhD thesis, Swiss Federal Institute of Technology, Switzerland.
- UN (United Nations)** (2020) UN Sustainable Development Goal 10: Reduce Inequality within and among Countries. Available from <https://sdgs.un.org/goals/goal10>
- Vogt M et al.** (2015) Integrating data on ethnicity, geography, and conflict: the ethnic power relations data set family. *Journal of Conflict Resolution* **59**, 1327–1342.
- Wang Y-T and Kolev K** (2019) Ethnic group inequality, partisan networks, and political clientelism. *Political Research Quarterly* **72**, 329–341.
- Weber M** (1976 [1922]) *Wirtschaft und Gesellschaft: Grundriss der Verstehenden Soziologie*. Tübingen: J.B.C. Mohr.
- Weidmann NB, Rød JK and Cederman L-E** (2010) Representing ethnic groups in space: a new dataset. *Journal of Peace Research* **47**, 491–499.
- Wucherpfennig J et al.** (2011) Politically relevant ethnic groups across space and time: introducing the GeoEPR dataset 1. *Conflict Management and Peace Science* **28**, 423–437.
- Ye F and Han SM** (2019) Does ethnic inequality increase state repression? *Canadian Journal of Political Science* **52**, 883–901.