Codebook — Nation-Building Policies Dataset

ETHNICGOODS Team.

Version: March 2026

Abstract

The Nation-Building Policies Dataset (NBP) maps state policies aimed at creating national identification and loyalty among the people situated within the territorial boundaries of nation-states. The dataset traces variations in nation-building strategies across time periods, groups, and countries from 1945 to 2020. This codebook describes the structure of the NBP dataset, the variables it contains, and coding rules used.

Citation When using this dataset in your research, please cite the following publication: vom Hau at el. [Dataset Intro Paper]

Dataset updates The latest version of the NBP dataset can be found **here.**

Contents

Background

The Nation-Building Policies (NBP) Research Platform is a multidisciplinary team of social scientists studying an under-researched dimension of political development. All states—whether strong or weak, democratic or autocratic—confront the challenge of nation-building: how to define national membership and manage ethnic and cultural diversity within their territories.

States have approached this challenge in markedly different ways. Some recognise minority rights through constitutional protections, language education, and religious freedom. Others treat the state as the embodiment of the majority group, promoting a single national language and religion. Still others actively seek to exclude minorities and erase their cultures.

Against this backdrop, the NBP Research Platform systematically maps variations in nation-building policies across countries and over time, identifying the political and historical factors that shape these policies and examining their consequences for collective identities, socioeconomic development, and social and political conflict.

Acknowledgements

The research was initially made possible through the ETHNICGOODS project, funded by the European Research Council (ERC) under the Horizon 2020 program (Grant Agreement No. 864333), and led by Prof. Matthias vom Hau at the Institut Barcelona d'Estudis Internacionals (IBEI).

The Nation-Building Policies (NBP) Dataset benefited from prior conceptual and data-gathering work by other researchers. The Ethnic Power Relations (EPR) Dataset Family and the All Minorities at Risk (AMAR) project were critical reference points for constructing a global country-group-year dataset. We are especially grateful to Lars-Erik Cederman and Manuel Vogt (EPR) and Jóhanna Birnir (AMAR) for their feedback throughout the project. Our research also drew on the Constitute Project, the Government-Sponsored Mass Expulsion (GSME) Dataset, and the Targeted Mass Killing (TMK) Dataset; we thank Zachary Elkins, Meghan Garrity, and Charles Butcher for permission to use these resources. We also acknowledge the support of Matthias Koenig (1971–2026) during the early stages of the project.

Contact Information

For further information, please visit: https://nation-building-research.github.io

Comments and requests may be addressed to:
Email: ethnicgoods@ibei.org
Mailing address:
Nation-Building Policies (NBP) Research Platform
Institut Barcelona d'Estudis Internacionals (IBEI)
UPF Campus de la Ciutadella
Ramon Trias Fargas 25–27
08005 Barcelona, Spain

1 Conceptual Framework

1.1 Definition of a Nation-Building Policy

The NBP dataset documents *de jure* state policies intended to promote collective national identity and regulate ethnic, religious, racial, and linguistic diversity within sovereign states. The unit of analysis is the country–group–year.

A **nation-building policy** is defined as any formal law, constitutional provision, executive order, or codified regulation enacted by a central or first-level subnational government that bears on the relationship between an identified population group and the national polity. Specifically, the policy must either:

(a) promote, regulate, restrict, or prohibit cultural, linguistic, or religious practices of an identified group;
(b) define the terms of national membership, recognition, or civic inclusion for such a group; or
(c) employ coercive instruments—including violence, demographic engineering, or cultural elimination—to alter the group's presence within or relationship to the body politic.

Two features distinguish nation-building policies from routine governance. First, the policy must be *group-referential*: it must target, benefit, or differentially affect an identifiable ethnic, religious, racial, or linguistic community. Universal education reforms, general immigration controls, or economic policies that incidentally affect ethnic groups are excluded unless designed to differentially affect a specific group. Second, the policy must bear on at least one of the two analytical dimensions—national exclusion or cultural institutionalisation—through which states define who belongs to the nation, on what terms, and with what consequences for cultural distinctiveness.

1.2 Policy Domains

The dataset records policies across five domains that existing scholarship has identified as critical for nation-building.

1. **Mass education.** Language of instruction, language classes, religious instruction, school segregation, and affirmative action in education.
2. **Citizenship.** Formal exclusion from and access to national citizenship.
3. **Laws regulating public identity expression.** Restrictions and bans on the public use of language (speech, naming, and media) and religious practice (dress, worship, proselytising, and publications).
4. **Constitutional provisions.** Group mentions (positive/neutral or negative), official language(s), and religion(s), secularism, and freedom of religious practice (the latter two at the country–year level).
5. **Eliminationist policies.** State-sponsored mass killings, forced relocation, spatial segregation (ghettos, reserves, and internment camps), and cultural elimination (destruction of artefacts and heritage sites, removal of children, language and religion suppression, resocialisation camps, forced change of livelihood, and restrictions on intangible cultural practices).

1.3 Scope Conditions

For a policy to be included in the NBP dataset, the following conditions must be jointly satisfied:

1. **State attribution.** A policy must be attributable to a central government, a first-level subnational authority (state, province, or region), or a quasi-state actor exercising effective

sovereign control within internationally recognised borders. Policies enacted by municipal governments, non-state armed groups without territorial control, or international organisations are excluded.

2. **De jure basis.** The policy must be grounded in a formal legal instrument—a constitution, statute, decree, executive order, or codified administrative regulation. De facto practices and informal norms are not coded. One exception is that violence and demographic engineering variables (Domain 5) are coded on the basis of documented events and secondary sources rather than legal instruments because mass killings and forced relocations are rarely codified in law. This exception is flagged throughout the relevant variable descriptions.
3. **Temporal coverage.** The policy must have been in force between 1945 (or the year of independence, whichever is later) and 2020 (or the last year the country existed).
4. **Country threshold.** Only sovereign states with populations exceeding one million are included. Colonial dependencies and mandates were excluded.
5. **Group relevance.** The targeted or affected group must appear in the NBP group list, which was constructed by consolidating the AMAR, EPR, and MR datasets (see Section 4.2).

1.4 Exclusions and Boundary Conditions

The following phenomena fall outside the dataset scope:

De facto discrimination without formal policy basis. Socioeconomic disadvantages, labour-market discrimination, or residential patterns arising from informal norms or market dynamics are not coded unless codified in law. De facto ethnic residential clustering is excluded; legally mandated spatial segregation is included.

Voluntary assimilation. Language shift, religious conversion, cultural adaptation driven by economic incentives or personal choice, and absent state coercion were excluded. This distinction is consequential for cultural elimination variables: an endangered language does not constitute evidence of suppression unless a prohibitory state policy can be identified.

Non-group-referential policies. Skill-based differentiated school systems (e.g. the German secondary model) were not coded as segregation. Universal reforms that incidentally affect groups are excluded, unless they are designed to differentially affect a specific community.

Non-state violence. Violence perpetrated by non-state actors without state sponsorship or direction was excluded from the violence and demographic engineering variables.

International and municipal policies. Policies enacted by international organisations (unless domesticated into national law) or municipal-level governments below the first administrative division were excluded.

1.5 Key Conceptual Distinctions

Two distinctions structure the dataset's coding logic.

Policy vs. outcome. The NBP records *de jure* policies, not their degree of implementation or downstream effects. A law mandating minority-language instruction is coded regardless of whether schools actually teach that language. Researchers interested in implementation can pair NBP with outcome-oriented data (e.g. census-based language-use statistics).

Restriction vs. elimination. The dataset distinguishes partial constraints on cultural practices (captured by the language and religious restriction variables) from systematic state campaigns aimed at extinguishing group distinctiveness (captured by the cultural elimination variables). The threshold is set at the policy level: a restriction limits the use of something, while elimination seeks to eradicate it. For example, the absence of a group's language from the official list of

languages of instruction does not, by itself, constitute cultural elimination; a statutory ban on minority-language instruction does.

2 Dataset Overview

2.1 Coverage

The complete NBP dataset comprises 77,040 country–group–year observations across 1,323 distinct groups in 163 countries, covering the period from 1945 (or the year of independence) to 2020 (or the last year the country existed). The number of countries and groups is not static, given events such as decolonization, secessions, or the arrival of new groups in a country (e.g., through migration).

2.2 Unit of Observation

The primary unit of observation is the country–group–year. If a country does not exist in a given year, that year is omitted; if a group is absent from a country during a specific period, the corresponding years are excluded. Most variables are recorded at the country–group–year level. Some variables are country-level (e.g., `NbConstitutions`, `ConstitutionalStatus`), and others are group-level, time-invariant characteristics (e.g., `LingGpType`, `IndigGp`).

2.3 Summary Variable Table

[Insert summary variable table here]

2.4 N/A and Special Codes

The dataset uses the following values to distinguish between different types of "not applicable" answers:

Code	Meaning
97	Not applicable—atheism (religious variables when the primary segment is non-religious).
98	Not applicable—private schools banned.
99	Not applicable (e.g., no constitution in force; group absent before 1945).
100	Not applicable—state failure (see below).
101	Not applicable—under other jurisdiction (e.g., Palestinian Arabs in Israel).

State Collapse

State collapse refers to a situation where no central authority is recognisable either internally or externally, caused by the total disintegration of state institutions.[1] During such periods, no government can enact or enforce nation-building policies. Therefore, all policy variables are set to 100 ("NA—state failure"). This coding applies to:

- Liberia, 1990–1996
- Sierra Leone, 1998–2002
- Somalia, 1991–2012

Under other jurisdiction

Code 101 ("NA—under other jurisdiction") applies to groups subject to external sovereign authority where the nominal state does not exercise effective policy control (currently: Palestinian Arabs in Israel).

[1] Following Call (2008) and the operationalisation used in EPR 2021.

3 Data Collection and Coding Process

3.1 Coding Process

The NBP dataset was coded primarily between 2022 and 2024 by a team of 22 coders recruited through a competitive process comprising a public call, written application screening, and interviews. Coders were predominantly Master's and doctoral students in the social sciences with relevant regional or linguistic expertise. Initial training was delivered through multi-day coding workshops ("codeathons"), held biannually in Barcelona. These sessions introduced coders to the project's codebook, containing variable definitions, coding instructions, and illustrative examples, and to the standardised data entry interface implemented in LimeSurvey. Following training, coders were assigned to regional teams of two to four members (e.g., Sub-Saharan Africa, Latin America), composed to match coders' language competencies and area expertise. Each team operated under the supervision of a senior researcher—either the principal investigator or one of three postdoctoral researchers—who convened regular meetings to resolve interpretive ambiguities and ensure consistent application of coding rules across cases.

The dataset adheres to the previously described *de jure* coding rule: coders were instructed to record what policies formally stipulate in law rather than how they are implemented or enforced in practice. Primary sources—constitutions, legislation, executive decrees, and official government programs—were therefore privileged over secondary or observational evidence. This principle has consistent implications across variable domains. In coding language-of-instruction policy, for instance, coders recorded whether a language was legally designated as a medium of instruction in relevant legislation or official policy documents, irrespective of whether that designation was consistently observed in classroom practice. Likewise, in coding the exclusion of groups from citizenship, coders attended to eligibility criteria as defined in law rather than practical or administrative barriers to obtaining citizenship or identity documents. Anticipated or planned policy changes were similarly excluded; only provisions formally adopted within the relevant period qualified for coding. This approach ensures that all coding decisions rest exclusively on the documentary record of formally enacted legal and policy provisions, maximising comparability across country-year observations and minimising reliance on potentially inconsistent implementation data.

3.2 Source Hierarchy

Coders drew on sources in the following order of priority:

1. Primary sources (laws, constitutions, statutes).
2. Academic sources (peer-reviewed articles, books, datasets).
3. NGO and intergovernmental reports (e.g., Human Rights Watch, UNESCO).
4. Non-legislative government sources, foreign and domestic (e.g., US State Department reports, national statistical offices).
5. Editorially independent, established newspapers and websites (e.g., Reuters, BBC).
6. Wikipedia (with preference for higher-order sources cited therein).

For violence questions, secondary (e.g., existing datasets) and NGO reports are preferred over government reports, given the well-documented tendency of governments to underreport state-sponsored violence.

3.3 Reliability Procedures

After each country's coding was completed, all observations underwent a comprehensive supervisor review in which every coding decision and supporting source justification was checked against both primary evidence and codebook criteria. This constituted a full verification pass rather than

a sample-based reliability check. Where inconsistencies emerged between a coder's decision and the supervisor's reading of the evidence or codebook, these were resolved through structured discussions, with the revised decision and its rationale recorded in the project documentation.

A second validation layer complemented this supervisory review. Coders were required to produce narrative country profiles describing the nation-building policies of their assigned cases and explaining the reasoning behind consequential coding choices. These profiles were reviewed by other team members, providing an additional layer of independent scrutiny. By requiring coders to articulate their interpretive logic in prose, the process systematically surfaced mismatches between qualitative accounts and coded values, identified inconsistencies across analogous cases, and served as a check against interpretive drift across coders and country contexts.

Senior researchers conducted a final consistency and coverage audit, verifying three dimensions of data quality: internal consistency within countries across variables and time points; temporal consistency within episodes, ensuring that onset, duration, and termination were coded coherently; and cross-national comparability for analogous policies, confirming that equivalent legal provisions received equivalent codes regardless of the regional context or coder responsible.

4 Country and Group Construction

4.1 List of Countries

Only countries with populations exceeding one million are included. Coverage runs from 1945 or the year of independence (whichever is later) to 2020 or the last year the country existed. Colonial dependencies and mandates are excluded. Countries that have undergone name changes are treated as the same entity; countries that dissolve into smaller units are treated as distinct entities from the date of independence or dissolution. Countries that expand territorially continue as the same entity. Table 1 presents the list of countries included in the dataset, the first and last year of inclusion in the dataset, and the number of distinct groups included. N.B.: Some groups are *not* included for all years a country is included in the dataset (see below).

4.2 Construction of the Group List

The basic unit of observation in NBP is the group–country–year, enabling within-country comparisons across groups and cross-national comparisons of group types, as well as aggregation to the country level. Group lists were constructed in two stages: consolidation of existing datasets followed by nesting.

Consolidation of Existing Datasets

NBP builds on three established datasets: **AMAR** (All Minorities at Risk), which selects groups based on cultural distinctiveness and social salience; **EPR** (Ethnic Power Relations), which covers politically relevant groups and may vary its listings over time to reflect shifts in mobilisation and ruling coalitions; and the **Minority Rights (MR) Global Directory**, which is used as a cross-check rather than a primary source.

The baseline includes all groups appearing in AMAR or EPR at any point between 1945 and 2020. Groups appearing in both datasets are treated as one, with identifiers from each recorded. Two categories may be excluded subject to PI approval: groups under 0.5% of the population that are culturally and politically similar to already-listed groups, and groups not meaningfully relevant to nation-building (e.g., non-resident foreign workers). The MR directory then cross-checks the baseline; a group absent from AMAR and EPR is added only if MR lists it and it constitutes at least 1% of the population for at least one decade (0.5% for large federal states, if spatially concentrated). Excluded groups are documented with brief explanatory notes.

Temporal boundaries: Groups are included from 1945 (or independence) to 2020 by default and cannot be dropped once included. Three exceptions apply: recent migrant groups enter from the onset of major migration flows; groups concentrated in territories lost through secession, annexation, or occupation exit at that point; and groups removed by mass expulsion also exit at that point.

Nesting of Groups

Nesting consolidates groups with relevant similarities into composite categories to reduce the coding burden and reflects the common state practice of applying uniform policies to clusters of groups. Nesting is permitted when at least one source dataset proposes the grouping or when case expertise supports it. For example, we followed AMAR in nesting Romanians, Bulgarians and Ukrainians into the category "Eastern Europeans." Groups may be nested even if they possess their own subnational administrative units. All nesting decisions require the approval of a senior researcher, and complex cases are reviewed collectively.

For each nested group, the two or three largest subgroups are identified using the Ethnologue,

EPR-ED, Joshua Project, or AMAR subgroup lists, as of 2020. For migrant nested groups, subgroups are defined by citizenship rather than ethno-linguistic categories.

4.3 Coding Rules

Periodisation Rules

Coders were asked to enter the period(s) during which specific policies (if any) were in force. The year in which an event occurred was captured as the date of the situation/event (start or end of a period). For example, in 1967, the Suharto regime instituted a ban on the public use of Chinese in Indonesia; the start of the policy was considered to be 1967. This rule ensured comparability across group-country-years.

Rules for Coding Nested Groups

Three rules governed the coding of variables for nested groups (composite groups containing two or three subgroups):

1. **Mode/average rule** (group characteristics): the most common characteristic among subgroups was coded for the group. If all three differed, the average was taken. If two subgroups showed different characteristics, the characteristic of the largest was coded.
2. **"If any. . . , then YES" rule** (policies, violence, and demographic engineering): if any subgroup was targeted by the policy or event, "Yes" was coded for the whole group.
3. **Subgroup-level coding** (linguistic and religious composition when the group does not appear as a single group in EPR-ED): each subgroup was coded as a separate entity.

Rules for Decentralised Government Systems

- **Bans and restrictions:** national-level policies were the focus; subnational variations were captured through "part of country" coding. Only first-level administrative divisions (states, provinces) were coded—not municipal-level variation.
- **Education policies:** jurisdiction was first established as national or subnational. If national, subnational regulations were set aside. If subnational, overview reports were drawn upon, with region-specific secondary sources consulted where necessary.
- **Affirmative action:** no differentiation was coded between national and subnational levels, as subnational quotas typically apply to the group as a whole.
- **Periodisation:** the earliest subnational adoption of a policy was taken as the start point, with the latest endpoint as closure.

5 Country-level variables

5.1 Background information

`Country`

Country name.

`ISO3`

ISO 3166-1 alpha-3 country code.

`CountryDateStart`

First year of the study period for the country.

Values:

1945 if the country became independent before 1945.
Year of formal independence if independence took place after 1945.

`CountryDateEnd`

Last year of the study period for the country.

Values:

2020 if the country still exists in 2020.
The year the country ceased to exist otherwise.

`CExpansion`

Whether the country experienced territorial expansion during the study period.

Values:

0 = No
1 = Yes

`CContraction`

Whether the country experienced territorial contraction during the study period.

Values:

0 = No
1 = Yes

Coding notes. Only "major" border changes were recorded: those resulting in the inclusion or exclusion of a group from the group list or a change in the size of a group. Land reclamation was not considered a border change.

`MergedWith`

Name of the country in which the merger occurred. Populated only during merger years.

`SecededFrom`

Name of the independent state from which this country seceded. Populated from the country's start date.

Coding notes. Colonial dependencies were not treated as a "former independent country." Both successor states record secession from the predecessor (for example, both North and South Vietnam recorded secession from Vietnam). If one entity retains the predecessor's name, the split is coded as territorial loss for the retaining entity and secession for the new entity.

5.2 Constitutions

`ConstitutionalStatus`

Constitutional status in the country-year (string variable).

Values:

"Constitution exists"
"Constitution was suspended"
"Country never had a constitution"
"No constitution in force yet"

Coding notes. Constitutional status and changes were identified using data from the Constitute project, supplemented by selected amendments and constitutions not listed on the Constitute website.

`NbConstitutions`

Number of constitutions in force during the study period.

`NbLangOfficial`

Number of official languages recognised in the constitution

`ReliFreedom`

Constitutional provision for religious freedom

Values:

0 = No
1 = Yes, but only for some groups
2 = Yes, universal recognition
99 = NA (no constitution in force)

`Secular`

Constitutional provision for secularism.

Values:

0 = No
1 = Yes, recognises separation of church/state
2 = Yes, explicit mention of secularism
99 = NA (no constitution in force)

`OfficialReli`

Official religion in the constitution

Values:

0 = No

1 = Yes, one religion or denomination
2 = Yes, multiple denomination(s) and/or religion(s)

Coding notes. For countries without constitutions (e.g. the UK, Israel, and Saudi Arabia), these variables were coded based on national laws. Sources include The Religion and State Project, the Government Religious Preference Dataset, ARDA National Profiles, and relevant secondary literature.

6 Group Characteristics

6.1 Group Dates and Presence

`GroupDateStart`

First year of group presence in the country. A group is considered "present" when it becomes a relevant cultural, political, or social group, not when the first individual arrives.

Coding notes. If the country became independent before 1945 and the group was present at that time, this variable takes the value 1945. If the country became independent after 1945 and the group was present at independence, the variable takes the value of the year of independence. For groups arriving later (e.g., labour migrants, refugees), the variable corresponds to the beginning of the relevant policy, war, territorial annexation, or population movement. For nested groups, the earliest arrival date among the subgroups was recorded.

`GroupDateEnd`

Last year of group presence in the country. The default is 2020 (or the end of the country). Differs from the default only if: (a) the country has experienced territorial loss and the group is no longer within its boundaries, (b) a policy completely removes the group, or (c) territorial contraction excludes the group.

6.2 Subgroups

`SubGroup[n]` $n = 1$–3

Names of the three largest subgroups that form a nested group.

Coding notes. The following sources were used to identify subgroups: Ethnologue ("ethnic population" information), EPR-ED (linguistic or religious make-up), Joshua Project, and the AMAR subgroup list. For migrant groups, the subgroups were defined based on citizenship categories (e.g., Chinese, Vietnamese, and Cambodians) rather than ethno-linguistic categories (e.g., Han Chinese and Viet people). Subgroup composition was assessed as of 2020 (or the last year the group/country was present).

`SizeSubGroup[n]` $n = 1$–3

Size of each subgroup (broad categories).

Values:

1 = Majority (50–100%)
2 = Large (30–49%)
3 = Medium (10–29%)
4 = Small (<10%)
5 = Very small (<1%)

N.B.: Some cases coded "no info."

Coding notes. Group size estimates draw on CIA World Factbooks, EPR, AMAR, Ethnologue, and national censuses. Where multiple sources exist, coders prioritized sources matching our group list and applied them consistently to avoid spurious change. For nested groups, size reflects the whole group, not summed subgroups. Population figures align to national jurisdiction, excluding territories under foreign control.

6.3 Match with AMAR and EPR

`GpInAMAR`

Match status with the All Minorities at Risk (AMAR) dataset

Values:

0 = Not in AMAR
1 = Matches AMAR group
2 = Split from AMAR group
3 = Corresponds to multiple AMAR groups
4 = Only subgroup(s) in AMAR

`AMARIDgroup`

AMAR `NUMCODE` for the entire group.

`AMARIDsubgroup[n]` $n = 1$–3

AMAR `NUMCODE` for the subgroups.

`GpInEPR`

Matching status with the Ethnic Power Relations (EPR) dataset

Values:

0 = Not in EPR
1 = Matches EPR group
2 = Split from EPR group
3 = Corresponds to multiple EPR groups
4 = Only subgroup(s) in EPR

Coding notes. EPR group lists may vary over time to reflect changes in mobilisation and ruling coalitions. The EPR "umbrella" variable reports the identifier of a nested group a group was originally part of. If a group is split off from a group listed in EPR (e.g., Mbundu from "Mbundu-Mestico"), EPR-ED was not used to determine the group's religious or linguistic make-up; Joshua Project was consulted instead.

`EPRIDgroup`

EPR `gwgroupid` for the entire group.

`EPRIDsubgroup[n]` $n = 1$–3

EPR `gwgroupid` for the subgroups.

6.4 Migration Status and Border Changes

This applies only if the group enters the list after 1945 or the independence year. If the group is present in 1945 or the independence year, all variables below are coded 99 (NA).

`ArrivedBorderChange`

Group arrived through a change in the country's borders.

Values:

0 = No
1 = Yes
99 = NA

`ArrivedPoliticalMigrantsRefugees`

Groups arrived as political migrants or refugees.

Values:

0 = No
1 = Yes
99 = NA

Coding notes. Political migrants are people who leave their country of origin because of war/conflict and/or exclusionary policies. They are often unable to return because of reasonable concerns about persecution. Political migrants may or may not be asylum seekers.

`ArrivedLabourMigrants`

The groups arrived as labour migrants.

Values:

0 = No
1 = Yes
99 = NA

Coding notes. Labour migrants are people who migrate from their home country to another country for work.

`MigrantBackground`

Whether the group had a migration background.

Values:

0 = No
1 = Yes

Coding notes. Coded 1 if the group either (a) arrived after 1945/independence via labour migration, political migration, or as refugees, or (b) received a substantial influx of new group members through migration after 1945/independence.

6.5 Language

`LingGpType`

Linguistic group type.

Values:

"Group associated with one language"
"Group associated with a single macrolanguage"
"Group associated with multiple languages"

Coding notes. A *macrolanguage* is defined as "multiple, closely related individual languages that are deemed in some usage contexts to be a single language" (e.g., Chinese, Serbo-Croatian, Arabic). "Group associated with a single macrolanguage" was used only if the group *as a whole* is associated with a macrolanguage (e.g., Kurds in Iraq associated with Kurdish). Coding started with Ethnologue: if the group name matched a language, that language was recorded. If not, EPR-ED was consulted. If the group was absent from EPR-ED, the Joshua Project was used.

`Lang1, Lang2, Lang3` *Languages associated with the group*

up to three languages associated with the group.

Coding notes. The following coding logic was applied:
- Based on the group name if it corresponds to a language spoken in the country; otherwise, based on EPR-ED.
- If subgroups were reported: `Lang1` corresponded to the largest subgroup.
- If no subgroups were reported: `Lang1` is the language with the most speakers within the group (ties broken alphabetically).
- If a single language: only `Lang1` was filled.

`LangMacro`

Macrolanguage, if applicable.

`AddiLang`

Additional language associated with group identity (entire group).

Coding notes. Recorded only if, while researching the language questions, coders found evidence of a second level of linguistic cleavage tied to the group's identity (for example, in Cameroon, ethnic groups are each associated with a distinct language *and* an overarching Anglophone/Francophone cleavage).

`StandardArabicLang`

Coded if a group or subgroup is associated with a spoken Arabic variety. Assumes an association with modern standard Arabic.

`EstPopLang_n` $n = 1$–3

Relative size of the language within the group. This applies only to non-nested groups.

`EthnologueStatusLang_n` $n = 1$–3

Language status according to Ethnologue (data collection: 2022/2024).

Values:

1 = Safe
2 = Endangered
3 = Extinct

Coding notes. For currently existing countries, the full Ethnologue EGIDS scale was used as the primary source (levels 0–10, from "International" to "Extinct"). The Agglomerated Endangerment Status (AES) from Glottolog serves as a cross-check when Ethnologue classifies a language as "Safe." For currently non-existent countries, secondary sources were used to assess status on the simplified three-point scale above. For groups associated with *multiple languages* (including nested groups), the "mode/average" rule was applied: the most common status among the listed languages was used. If all three differed, the average was computed using numeric conversions from Ethnologue (6a = 6.33, 6b = 6.67, 8a = 8.33, 8b = 8.67) and converted back to the corresponding status category.

`OtherSpokenLang`

Reported if the status of `Lang1`/`Lang2`/`Lang3` in Ethnologue or AES-Glottolog is "6b. Threatened (Vulnerable)" or worse, and the coder found evidence that group members primarily speak another language.

N.B.: Many groups speak additional languages not captured here (e.g., migrants speaking the national language, multilingual groups like Catalans). This variable is populated only if the other spoken language is distinct from languages already associated with the group.

Coding notes. Values derive from Ethnologue's "Language Use" section, based on keywords indicating language shift or secondary use. Where sources report both shift and co-use, the language shifted to takes precedence. When no group-specific information was available, the country's principal language was coded.

`DateLanguageShift`

Recorded if Ethnologue status is "7. Shifting (Definitely Endangered)" or worse.

Values:

"Before 1945"
"1940s/1950s"
"1960s/1970s"
"1980s/1990s"
"2000s/2010s"

Coding notes. Based on UNESCO's (2003) criteria for "definitely endangered, " the language was used mostly by the parental generation and up. The search keywords were "language loss" or "language shift" plus the language name. AMAR provided supplementary information (variable `LANG`). For groups associated with multiple languages, the earliest date at which one listed language was considered "definitely endangered" was reported.

6.6 Religion

`Reli[n]` $n = 1$–3

The three largest religious segments within the group are ranked by size (religion 1 is the largest).

Coding notes. The coding process started with the EPR-ED dataset that was used to identify the top three religions and their EPR-ED codes. When a group was not found in the EPR-ED, the Joshua Project was consulted. EPR-ED was checked against the Joshua Project; if the findings conflicted, the Joshua Project records were reported. Atheism was treated as functionally equivalent

to a religion. For nested groups, the NBP presents the composition of the group as a whole (based on the religious composition of the three largest subgroups).

`Reli[n]Size` $n = 1$–3

Size of each religious segment (broad categories from the EPR-ED).

`PastReli[n]` $n = 1$–4

Religion(s) associated with the identity of the entire group or segments of the group in the past ranked alphabetically. Not coded for groups entering the list after `CountryStartDate` or if no religion was strongly associated with a group's past identity.

Coding notes. Recorded based on secondary sources, demographic surveys, or qualitative descriptions. Religious beliefs prior to 1800 were not considered. This means that large-scale conversion was only captured when it occurred between 1800 and 1945/independence (e.g., conversion from African traditional religions to Christianity), but not during an earlier time frame (e.g., conversion from traditional religions to Roman Catholicism in Latin America).

`Reli1AssociatedPast`

Whether `Reli1` was associated with the identity of the group in the past.

Values:

0 = `Reli1` not associated with group identity in the past, or practised by fewer than 80% of group members.
1 = `Reli1` practised by more than 80% of group members and associated with the identity of the group or of one of its subgroups in the past.

Coding notes. Indicators of a strong association included a national church, liturgy in the native language, a long history of practising a given religion, or overlap between religious and group identity.

6.7 Other Characteristics

`RacialGpStudyPeriod`

Whether the group was perceived as racial at any point during the study period.

Values:

1 = Yes
0 = No

Coding notes. Racial groups are those to whom external actors—including states and other ethnic groups—attribute shared physical traits (e.g., skin colour, hair texture, and facial features). The NBP dataset combines attention to the legacies of colonial racial domination with coding strategies used in other group-based datasets (AMAR and EPR). A group was coded as a racial group if (a) its name referred to a phenotype ("Black," "White"), (b) AMAR or EPR classified it as originating from a distinct world region, and/or (c) the secondary literature provided evidence of colonial-era racial classification. This characteristic is expected to rarely vary across subgroups of a nested group.

`RacialGpPast`

Whether the group was perceived as a racial group in the past (before 1945/independence).

Values:

1 = Yes
0 = No
99 = NA (group not present in country before the 1945/independence).

`CasteGp`

Whether the group is a caste group.

Values:

1 = Yes
0 = No

`IndigGp`

Whether the group is indigenous.

Values:

1 = Yes
0 = No

Coding notes. This definition draws on the ILO Convention 169 (1989) and the International Working Group for Indigenous Affairs (IWGIA). A group was considered indigenous if its members (1) self-identify as indigenous, (2) trace descent to populations that inhabited the country or wider region at the time of conquest/colonisation or were marginalised in the creation of present nation-states, and (3) retain at least some of their own social, economic, cultural, and political institutions. Classification drew primarily on IWGIA country summaries and publications. For nested groups, indigenous status was assigned at the level of the encompassing group rather than its subcomponents using the "mode/average" rule.

`SpatialConc`

Whether the group was spatially concentrated in a given year. Time-varying.

Values:

1 = Yes
0 = No

Coding notes. Coding was based on the `GROUPCON` variable in AMAR (coded affirmative if GROUPCON ≥ 2). If unavailable, the Joshua Project was consulted ("Location in Country"), SIDE data, EPR Atlas, or Minority Rights. The "mode/average" rule was used for nested groups.

`SDM`

Whether the group had an active self-determination movement in a given year. Time-varying.

Values:

1 = Yes
0 = No
99 = NA (applies only to years after 2012)

Coding notes. Based on Sambanis et al.'s (2018) Self-Determination Movements Dataset. A self-determination movement is a political organisation that makes claims for increased self-determination, including movements for national independence, mergers with another state, or internal autonomy. The variable `active` (`active`=1 $\Rightarrow$ Yes; `active`=0 $\Rightarrow$ No) indicates whether the movement is active. Coverage ends at 2012. For nested groups, the "if any. . . , then YES" rule was used.

`CoreGp`

Whether the group was the core group in a given year. Time-varying.

Values:

1 = Yes
0 = No

Coding notes. Core groups have access to political power and use it to align the state's identity with their own. To code a group as a core group, two criteria must both be met: (1) the group was coded as "dominant" or "monopoly" in EPR, or AMAR's `one_dominant_group` = 1 for the group; and (2) the group was represented as the nation's core in the constitution, the group's language is the official language, or the group is the titular nation in the country's historiography. If criterion (1) was met but criterion (2) was not—for instance, if political dominance reflects coalition politics rather than core-group status—the group was *not* coded as the core group.

6.8 Ranking of Identity Markers

Each variable below ranks the salience of a given dimension of group identity, as of 2020 (or the last year the group was present).

`LangIDmarker`

The role of language as an identity marker

Values:

0 = Not an identity marker
1 = Most important
2 = Second most important
3 = Least important

`ReliIDmarker`

Religion as an identity marker. Same coding.

`RaceIDmarker`

Race as an identity marker. Same coding.

`CasteIDmarker`

Caste as an identity marker. Same coding.

`OtherIDmarker`

Other identity marker. Same coding.

Coding notes. Ranking proceeded in three steps. (1) Relevance: Markers that applied to the group in 2020 were identified. Language was relevant only if group members still speak it; religion only if the group is not predominantly non-religious; race/caste only if the group is perceived as a racial or caste group. (2) Homogeneity: If the group is internally divided on a dimension (e.g., split between Protestant and Catholic), that marker is less important. (3) Distinctiveness: Among the relevant markers, those that set the group apart from other groups in the country were ranked as the most important. Multiple markers may share the same rank. "Other" markers may include regional identity, socioeconomic status, or legal status (e.g., lack of citizenship). This coding was based on the coder's accumulated knowledge; consulting additional literature was not required.

6.9 Demographic Size

`SizeApprox`

Group size as a percentage of the national population (categorical).

Values:

1 = Supermajority (70–100%)
2 = Majority (50–69%)
3 = Large (30–49%)
4 = Medium (10–29%)
5 = Small (<10%)
6 = Very small (<1%)

Coding notes. Sources consulted included CIA Factbooks from different decades, EPR, AMAR, censuses or surveys, and Ethnologue. If multiple sources were available, the source with the group list that most closely matched the NBP list was preferred, and the same source was used across decades where possible. For nested groups, the size of the group as a whole was reported, not the sum of the three largest subgroups.

`SizeEst`

Group size as a percentage of the national population (numeric). Where no figure is available, it is coded as `missing`.

`EstQuality`

Quality assessment of the underlying size estimate.

Values:

1 = Reliable source AND group list close to the NBP one
2 = Reliable source but estimate only for one group
3 = Reliable source but group list not close to NBP
4 = Unreliable
5 = No data

7 Policy Variables

7.1 Segregation and Affirmative Action

`SchoolSeg`

Whether the school system was segregated with respect to this group.

Values:

0 = No segregation
1 = Segregation (part of country)
2 = Segregation (whole country)

Coding notes. The focus was on formal laws and regulations that mandate ethnic, racial, or religious segregation. De facto ethnic segregation or skill-based differentiated systems (e.g., the German secondary school system) were coded as 0 (no segregation). "Whole country" refers to the parts controlled by the government. For nested groups, the "if any..., then YES" rule was used. In decentralised systems, if segregation was a central government policy, "whole country" was coded; if

it varies by subnational unit, "part of country."

`AfAction`

Affirmative action for groups in the field of education.

Values:

0 = No affirmative action
1 = Access to higher education
2 = Quota teacher recruitment
3 = Access + teacher quota
100 = NA – failed state

Coding notes. This variable captures preferential access based on group membership that is *mandated by the state*. It includes quotas, lower admission thresholds, and bonus points on entrance examinations. It does *not* cover parliamentary quota systems, poverty alleviation programs, or special funding for minority groups. Optional policies and policies mandated by private entities are coded as "No." In decentralised systems, subnational policies were included because they typically apply to the group as a whole. For nested groups, the "if any. . . , then YES" rule was used.

7.2 Education Policies

7.2.1 Language in Education

`AnyLangUsed`

Whether any language associated with the group is used as a language of instruction (LOI) or in language classes (LC) in public schools. Excludes `OtherSpokenLang`.

Values:

1 = Yes
0 = No

Coding notes. Coding was based on laws, decrees, programs, or informal regulations, and not on the extent of actual implementation, intentions, or pilot projects. Sources include the UNESCO World Survey of Education, L' Aménagement Linguistique dans le Monde, national education laws, and region-specific secondary literature on language policy.

`LOIPrimary[n] / LOISecondary[n]` *n corresponds to associated language*

Language of instruction at the primary and secondary levels.

Values:

0 = Not a language of instruction
1 = National (throughout the country)
2 = Local (in part of the country)
100 = NA – failed state

Coding notes. LOI refers to the medium through which students are taught across subjects—not a single subject taught in the language. If the language was used as a LOI for only some grades within primary school, it was still coded as an LOI at the primary level. In decentralised systems, coding as "national" required evidence that national law designated it as a country-wide LOI, or that all subnational regulations treated it as such. Coding as "local" required evidence from at least one subnational unit.

`LCPrimary[n] / LCSecondary[n]` *n corresponds to associated language*

Language classes at the primary and secondary levels

Values:

0 = Not used in language class
1 = Mandatory
2 = Optional

Coding notes. If a language class was mandatory for some grades and optional for others, the highest option was chosen (mandatory > optional). "On demand" classes not limited to part of the country were coded as "throughout the country." In decentralised systems, if a language is mandatory in some subnational units and optional in others, the highest option is chosen.

`LangPrivateSchools[n]` *n corresponds to associated language*

Use of the associated language in private schools

Values:

0 = Complete ban
1 = Only in language classes
2 = Fully authorised
98 = Private schools banned
100 = NA – failed state

Coding notes. Only laws and decrees that regulated private education were considered. Coders were instructed not to consider international or consular schools. A "private school" is an institution controlled and managed by a non-governmental organisation (UNESCO definition).

7.2.2 Religion in Education

`AnyReliGroupTaught`

Whether any denomination or religion is associated with the group is taught in religious instruction (RI) classes in public schools.

Values:

0 = No RI related to the group's religious make-up
1 = At least one other denomination within the same faith is taught
2 = At least one denomination/religion directly associated with the group is taught

`ReliInst[n]` *n corresponds to associated faith/denomination*

Whether an associated religion is taught in public schools

Values:

−1 = No RI
0 = Faith not taught in RI
1 = Faith taught but not denomination
2 = Denomination taught
97 = NA – atheism
100 = NA – failed state

Coding notes. Sources: Religion and State Project (variable `VED1X`), Religlaw, Annual Reports on International Religious Freedom, and regional compilations (e.g., *Religious Education at Schools in Europe*). General religious studies courses were not counted as instruction in a specific religion. If the denomination itself is not listed but the broader religion is taught (e.g., "Christian education"), "Yes" was coded for all denominations within that faith. In decentralised systems, if RI is a subnational matter and at least one subnational unit teaches a religion, that religion is considered to be taught. The highest option (mandatory > optional) was always selected.

`ReliRIWhich[n]` *n corresponds to* ***`Reli[n]`***

How associated religion is taught.

Values:

1 = Optional class
2 = Mandatory for some
3 = Mandatory for all
99 = NA – not taught

`OtherDenRIWhich[n]` *n corresponds to* ***`Reli[n]`***

Where `Reli[n]` is not itself taught but another denomination within the same faith is: how is that denomination taught? The strongest value recorded.

Values:

1 = Optional class
2 = Mandatory for some
3 = Mandatory for all

`ReliPrivateSchools[n]` *n corresponds to associated faith/denomination*

Status of the associated religion/denomination in private schools

Values:

−1 = No Religious Instruction
0 = Ban on this religion in RI
1 = Can be taught in RI
97 = NA – atheism
98 = Private schools banned
100 = NA – failed state

N.B.: Coded at denomination level, not faith level.

7.3 Linguistic and Religious Restrictions/Bans

7.3.1 Language Restrictions

`AnyRestriLang[n]` *n corresponds to associated language*

Whether any restrictions or bans apply to the associated language.

Values:

0 = No
1 = Yes

Coding notes. Language restrictions are *formal laws or executive orders* that limit the use of a given language. A language is "totally banned" if it is prohibited anytime and anywhere; it is "restricted" if it is allowed only in limited times or places. The coding covers three domains: public speaking, naming, and media use. In decentralised systems, national-level bans were checked first, then secondary literature; if a restriction existed in at least one subnational unit but was not a national ban, it was coded as "in part of the country." For macrolanguages, if any dialect listed in the coding faced no restrictions, the macrolanguage was coded as unrestricted.

`RestriPublicSpeaking[n]` *n corresponds to associated language*

Restrictions on the public use of the language in speech.

Values:

1 = Total ban (whole country)
2 = Total ban (part of country)
3 = Restrictions (whole country)
4 = Restrictions (part of country)

Coding notes. "Public speaking" refers to statements made in public spaces (e.g., speeches in public squares, announcements in transportation hubs, or public prayers in religious buildings); private conversations and speech in private spaces were excluded. Values 3 or 4 (Restrictions) were assigned when the language was permitted in some public contexts but prohibited in others.

`RestriNaming[n]` *n corresponds to associated language*

Restrictions on the use of the language in naming.

Values:

1 = Total ban (whole country)
2 = Total ban (part of country)
3 = Restrictions (whole country)
4 = Restrictions (part of country)

Coding notes. Covers the language(s) permitted for naming persons, local administrative units, and organisations. Values 3 or 4 (Restrictions) were assigned when prohibitions applied to only one naming dimension (e.g., personal names but not place names, or vice versa). Whether the group's writing system was accepted for official registration was not considered.

`RestriMediaUse[n]` *n corresponds to associated language*

Restrictions on the use of the language in the media.

Values:

1 = Total ban (whole country)
2 = Total ban (part of country)
3 = Restrictions (whole country)
4 = Restrictions (part of country)

Coding notes. Covers broadcasting (television and radio) and print newspapers. Values 3 or 4 (Restrictions) were assigned when prohibitions applied to only one media domain (e.g., broadcasting but not print, or vice versa).

`LangRestri[n]` *n corresponds to associated language*

Other language restrictions.

Values:

1 = Total ban (whole country)
2 = Total ban (part of country)
3 = Restrictions (whole country)
4 = Restrictions (part of country)

7.3.2 Religious Restrictions

`AnyRestriReli[n]` *n corresponds to associated religion*

Whether any restrictions or bans apply to the associated religion.

Values:

0 = No
1 = Yes

`RestriReligiousWear[n]` *n corresponds to associated religion*

Restrictions on religious dress

Values:

0 = No restrictions
1 = Ban (whole country)
2 = Ban (part of country)
3 = Restrictions (whole country)
4 = Restrictions (part of country)
97 = NA – atheism

`RestriWorshipPlaces[n]` *n corresponds to associated religion*

Restrictions on places of worship. Same coding as `RestriReligiousWear[n]`.

`RestriReliPubli[n]` *n corresponds to associated religion*

Restrictions on public religious expression or proselytising

Values:

0 = No restrictions
1 = Ban/restrictions (whole country)
2 = Ban/restrictions (part of country)
97 = NA – atheism

`RestriConversions[n]` *n corresponds to associated religion*

Restrictions on religious conversion. The same coding as `RestriReliPubli[n]`.

Coding notes. Sources: Annual Reports on International Religious Freedom (US State Department), Religion and State Project, and secondary literature. National-level restrictions were checked first; in decentralised systems, if a restriction exists in at least one subnational unit, coded as "in part of the country."

7.4 Constitutional Provisions

`GroupReco`

Whether the group is mentioned in the Constitution.

Values:

−1 = Negative mention
0 = No mention
1 = Positive/neutral mention

Coding notes. Based on Constitute Project. The constitutional text was searched for references to groups that are part of the NBP group list. A mention was considered "positive/neutral" if the group was recognised, protected, or included in the definition of the nation. A mention was considered "negative" if the group is mentioned for purposes of political disenfranchisement, segregation, or exclusion. Broad references (e.g., "the people of [Country]") did not constitute group mentions. For umbrella references (e.g., "indigenous peoples"), all subgroups subsumed under the umbrella were coded as mentioned. For nested groups, the "mode/average" rule was used.

`GroupRecoLevel`

Specificity of constitutional mentions.

Values:

1 = Exact mention (the group is named directly)
2 = Umbrella group mentioned (e.g., "indigenous peoples" but not specific subgroups)
98 = Group not mentioned

`LangReco[n]` *n corresponds to associated language*

Whether the associated language holds official status in the constitution.

Values:

0 = Not an official language in the constitution
1 = Official language in the constitution
99 = NA – no constitution in force
100 = NA – failed state

Coding notes. Based on the Constitute Project. If no official language was mentioned, it was coded as 0. For federations, only the federal constitution was considered. The term "state language" (common in former Soviet states) was treated as equivalent to "official language." The protection of language use or designation as a medium of communication in schools or parliament did *not* constitute official status. When a distinction existed between "official" and "national" languages, the "official" languages were coded.

`ReliOfficial[n]` *n corresponds to associated religion*

Whether the associated religion is defined as the official state religion.

Values:

0 = No
1 = Other denomination of the same faith is official
2 = Faith is official
3 = Religion/denomination is official
97 = NA – atheism

99 = NA – no constitution in force

Coding notes. Financial support, endorsement, recognition, or registration requirements do not qualify as official religion status. In decentralised systems, the "if any. . . , then YES" rule was applied.

`ReliAcknow[n]` *n corresponds to associated religion*

Whether the associated religion is acknowledged in the Constitution (short of official status).

Values:

0 = No
1 = Other denomination acknowledged
2 = Faith acknowledged
3 = Religion/denomination acknowledged
97 = NA – atheism
99 = NA – no constitution in force

Coding notes. A religion was coded as "acknowledged" if it was not the official religion but its name was mentioned in the constitution, or if the constitution included references to a specific religion (e.g., in Ireland's 1937 Constitution, Roman Catholic was acknowledged because the preamble states that "In the Name of the Most Holy Trinity, from Whom is all authority and to Whom, as our final end, all actions both of men and States must be referred, We, the people of Éire, Humbly acknowledging all our obligations to our Divine Lord, Jesus Christ.").

7.5 State-Based Violence Against Group Members

`ViolenceAgainstGroup`

Summary indicator: Whether any state-based violence was directed at the group in a given year. Equals 1 if `LowLevelViolence` = 1 or 2, or if `MassViolence` = 1 or 2.

Values:

0 = No
1 = Yes

`LowLevelViolence`

Low-level state violence against the group

Values:

0 = No
1 = Yes – some members targeted
2 = Yes – all members targeted
99 = NA – no violence recorded

Coding notes. *State-based violence* is the use of armed force by the government resulting in the intentional death of at least 100 non-combatants from the group in a period of sustained violence. Violent tactics could be employed by any government agency (military, police, or special security forces). Sources included: Anderton's compilation of mass atrocities, Targeted Mass Killing (TMK) Dataset, UCDP, Political Instability Task Force, and secondary literature. For nested groups, the "if any. . . , then YES" rule was used.

`MassViolence`

Mass State Violence against the Group. The same coding as `LowLevelViolence`.

Coding notes. A mass killing is defined as "any event in which the actions of state agents result in the intentional death of at least 1,000 noncombatants from a discrete group in a period of sustained violence" (Ulfelder and Valentino 2008, p. 2). 2). 2).

`ViolenceDuringWar`

Whether violence occurred in the context of armed conflict.

Values:

0 = No
1 = Yes – during civil war
2 = Yes – during international war
99 = NA – no violence recorded

Coding notes. A war is defined by the Correlates of War (COW) project as sustained combat involving organised armed forces, resulting in at least 1,000 battle-related fatalities within a twelve-month period.

`NameCivilWar`

Name of the civil war, if applicable.

`NameIntWar`

Title of the international war, if applicable.

7.6 Demographic Engineering

7.7 Relocation and Segregation

`Relocation`

Forced relocation of group members

Values:

0 = No relocation
1 = Relocation (domestic)
2 = Relocation (international)
3 = Relocation (domestic and international)

Coding notes. *Forced relocation* refers to a systematic, government-sponsored policy to remove an ethnic, racial, religious, or national group without individual legal review and without recognizing a right to return. This includes targeted movements, such as deportations, resettlements, and population transfers, that originate from state policy and may displace people internally or across borders. This variable was coded by first checking the Government-Sponsored Mass Expulsion Dataset (GSME) for relevant cross-border expulsions and then consulting secondary sources to identify internal displacements. An event was coded only when the state systematically removed a discrete group because of shared group characteristics (not incidental displacement from violence, famine, or natural disasters) and met an annual 1,000-person threshold.

`SpatialSeg`

Summary indicator: whether any form of legally mandated spatial segregation was imposed on the group in a given year. Equals 1 if any of `SegGhettos`, `SegReserves`, `SegIntCamps`, or

`SegOther` = 1.

Values:

0 = No
1 = Yes

Coding notes. *Spatial segregation* refers to legally mandated, group-based physical separation of an ethnic group—or the preponderance of its members—from the core group or other groups, implemented and, where necessary, enforced by the state. The defining criterion is the presence of a formal legal or policy instrument compelling separation; de facto residential clustering driven by economic stratification or voluntary association was not considered.

`SegGhettos`

Legally designated urban residential zones to which group members are confined or from which the core group was formally excluded.

Values:

0 = No
1 = Yes

Coding notes. This variable excludes de facto segregation—such as ethnically homogeneous residential enclaves arising from voluntary settlement patterns—and policies that merely encourage voluntary separation, including housing incentives or the localised provision of group-specific services. The defining criterion is a formal legal or administrative instrument that mandates residential confinement on the basis of group membership. Coders recorded whether such a policy was in force in a given year and, if so, which group or groups were subject to it.

`SegReserves`

Spatially bounded territories—including reservations, reserves, and homelands—to which a group is legally assigned or restricted, typically restricting freedom of movement and land rights beyond the designated area.

Values:

0 = No
1 = Yes

`SegIntCamps`

Internment or detention camps in which group members are collectively confined based on group membership rather than on individual legal proceedings. Distinct from penal incarceration; requires evidence that confinement is group-targeted rather than individually adjudicated.

Values:

0 = No
1 = Yes

`SegOther`

Other formally enacted arrangements of group-based spatial separation not captured by the above categories. Coders must provide a source-supported description of the policy mechanism in `SegOtherWhich`.

Values:

0 = No
1 = Yes

7.7.1 Cultural Elimination

`Cultural Elimination`

Summary indicator: Whether any cultural destruction policies were directed at the group in a given year. 0 = No; 1 = Yes. Derived: equals 1 if any component variable below = 1.

Coding notes. *Cultural elimination* refers to state policies targeting language, religion, institutions, material culture, and collective memory with the intent to extinguish the cultural distinctiveness of a specific group, while stopping short of systematic killing or mass forced displacement. The concept is specified by intent at the *policy* level rather than by outcome: a suppression campaign that fails to extinguish a minority tongue still constitutes cultural elimination if it reflects deliberate state policy aimed at that objective. The policy must be attributable to a central government or first-level subnational authority. Four phenomena were explicitly excluded. (1) *Routine discrimination*: unequal treatment without an identifiable eliminationist policy program (e.g., underinvestment in minority-language schooling through neglect, as distinct from criminalising minority-language instruction).(2) *Voluntary assimilation*: language shift or cultural adaptation driven by economic incentives or personal choice, absent state coercion. (3) *Demographic engineering through population movement*: state-directed settlement of ethnic majorities into minority-inhabited regions; the primary mechanism here is population relocation, placing it beyond the scope of this variable. (4) The mere absence of a group's language from the official list of languages of instruction was *not* coded as cultural elimination. Statutory bans on minority-language instruction or coercive transfer of children to institutions where the group's language and culture were systematically prohibited were coded as cultural elimination.

The component variables are:

`DestructionArtifacts`

State-directed destruction of moveable cultural assets of symbolic significance (e.g., books, archives, artworks, religious objects). 0/1. Excludes collateral damage from combat. Requires evidence of intent to target the group's cultural heritage, tied to group identity.

`DestructionSites`

State-directed or state-sanctioned demolition of immovable heritage of symbolic significance (e.g., religious shrines, monuments, architecture tied to group identity). 0/1. Excludes non-state actor violence unless state-supported.

`RemovalChildren`

Systematic, government-organised coerced transfer of children from the targeted group to families or institutions of the dominant group. 0/1. Threshold for inclusion: systematic scale ($\geq$ 1,000 transfers per year) or codified policy mandating removal.

`ForcedChangeWayOfLife`

State-imposed coercive alteration of traditional livelihoods, social organisation, or dwelling patterns (e.g., enforced sedentarisation, collectivisation targeting specific ethnic groups, prohibition of customary occupations). 0/1. Voluntary adaptation was excluded.

`LanguageSuppression`

State-enforced suppression of a minority group's native language through prohibition of its use in public administration, education, or media; criminalisation of daily speech; mandatory replace-

ment of toponyms; or bans on publication in the language. 0/1. Code only outright bans, not partial restrictions or lack of state support. Changing of place names (*ChangingToponyms*) is subsumed under this variable when part of a broader language suppression program; it was not coded independently.

`ReligionSuppression`

State-enforced prohibition of religious practice, encompassing bans on worship, religious education, clergy activities, possession of sacred texts, forced conversion, or public expression of religious identity. 0/1. The variable subsumes *forced conversion* when it is state-directed.

`ResocializationCamps`

Establishment of involuntary detention facilities designed to replace group identity with state-approved cultural norms through ideological re-education. 0/1. Distinct from standard penal incarceration; requires evidence that the camp system specifically targets an ethnic or religious group for identity reconstruction.

`IntangibleCultureRestrictions`

Legal prohibitions on non-material cultural practices constitutive of group identity, including traditional dress, music, festivals, and art forms, not already captured under `LanguageSuppression` or `ReligionSuppression`. 0/1. Focuses on public symbolic erasure codified in law or executive orders.

7.8 Citizenship

`GpExcludedCitizenship`

Whether the group is formally excluded from citizenship.

Values:

0 = No
1 = Yes

Coding notes. This captures *de jure* citizenship rights, not barriers to accessing identity documents. When multiple citizenship categories exist, the coding concerns "full citizenship." Sources: Global Nationality Laws Database (laws from 1985/1989 onwards, plus country profiles). For nested groups, the "if any. . . , then YES" rule was used.

A Tables

Table 1: List of countries included in NBP, by world region

Country	Start	End	Nb Groups[a]
Africa			
Algeria	1962	2020	5
Angola	1975	2020	11
Benin	1960	2020	7
Botswana	1966	2020	11
Burkina Faso	1960	2020	14
Burundi	1962	2020	3
Cameroon	1960	2020	8
Central African Republic	1960	2020	8
Chad	1960	2020	17
Congo (Republic of)	1960	2020	8
Democratic Republic of Congo	1960	2020	35
Djibouti	1977	2020	5
Egypt	1945	2020	7
Equatorial Guinea	1968	2020	5
Eritrea	1993	2020	8
Eswatini	1968	2020	2
Ethiopia	1945	2020	18
Gabon	1960	2020	13
Gambia	1965	2020	8
Ghana	1957	2020	10
Guinea	1958	2020	6
Guinea-Bissau	1974	2020	7
Ivory Coast	1960	2020	6
Kenya	1963	2020	18
Lesotho	1966	2020	2
Liberia	1945	2020	14
Libya	1951	2020	6
Madagascar	1960	2020	12
Malawi	1964	2020	8
Mali	1960	2020	8
Mauritania	1960	2020	4
Mauritius	1968	2020	5
Morocco	1956	2020	4
Mozambique	1975	2020	10
Namibia	1990	2020	13
Niger	1960	2020	6
Nigeria	1960	2020	15
Rwanda	1962	2020	3
Senegal	1960	2020	5
Sierra Leone	1961	2020	11
Somalia	1960	2020	7
South Africa	1945	2020	15

continued

Table 1: List of countries included in NBP, by world region

Country	Start	End	Nb Groups[a]
South Sudan	2011	2020	11
Sudan	1956	2020	20
Tanzania	1961	2020	24
Togo	1960	2020	5
Tunisia	1956	2020	5
Uganda	1962	2020	16
Zambia	1964	2020	13
Zimbabwe	1965	2020	8
Americas			
Argentina	1945	2020	4
Bolivia	1945	2020	4
Brazil	1945	2020	6
Canada	1945	2020	8
Chile	1945	2020	4
Colombia	1945	2020	3
Costa Rica	1945	2020	3
Cuba	1945	2020	3
Dominican Republic	1945	2020	2
Ecuador	1945	2020	4
El Salvador	1945	2020	2
Guatemala	1945	2020	4
Haiti	1945	2020	2
Honduras	1945	2020	3
Jamaica	1962	2020	2
Mexico	1945	2020	6
Nicaragua	1945	2020	5
Panama	1945	2020	4
Paraguay	1945	2020	3
Peru	1945	2020	6
Trinidad and Tobago	1962	2020	3
United States of America	1945	2020	9
Uruguay	1945	2020	2
Venezuela	1945	2020	4
Asia			
Afghanistan	1945	2020	11
Armenia	1991	2020	3
Azerbaijan	1991	2020	5
Bahrain	1971	2020	8
Bangladesh	1971	2020	7
Cambodia	1953	2020	5
China	1949	2020	38
Cyprus	1960	2020	2
Georgia	1991	2020	6
India	1947	2020	23
Indonesia	1950	2020	31

continued

Table 1: List of countries included in NBP, by world region

Country	Start	End	Nb Groups[a]
Iran	1945	2020	16
Iraq	1945	2020	9
Israel	1948	2020	6
Japan	1945	2020	6
Jordan	1946	2020	7
Kazakhstan	1991	2020	8
Kuwait	1961	2020	12
Kyrgyz Republic [Kyrgyzstan]	1991	2020	5
Laos	1953	2020	5
Lebanon	1946	2020	12
Malaysia	1957	2020	7
Mongolia	1945	2020	4
Myanmar [Burma]	1948	2020	14
Nepal	1945	2020	7
Oman	1951	2020	7
Pakistan	1947	2020	10
People's Democratic Republic of Yemen	1967	1990	5
Philippines	1946	2020	4
Qatar	1971	2020	8
Saudi Arabia	1945	2020	16
Singapore	1965	2020	4
South Korea	1948	2020	2
South Vietnam	1954	1975	3
Sri Lanka	1948	2020	5
Syria	1946	2020	7
Taiwan	1945	2020	4
Tajikistan	1991	2020	6
Thailand	1945	2020	7
Timor Leste	2002	2020	2
Turkey	1945	2020	7
Turkmenistan	1991	2020	4
USSR	1945	1990	47
United Arab Emirates	1971	2020	10
Uzbekistan	1991	2020	7
Vietnam	1945	2020	10
Yemen (Arab Republic of Yemen)	1945	2020	7
Europe			
Albania	1945	2020	5
Austria	1955	2020	5
Belarus	1991	2020	4
Belgium	1945	2020	6
Bosnia and Herzegovina	1992	2020	4
Bulgaria	1945	2020	5
Croatia	1991	2020	6
Czech Republic	1993	2020	4

continued

Table 1: List of countries included in NBP, by world region

Country	Start	End	Nb Groups[a]
Czechoslovakia	1945	1992	6
Denmark	1945	2020	3
Estonia	1991	2020	4
Finland	1945	2020	4
France	1945	2020	19
Germany (Democratic Republic)	1949	1990	1
Germany (Federal Republic)	1949	2020	11
Greece	1945	2020	6
Hungary	1945	2020	5
Ireland	1945	2020	4
Italy	1945	2020	10
Kosovo	2008	2020	6
Latvia	1991	2020	6
Lithuania	1991	2020	4
Moldova	1991	2020	6
Netherlands	1945	2020	8
North Macedonia	1991	2020	7
Norway	1945	2020	3
Poland	1945	2020	7
Portugal	1945	2020	3
Romania	1945	2020	4
Russia	1991	2020	39
Serbia	2006	2020	8
Slovakia	1993	2020	3
Slovenia	1991	2020	5
Spain	1945	2020	9
Sweden	1945	2020	5
Switzerland	1945	2020	9
Ukraine	1991	2020	11
United Kingdom	1945	2020	13
Yugoslavia	1945	2006	9
Oceania			
Australia	1945	2020	9
New Zealand	1947	2020	4
Papua New Guinea	1975	2020	10

[a] Some groups are not included for all years a country is included in the dataset.
Source: NBP dataset, authors' computation.