Standardized ethnically attributed mass surveys (SEAMS) dataset

v1.0

Codebook

Andreas Juon, June 2023

Table of contents

1. Introduction 2
   1.1. Purpose of SEAMS 2
   1.2. Included files 3
2. Coverage 4
3. Variable list 6
4. Standardized variables on socio-political attitudes: question items 8
   4.1. Political attitudes 9
   4.2. Social attitudes 11
5. Standardized codes for survey respondents' ethno-cultural identities 13
   5.1. Manual coding of direct links (respondent ethnicity, region, language, religion, and phenotype) 14
   5.2. Semi-automatic ethnic attribution, based on direct and indirect links 14
References 18

Suggested citation:
When using SEAMS' standardized variables on government (dis-)satisfaction, perceptions of discrimination, social attitudes, and basic demographic information or when using its standardized ethnic/region/language/religion/phenotype IDs, please cite the following publication:

> Juon, Andreas (2023) Inclusion, Recognition, and Inter-Group Comparisons: The Effects of Power-Sharing Institutions on Grievances. *Journal of Conflict Resolution* (online first, January). https://doi.org/10.1177/00220027231153583.

Please also make sure to cite all underlying original survey data used (see section 2).

1. Introduction

1.1. Purpose of SEAMS

SEAMS has two main purposes:

- First, it provides **standardized cross-national information for major public opinion concepts**, based on comparable survey items from numerous global and regional mass surveys (cf. figure 1). Thereby, SEAMS enables researchers to study cross-national relationships between citizens' socio-political attitudes and time-varying outcomes at the country-level. For instance, SEAMS (v1.0) assembles information on respondents' satisfaction with and trust in governing institutions from 90 unique survey waves, which together cover a total of 1284 country years. This enables researchers to examine how time-variant country-level characteristics, such as economic growth or political institutions, affect citizens' satisfaction with the government. Vice-versa, it also enables researchers to test the consequences of citizens' *aggregate* attitudes on country-level outcomes, for instance how citizens' average satisfaction with the government affects the stability of political institutions or the risk of major social conflict.
- Second, SEAMS also provides **systematic information on respondents' ethnicity, region of residence, language, religion, and phenotype**. This information is connected to major existing datasets, such as the Ethnic Power Relations Dataset (EPR, Vogt et al. 2015), the Constitutional Power-Sharing Dataset (CPSD, Juon 2020), EPR-Ethnic Dimensions (EPR-ED, Vogt et al. 2015), and to standardized lists of administrative regions from the Federal Information Processing Standard (FIPS) PUB 10-4. Thereby, SEAMS enables researchers to conduct even more fine-grained, *sub*national analyses. For instance, they might examine how the representation of ethnic minority representatives, religious and linguistic rights, or regional autonomy shapes affected citizens' satisfaction with the government (cf. Juon 2023 for an example).

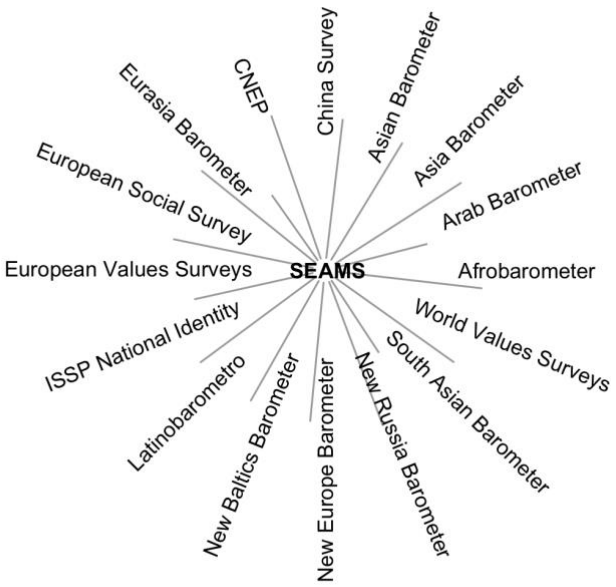

Figure 1. SEAMS: All integrated surveys.

Researchers can connect SEAMS to their data in two main ways. First, if their interest is in SEAMS' standardized attitudinal variables, they can merge additional country- or group-level information to SEAMS using each respondent's **country, ethnic, region, language, religion, and phenotype IDs**. Conversely, if their interest is in identifying the ethnicity of respondents in a given survey (wave), they can merge SEAMS to the original survey data of interest using SEAMS' **respondent ID variable**.[1]

The current version of SEAMS integrates information from **98 unique survey waves**, which together cover **2'071'315 respondents nested in 1372 country years and 148 countries**. It contains information on respondents' (dis-)satisfaction with the government, perceptions of belonging to a discriminated group, generalized trust, and political interest. SEAMS also contains demographic variables, which currently cover information on respondents' gender, age, education level, and urban areas of residence.

Future releases are planned, which aim to expand SEAMS in three ways. First, they will include **additional standardized variables**, including information on respondents' party choice, connected to the V-Parties dataset (Lindberg et al. 2022), respondents' left-right orientation, and their degree of ethnic identification. Second, **more surveys** will be added, including the most recent waves of all survey projects presently covered by SEAMS. Finally, future releases will provide **more fine-grained information on the heterogeneous original question items** used to construct SEAMS' standardized variables, which can be used to construct more explicit measurement models of these attitudes (cf. Claassen 2019).

1.2. Included files

- **seams_v1.0.csv:** Main SEAMS data file which contains all individual-level SEAMS variables (see section 3).
- **seams_sources.csv:** Overview on integrated survey waves and sources (cf. section 2), including names of all original survey files used.
- **seams_ethnic_attribution_wout_region**: Calculated demographic probabilities of respondents belonging to a given ethnic group, depending on respondents' country, language, religion, and phenotype (see section 5 for details). NA's (missing information on respondent attributes) are coded as "-999".
- **seams_ethnic_attribution_with_region**: Calculated demographic probabilities of respondents belonging to a given ethnic group, depending on respondents' country, region of residence, language, religion, and phenotype (see section 5 for details). NA's (missing information on respondent attributes) are coded as "-999".
- **seams_region_codes**: Yearly list of all administrative and survey regions, alongside their FIPS/GEC code (see section 5).

[1] Some survey projects did not provide respondent IDs. For these surveys, it is possible to use the "rownumber" variable, which indicates each respondent's position in the original survey file used (see file "seams_sources.csv" for names and sources of all source files).

2. Coverage

SEAMS aims to comprehensively cover two types of surveys projects:

- First, **large global mass surveys**, including the World Values Surveys (Inglehart et al. 2014), the European Values Survey (EVS 2020), the International Social Survey Programme (ISSP) National Identity modules (ISSP Research Group 2010-2015), augmented by its spin-off, the China Survey (Harmel & Yeh 2015),[2] and the Comparative National Elections Project (CNEP).[3]
- Second, a series of **mass surveys covering specific world regions**, including *Africa* (the Afrobarometer (cf. Afrobarometer Data 1999-2016)), the *Middle East and North Africa* (the Arab Barometer),[4] *Asia* (the Asia Barometer (cf. Inoguchi & Fuji 2008), Asian Barometer, South Asia Barometer,[5] and Eurasia Barometer),[6] *Europe* (the European Social Survey (cf. Norwegian Centre for Research Data, Norway 2002-2018), the new Baltics Barometer (cf. Rose 2010a), the New Europe Barometer (cf. Rose 2010b), and the New Russia Barometer (cf. Rose 2010c)), and *Latin America* (the Latinobarometro).[7]

Table 1 gives an overview on all integrated surveys waves, including the number of countries, years, and total number of respondents contained in them. For detailed information on the sources and specific files used, see file "seams_sources.csv".

Table 1. SEAMS coverage: Surveys, countries, and respondents.

survey	waves	years	countries	respondents
Afrobarometer	1-6	1999-2015	37	206850
Arab Barometer	1-5	2007-2018	13	63345
Asia Barometer	1-5	2003-2007	28	45094
Asian Barometer	1-4	2001-2015	13	60135
China Survey	1	2008-2008	1	3989
Comparative National Elections Project	1	1992-2018	21	70456
Eurasia Barometer	1	2010-2010	9	18000
European Social Survey	1-8	2002-2017	36	393495
European Values Survey	1-4	1981-2009	47	164997
ISSP National Identity	1-3	1995-2013	44	122358
Latinobarometro	1-21	1995-2018	19	431148
New Baltics Barometer	1-6	1993-2004	3	21601
New Europe Barometer	1-7	1991-2004	17	76492
New Russia Barometer	1-18	1992-2009	1	34071
South Asian Barometer	1-2	2005-2013	5	19059
World Values Survey	1-6	1981-2014	95	340225

[2] The China Survey is a project of the College of Liberal Arts at Texas A\&M University, in collaboration with the Research Center for Contemporary China (RCCC) at Peking University.

[3] Available at: <https://u.osu.edu/cnep/surveys/surveys-through-2012/> (accessed on: 17.5.2020).

[4] Available at: <https://www.arabbarometer.org/> (accessed on: 17.5.2020).

[5] Both available at: <http://asianbarometer.org/> (accessed on: 17.5.2020).

[6] Through the Global Barometer Surveys Project, available at: <https://www.globalbarometer.net/>, cf. <http://office.eurasiabarometer.org/> (both accessed on: 17.5.2020).

[7] Available at: <http://www.latinobarometro.org/latContents.jsp> (accessed on: 17.5.2020).

In total, SEAMS covers **148 countries with, on average, 9.7 unique survey waves for each country**. Apart from a small number of states in Africa and in the Middle East, no large countries are systematically missing. The number of unique survey waves per country ranges from 1 for several states in Central Asia up to 35 for Russia. Figure 2 gives a geographic overview over the number of survey waves per country contained in SEAMS.

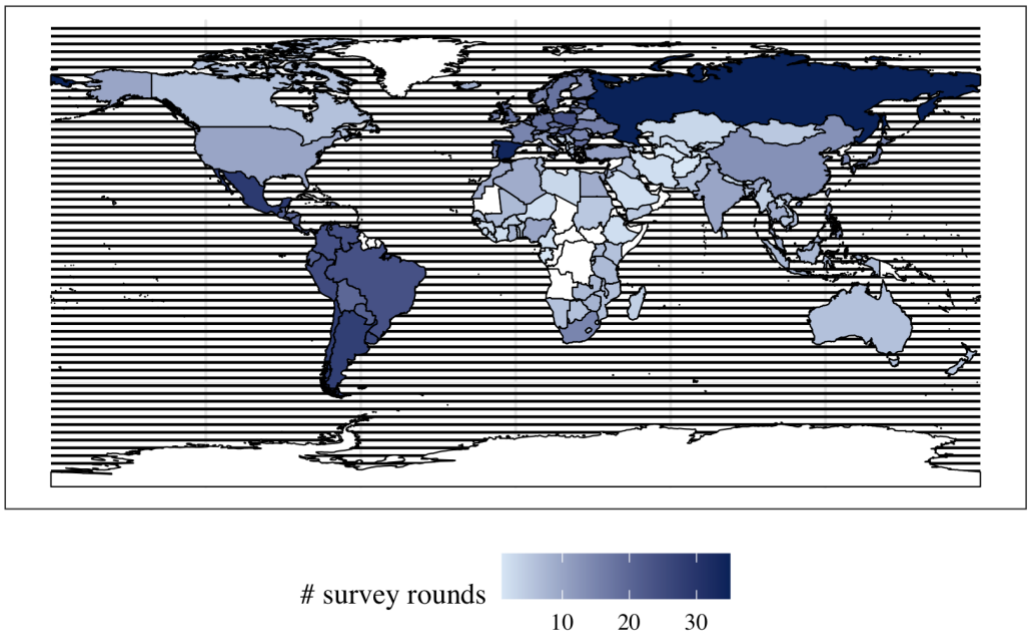

Figure 2. SEAMS coverage: Map of covered countries and number of integrated survey rounds per country.

3. Variable list

Table 2 lists all variables included in the current version of SEAMS. *Identifying* variables enable researchers to identify the source survey (wave), along with each respondent's original ID and their ethno-cultural identity (see section 5 for details). *Standardized variables on political and social attitudes* cover major public opinion concepts of interest, such as (dis-)satisfaction with the government and generalized trust (see section 4 for details). Finally, *standardized variables on demographic characteristics* include variables typically used as controls, such as age and education level of each respondent.

Table 2. List of variables contained in SEAMS.

variable	type	range	description
1. Identifying variables			
survey	character	NA	Survey name (abbreviated).
round	integer	{1, 21}	Survey round.
id_resp	integer	NA	Respondent ID in original survey file, where available.
rownumber	integer	{1, 341271}	Row number in original survey file (for merging data to survey waves where no respondent ID was provided).
gwid	integer	NA	Gleditsch/Ward country code.
country	character	NA	Name of the country.
year	numeric	{1981, 2018}	Year.
month	numeric	{1, 12}	Month.
group	character	NA	Name of the attributed EPR group.
gwgroupid	character	NA	ID of the attributed EPR group (based on all available information, including region, language, religion, and phenotype (see section 5.2 for details).
gwgroupid_self	character	NA	ID of the attributed EPR group, exclusively based on manually-coded, direct information from self-identification question items only.
gwgroupid_presumed	character	NA	ID of the attributed EPR group, exclusively based on manually-coded, direct information from other question items only.
probability	numeric	{0,1}	Demographic probability of correct ethnic identification of *gwgroupid*, considering ethnic segments identified by EPR-ED as exhaustive and ignoring unknown segments (see section 5.2 for details).
probability_strict	numeric	{0,1}	"Strict" demographic probability of correct ethnic identification of *gwgroupid*, considering segments not identified by EPR-ED (see section 5.2 for details).
groupsum	integer	NA	Number of respondents in group and unique survey round.
fips	character	NA	FIPS/GEC code of administrative unit; in case of custom survey region a unique code in the format "[Country abbreviation]_Sxy_xyz"; "-999" if information not available.
rump	integer	{0,1]	1 if the region is a "rump" region, to the exclusion of subordinate administrative areas (e.g., Yugoslavia's SR Serbia without Kosovo), and 0 otherwise.
religion	character	NA	EPR-ED code for respondent's religion.
language	character	NA	Ethnologue code for respondent's language, wherever possible harmonized with EPR-ED.
phenotype	character	NA	EPR-ED code for respondent's phenotype.

Table 2. List of variables contained in SEAMS.

variable	type	range	description
2. Standardized variables on political attitudes			
unsat_gov	integer	{0,1}	Dichotomous variable, taking the value 1 if the respondent is dissatisfied or does not trust the government. Don't know's are coded as 0.
d_unsat_gov	ordinal	{-2,2}	Factor variable capturing original response levels of question items measuring respondent's degree of satisfaction and trust in the government. Positive values indicate more trust/satisfaction, negative values indicate less trust/satisfaction. Don't know's take value 0. Levels depend on the question item of each survey.
disc_grp	integer	{0,1}	Dichotomous variable taking the value 1 if respondent states their group is discriminated against, and 0 otherwise. Don't know's are coded as 0.
d_disc_grp	ordinal	{-2,2}	Factor variable capturing original response levels of question items measuring respondent's perception of discrimination. Positive values indicate a lower perception of discrimination, negative values indicate higher perceptions of discrimination. Don't know's take value 0. Levels depend on the question item of each survey.
3. Standardized variables on social attitudes			
trust_ppl	integer	{0,1}	Dichotomous variable taking the value 1 if respondent generally trusts other people, and 0 otherwise. Don't know's are coded as 0.
d_trust_ppl	ordinal	{-2,2}	Factor variable capturing original response levels of question items measuring respondent's generalized trust. Positive values indicate a stronger trust, negative values indicate lower trust. Don't know's take value 0. Levels depend on the question item of each survey.
interest_pol	integer	{0,1}	Dichotomous variable taking the value 1 if respondent is interested in politics, and 0 otherwise. Don't know's are coded as 0.
d_ interest_pol	ordinal	{-2,2}	Factor variable capturing original response levels of question items measuring respondent's political interest. Positive values indicate a stronger interest, negative values indicate lower interest. Don't know's take value 0. Levels depend on the question item of each survey.
4. Standardized variables on demographic characteristics			
gender	integer	{0,1}	1 if respondent is identified/identifies as female, 0 if respondent is identified/identifies as male.
age	integer	{0,150}	Age of respondent, in most cases calculated by subtracting respondent's year of birth from year of survey administration.
educ	ordinal	{0,3}	Education level of respondent. 0 = no formal education or incomplete primary; 1 = complete primary or compulsory education; 2 = complete secondary education; 3 = complete tertiary education.
urban	integer	{0,1}	1 if respondent lives in an urban area, 0 otherwise.

4. Standardized variables on socio-political attitudes: question items

The first purpose of SEAMS is to provide **standardized information on major public opinion concepts** (see section 2). For this purpose, comparable question items in its underlying survey sources are identified, standardized, and recoded as binary variables (see section 3 for a full list).

While the aim is to identify question items that are as similar as possible, in many cases the **question wordings and their answer categories differ** from survey to survey, in some cases substantially so (cf. figure 3, which depicts the most frequent words included in question items and answer categories underlying the SEAMS variable for *government dissatisfaction*). Researchers may hence want to subset the data and exclude survey waves whose question wordings are not suitable for their purpose. For instance, when using SEAMS' variable on government dissatisfaction, researchers may want to exclude question items that ask for respondents' trust in the government (as opposed to satisfaction with the government).

To facilitate this task, tables 3 to 6 provide the **wording and answer categories for all underlying question items** used to code SEAMS' standardized variables for government dissatisfaction, perceptions of discrimination, political interest, and generalized trust. In all cases, the underlined answer categories are coded as 1, while other answer categories and Don't Know's are coded as 0.[8]

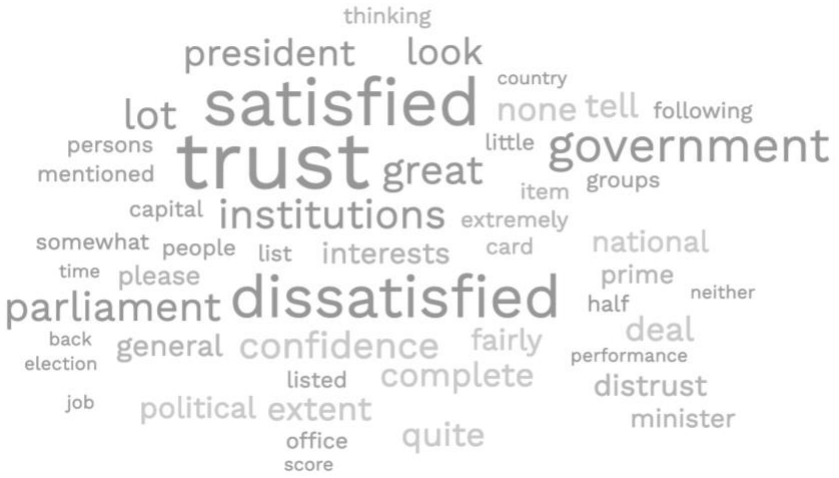

Figure 3. Word cloud for question items and answer categories underlying standardized variable for government dissatisfaction.

[8] Don't know answers can be recoded or excluded from the analysis using the more fine-grained ordinal variables which are provided for each of these concepts (e.g., when analyzing government (dis-)satisfaction with *unsat_gov*, researchers can exclude don't knows by excluding respondents whose *d_unsat_gov* variable takes the value 0).

4.1. Political attitudes

4.1.1. Government dissatisfaction

Table 3. Survey items: Government dissatisfaction (*unsat_gov*).

Survey	Wave	Question	Categories
Afro Barometer	1	Since the last election, how satisfied have you been with the performance of the President [of country] / your member of parliament?	Very unsatisfied / Somewhat unsatisfied / Somewhat satisfied / Very satisfied
Afro Barometer	2-6	How much do you trust each of the following, or haven't you heard enough about them to say: The President / Parliament?	Not at all / A little bit / A lot / A very great deal
Arab Barometer	1-5	Using a 10 point scale, how satisfied are you with the performance of the current government?	1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10
Arab Barometer	1-5	How much trust do you have in the PM or President / Parliament?	A great deal / quite a lot / not very much / none at all
Asia Barometer	1-5	Please indicate to what extent you trust the following institutions to operate in the best interests of society. The central government / Parliament	Don't trust at all / Don't really trust / Trust to a degree / Trust a lot
Asian Barometer	1-4	How satisfied or dissatisfied are you with the current government?	Very dissatisfied / Somewhat dissatisfied / Half and Half / Somewhat satisfied / Very satisfied
Asian Barometer	1-4	How much trust do you have in the national government? / in parliament?	None at all / Not very much / Quite a lot / A great deal
Asian Barometer	3-4	How much trust do you have in the PM / President?	None at all / Not very much / Quite a lot / A great deal
China Survey	1	Please tell us how satisfied or unsatisfied you are with each of the following: Central government	0 = not satisfied at all / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 = satisfied very much
Comparative National Elections Project	IV	Thinking back to the time of the election, to what extent were you satisfied or dissatisfied with the performance of the President/Prime Minister?	Very dissatisfied / dissatisfied / Neither satisfied nor dissatisfied / Very satisfied
European Social Survey	1-8	Now thinking about the [country] government, how satisfied are you with the way it is doing its job?	0 = extremely dissatisfied / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 = extremely satisfied
European Social Survey	1-8	Please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. Parliament	0 = no trust at all / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 = complete trust
Eurasia Barometer	1	How much trust do you have in the President / Government / Parliament?	None at all / Not very much / Quite a lot / A great deal
European Values Survey	1-4/4	Please look at this card and tell me, for each item listed, how much confidence you have in them. The government [in your capital] (1-4) / Parliament (4)	None at all / Not very much / Quite a lot / A great deal
European Values Survey	3	How satisfied are you with people in the national office?	Very dissatisfied / fairly dissatisfied / fairly satisfied / very satisfied
Latino Barometer	1,2,7-21	Tell me how much confidence you have in each of the following groups, institutions or persons mentioned on the list: the Government	No confidence / a little / some / a lot
Latino Barometer	1-21	Tell me how much confidence you have in each of the following groups, institutions or persons mentioned on the list: the National Congress	No confidence / a little / some / a lot
New Europe Barometer	2-7	To what extent do you trust each of these political institutions to look after your interests? Government (2- 5) / President (3-7) / Prime Minister (5-6) / Parliament (2-7)	No trust at all / Great trust

Table 3. Survey items: Government dissatisfaction (*unsat_gov*).

Survey	Wave	Question	Categories
New Baltic Barometer	1, 3	To what extent do you trust each of these political institutions to look after your interests? Cabinet of ministers (3) / President (1, 3) / Parliament (1, 3)	Complete distrust / general distrust / general trust / complete trust
New Baltic Barometer	5, 6	To what extent do you trust each of these political institutions to look after your interests? Prime Minister	No trust at all / Great trust
New Russia Barometer	1, 3, 5-18	To what extent do you trust each of these political institutions to look after your interests? Government (3) / President (2, 3, 5-12, 14, 15) / Supreme Soviet (2) / Duma (3, 5-18)	Complete distrust / general distrust / general trust / complete trust
South Asia Barometer	1-2	How much trust do you have in the national government? / in parliament?	None at all / Not very much / Quite a lot / A great deal
South Asia Barometer	2	How satisfied or dissatisfied are you with the current government?	Very dissatisfied / Somewhat dissatisfied / Half and Half / Somewhat satisfied / Very satisfied
South Asia Barometer	2	How much trust do you have in the President? / in the PM?	None at all / Not very much / Quite a lot / A great deal
World Values Survey	1/3-6	Please look at this card and tell me, for each item listed, how much confidence you have in them. The government [in your capital] (3-6) / Parliament (1-6)	None at all / Not very much / Quite a lot / A great deal
World Values Survey	3, 4	How satisfied are you with people in the national office?	Very dissatisfied / fairly dissatisfied / fairly satisfied / very satisfied

4.1.2. Perceptions of discrimination

Table 4. Survey items: Feeling discriminated (*disc_grp*).

Survey	Wave	Question	Categories
Afrobarometer	1-6	How often, if ever, are [respondent's ethnic group] treated unfairly by the government?	Never / Sometimes / Often / Always
Asian Barometer	3-4	All citizens from different ethnic communities in [country] are treated equally by the government.	Strongly agree / Somewhat agree / Somewhat disagree / Strongly disagree
European Social Survey	1-8	Would you describe yourself as being a member of a group that is discriminated against in this country?	Yes / No
ISSP-N	1-3	How proud are you of [country's] fair and equal treatment of all groups in society?	Not proud at all / Not very proud / Somewhat proud / Very proud
Latinobarometro	14, 15, 16, 18	Would you describe yourself as being a member of a group that is discriminated against in this country?	Yes / No
New Europe Barometer	5	Non-citizens and minority nationalities are being badly treated here.	Strongly agree / Agree / Disagree / Strongly disagree
New Baltics Barometer	1	Non-citizens and minority nationalities are being badly treated here.	Strongly agree / Agree / Disagree / Strongly disagree

4.2. Social attitudes

4.2.1. Political interest

Table 5. Survey items: Political interest (*interest_pol*).

Survey	Wave	Question	Categories
Afrobarometer	1-6	How interested are you in politics and government/How interested would you say you are in public affairs?	Not interested/hardly / Now and then / Somewhat interested/some of the time / Very interested/most of the time
Arab Barometer	1-5	Generally speaking, how interested would you say you are in politics? / In general, to what extent are you interested in politics?	Not interested/Very uninterested / Little interested/Uninterested / Interested / Very interested
Asian Barometer	1-4	How interested would you say you are in politics?	Not at all interested / Not very interested / Somewhat interested / Very interested
The China Survey	1	How interested would you say you are in political matters?	Not at all interested / Not very interested / Somewhat interested / Very interested
Comparative National Elections Project	1	Would you say that you are very, somewhat, not very or not at all interested in politics?	Not at all interested / Not very interested / Somewhat interested / Very interested
European Social Survey	1-8	How interested would you say you are in politics – are you…	Not at all interested / Hardly interested / Quite interested / Very interested
European Values Survey	2-4	How interested would you say you are in politics?	Not at all interested / Not very interested / Somewhat interested / Very interested
Latinobarometro	1-6, 8-10, 12, 14, 15, 17	How interested are you in politics?	Not at all interested / A little interested / Fairly interested / Very interested
New Baltics Barometer	1-3, 6	How interested would you say you are in politics?	Not at all interested / A little interested / Somewhat interested / Very interested
New Europe Barometer	2, 7	How interested would you say you are in politics?	Not at all interested / A little interested / Somewhat interested / Very interested
New Russia Barometer	1, 4, 5, 11, 14-18	How interested would you say you are in politics?	Not interested / Little interested / Interested / Very interested
South Asia Barometer	1	How frequently do you discuss politics with your friends and colleagues, often, occasionally, or never?	Never / Occasionally / Often
South Asia Barometer	2	How interested would you say you are in politics?	Not at all interested / Not very interested / Somewhat interested / Very interested
World Values Survey	1-6	How interested would you say you are in politics? Are you	Not at all interested / Not very interested / Somewhat interested / Very interested

4.2.2. Generalized trust

Table 6. Survey items: Generalized trust (*trust_ppl*).

Survey	Wave	Question	Categories
Afrobarometer	1, 3-5	Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	Can't be too careful / Most people can be trusted
Arab Barometer	1-5	Generally speaking, would you say that most people can be trusted?	You must be very careful in dealing with people / Most people can be trusted
Asia Barometer	1-5	Generally, do you think people can be trusted or do you think that you can't be too careful in dealing with people (that it pays to be wary of people)?	Can't be too careful in dealing with people / Most people can be trusted
Asian Barometer	1-4	General speaking, would you say that "Most people can be trusted" or "you can't be too careful in dealing with them"?	One can't be too careful in dealing with them/people / Most people can be trusted
The China Survey	1	Generally speaking, would you say that most people can be trusted, or you can't be too careful in dealing with them?	You can't be too careful / Most people can be trusted
European Social Survey	1-8	Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful3 in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.	0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10
Eurasia Barometer	1	What is the degree to which you agree with the statement that a majority of people can be trusted?	You must be very careful in dealing with people / Most people can be trusted
European Values Survey	1-4	Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	Can't be too careful / Most people can be trusted
Latinobarometro	2-21	Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	One can never be too careful when dealing with others / You can trust most people
New Baltics Barometer	2-3	People often hold different opinions about relations with others; for each pair of statements, please say which is closer to your own way of thinking: Most people can be trusted OR You cannot be too careful in dealing with people.	You can't be too careful in dealing with people / Most people can be trusted
New Europe Barometer	5-7	There are many different institutions in this country, for example, government, courts, police, civil servants. Please show me on this scale how great is your personal trust in each of these institutions. "1" represents no trust and "7" represents great trust. Most people in this country / you know	1 / 2 / 3 / 4 / 5 / 6 / 7
New Russia Barometer	8, 9, 11, 14, 15, 18	There are many different institutions in this country, for example, government, courts, police, civil servants. Please show me on this scale how great is your personal trust in each of these institutions. "1" represents no trust and "7" represents great trust. Most people you meet	1 / 2 / 3 / 4 / 5 / 6 / 7
World Values Survey	1-6	Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?	Need to be very careful / Most people can be trusted

5. Standardized codes for survey respondents' ethno-cultural identities

The second purpose of SEAMS is to provide **systematic information on respondents' ethnicity, region of residence, spoken language, practiced religion, and phenotype**.[9] To facilitate future analyses, this information is captured by standardized codes, which link to existing datasets. SEAMS' ethnic identity variables (*gwgroupid_self*, *gwgroupid_presumed*, and *gwgroupid*) link to the Ethnic Power Relations dataset (EPR, Vogt et al. 2015) and to the Constitutional Power-Sharing Dataset (CPSD, Juon 2020), its *language* variable links to Ethnologue, and its *religion* and *phenotype* variables link to the EPR-Ethnic Dimensions dataset (EPR-ED, Vogt et al. 2015).

SEAMS provides two types of linking variables. First, **direct linking variables** to ethnicity (*gwgroupid_self*, *gwgroupid_presumed*), region (*fips*, *rump*), *language*, *religion*, and *phenotype* are manually coded based on the available question items in the underlying survey waves. Second, a **more exhaustive linking variable is coded in a semi-automatic manner for ethnicity** (*gwgroupid*). This takes into account *all* available information, including (indirect) information on respondents' ethnicity via their observed region of residence, language, religion, and phenotype. Figure 4 schematically visualizes these links.

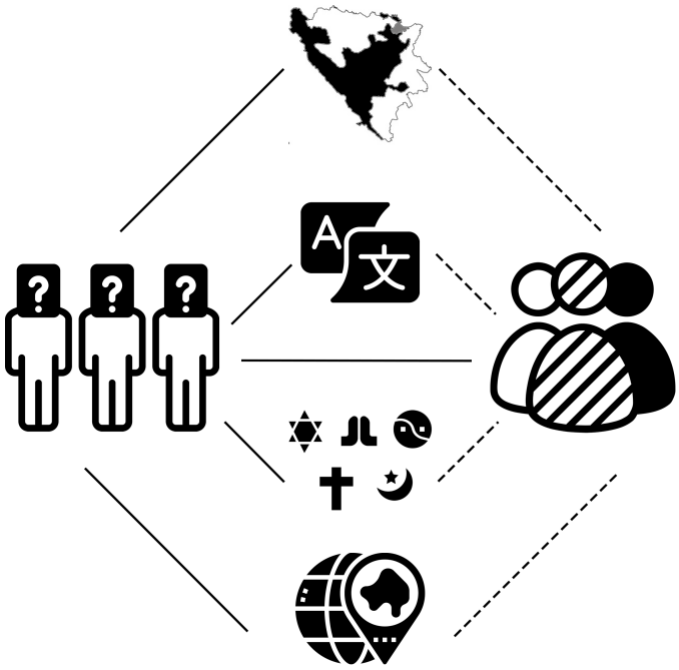

Figure 4. Direct links of SEAMS survey respondents to ethnicity, region, language, religion, and phenotype (solid lines) and indirect links to ethnicity via region, language, religion, and phenotype (dotted lines).

[9] In line with EPR-ED (Vogt et al. 2015), phenotype is understood as a measure of ethnic groups' (perceived) origins from particular world regions, such as Europe, Sub-Saharan Africa, Oceania, etc.

5.1. Manual coding of direct links (respondent ethnicity, region, language, religion, and phenotype)

In the original surveys integrated into SEAMS, all available variables pertaining to respondents' ethno-cultural identities and region of residence were identified to manually code **direct links**. These variables are either question items that directly ask respondents for their self-identification (e.g., ethnicity, religion practiced) or variables that were coded by the survey administrators themselves (e.g., sampled administrative region). Based on these variables and on manual coding, SEAMS provides the following, standardized identity variables:

- *gwgroupid_self*: **ethnic group** ID that links to EPR and CPSD (Vogt et al. 2015, Juon 2020), based exclusively on question items that ask respondents to self-identify with an ethnic group;
- *gwgroupid_presumed*: **ethnic group** ID that links to EPR and CPSD (Vogt et al. 2015, Juon 2020), based exclusively on question items that ask for respondents' "objective" cultural characteristics;
- *fips*: FIPS/GEC code of the **administrative region, statistical region, or survey-specific geographic zone** in which the respondent was resident or was surveyed (see file "seams_region_codes.csv" for full list of all region codes, region types, and names);
- *rump*: dummy variable that designates **"rump" regions**, which do not contain certain subordinate regions, as these are sampled separately in the survey (e.g. Yugoslavia's SR Serbia without Kosovo);
- *language*: Ethnologue code for respondent's **language**, wherever possible harmonized with EPR-ED;
- *religion*: EPR-ED (Vogt et al. 2015) code for respondent's **religion**;
- *phenotype*: EPR-ED (Vogt et al. 2015) code for respondent's **phenotype**.

5.2. Semi-automatic ethnic attribution, based on direct and indirect links

5.2.1. Rationale

Based on the combined information of these direct linking variables, SEAMS also provides a **more extensive coding of respondents' ethnic identity** (Vogt et al. 2015; Juon 2020). This is based not only on respondents' self-identification (*gwgroupid_self* / *gwgroupid_presumed*), but takes into account *all* available information, including respondents' region of residence, language, religion, and phenotype, as coded above. This more extensive measure is helpful for cross-national survey research that is interested in the political and social attitudes of ethnic group members, as explicit questions on respondent ethnicity are available only for a small subset of surveys (e.g., the Afrobarometer).

Most existing studies facing a similar challenge (e.g., Bühlmann & Hänni 2012; Elkins & Sides 2007; Hänni 2017; Robinson 2014) similarly combine question items on ethnicity, region, language, religion, and phenotype, and use these to manually attribute respondents to their ethnic groups. For SEAMS, the combination of a large number of mass survey sources means that such a manual attribution procedure might not only be unfeasible, but might, due to different information provided across survey types, also result in inconsistent classifications.

Instead, SEAMS conducts a **semi-automatic ethnic attribution procedure** that combines explicit self-identification question items asked in some of the surveys (most consistently in the Afrobarometer) with all available information on respondents' resident region, religion, language, and phenotype provided in other surveys (see above). These were compared to the corresponding demographic data from the EPR-Ethnic Dimensions dataset (Vogt et al. 2015) and regional settlement patterns derived from the Geo-EPR

dataset (Vogt et al. 2015) to match respondents and ethnic groups.[10] In a nutshell, respondents were attributed to the ethnic group that offers the closest demographic match to their observed characteristics (e.g., spoken language or practiced religion).

5.2.2. Semi-automatic attribution: step-wise approach

In a first step, the information provided by EPR and CPSD was used to calculate the **population shares of each group's within-group cleavages**. EPR-ED provides information on three cleavages: the languages spoken by group members, the religions practiced by them, and their phenotype. It codes both type (e.g., which languages do group members speak?) and relative population shares of up to the three largest sub-segments arising from each cleavage. Using this information, the population shares of each cleavage in each country year were calculated, given by their religion, language, and phenotype.

In a second step, the **demographic proportion of each group settling in each region** used in the different mass surveys was calculated. This was done by spatially intersecting the ethnic settlement patterns provided by the EPR dataset (Vogt et al. 2015) with the territorial boundaries of the regions used in the various surveys.[11]

In a third step, both sources of information were combined to calculate the **population shares of each group's segments (given by religion, language, and phenotype) in each region** used in the surveys.

Using the information assembled in these three steps, the **demographic probability that a respondent is part of a given ethnic group was calculated**, given their available information on the corresponding characteristics.

Two probability measures were calculated. A first measure, *probability*, calculates the demographic probability that a respondent is correctly identified, while ignoring any ethnic segments not identified by EPR and EPR-ED (relevant in cases where a large proportion of the country is not considered as "politically relevant" by EPR or where groups have more than the three cultural cleavages coded by EPR-ED). Conversely, a second variable, *probability_strict*, re-calculates the demographic probability while considering the non-identified segments.

The probability that a respondent *r* with a vector of characteristics *R* belongs to a given group *x* is given by the following formulas (depending on whether it is calculated as *probability* or as *probability_strict*):

$$Probability(r \in x) = \frac{\alpha_{x,territory\ r} \times \beta_{x,religion\ r} \times \gamma_{x,language\ r} \times \delta_{x,phenotype\ r}}{\sum_{x=1}^{X}(\alpha_{x,territory\ r} \times \beta_{x,religion\ r} \times \gamma_{x,language\ r} \times \delta_{x,phenotype\ r})}$$

$$Probability\ strict(r \in x) = \frac{\alpha_{x,territory\ r} \times \beta_{x,religion\ r} \times \gamma_{x,language\ r} \times \delta_{x,phenotype\ r} + \varepsilon_x}{\sum_{x=1}^{X}(\alpha_{x,territory\ r} \times \beta_{x,religion\ r} \times \gamma_{x,language\ r} \times \delta_{x,phenotype\ r} + \varepsilon_x)}$$

[10] Both datasets were augmented to account for the small number of additional ethnic groups contained in the CPSD (see Juon 2023).

[11] For this purpose, I predominantly relied on the polygons for administrative units provided the Database of Global Administrative Areas (GADM), available online under <http://gadm.org/data.html>. Where these units changed over time or where the surveys used other territorial units, I handcoded these through a combination of backwards aggregation of administrative boundaries and manually georeferenced historical or contemporary maps. I weighted the spatial overlaps attained through this procedure with their local population densities, proxied by the 2015 version of the Gridded Population of the World Version 3 database, provided by the Center for International Earth Science Information Network and available at < http://sedac.ciesin.columbia.edu/gpw>.

In other words, it is calculated by dividing the population shares of group *x* with characteristics corresponding to respondent *r* by the total population shares of all groups with the same characteristics. $\alpha_{x,territory\,r}$, $\beta_{x,religion\,r}$, $\gamma_{x,language\,r}$, and $\delta_{x,phenotype\,r}$ are the respective shares of group *x* with characteristic *r*.

For *probability_strict*, ε_x is a factor to account for the demographic shares of group *x*'s **unidentified sub-segments** (i.e., for who information on a characteristic is missing). This is relevant in cases where a substantial proportion of the population is not considered "politically relevant" by EPR or where the population shares of the three characteristics provided by the augmented EPR-ED do not add up to 1. The addition of this factor to both nominator and denominator serves to account for the fact that these unidentified sub-segments might possess these same characteristics *R*, yet remain unidentified.

The supplementary files "seams_ethnic_attribution_wout_region.csv" and "seams_ethnic_attribution_with_region.csv" provide all calculated (strict) probabilities for all countries covered by EPR between 1981 and 2018.

5.2.3. Illustrative example

Figure 2 provides an illustrative example for this attribution procedure. It shows the population shares of all Bosnian ethnic groups in EPR/CPSD and the proportion within each group that possesses a certain characteristic (region, religion, language, phenotype) in 2018. It indicates that some groups are easily distinguishable from others, given certain observed characteristics, while this is more difficult for others.

For example, a respondent speaking the Bosnian language would almost unequivocally be attributed to the Bosniak group (*probability_strict* = [0.5 * 1] / [0.5 * 1 + 0.03]). Similarly, a respondent speaking Serbian would be attributed to the Serb group with high probability (*probability_strict* = [0.31 * 1] / [0.31 * 1 + 0.0005 * 1 + 0.03]).

In other cases, the attribution would be less straightforward: an otherwise unidentified respondent living in the Federacija would only be attributed with a comparably small probability to the Bosniak group (*probability_strict* = [0.5 * 0.95] / [0.5 * 0.95 + 0.31 * 0.04 + 0.15 * 0.99 + 0.01 * 0.56 + 0.0008 * 0.56 + 0.0005 * 0.56 + 0.03 * 0.56]).

5.2.4. Validation

To **validate** the attribution procedure, the relative population shares of the ethnic groups according to EPR/CPSD were compared with those obtained for each country's unique survey wave. The very high correlation of r=0.986 indicates a high agreement between both and is comparable to manual classifications (cf. Bühlmann & Hänni 2012).[12]

[12] Without discarding any group or country survey year entirely (see main article), this correlation is still very high at r=0.890.

		Bosniaks	Serbs	Croats	Roma	Albanians	Montenegrins	Other
		0.5	**0.31**	**0.15**	**0.01**	**0.0008**	**0.0005**	**0.03**
Territory	Federacija	0.95	0.04	0.99	0.56	0.56	0.56	0.56
	Republika	0.03	0.94	0.01	0.43	0.43	0.43	0.43
Religion	Sunni Islam	1	0	0	0.4	0.6	0	N/A
	Serb Orthodox	0	1	0	0	0	1	N/A
	Catholic	0	0	1	0.3	0.1	0	N/A
	Protestant	0	0	0	0.3	0.2	0	N/A
Language	Bosnian	1	0	0	0	0	0	N/A
	Serbian	0	1	0	0	0	1	N/A
	Croat	0	0	1	0	0	0	N/A
	Romany	0	0	0	1	0	0	N/A
	Tosk Albanian	0	0	0	0	0.62	0	N/A
	Gheg Albanian	0	0	0	0	0.38	0	N/A
Phenotype	European	1	1	1	1	1	1	N/A

Figure 2. Attribution of respondents to ethnic groups, example of Bosnia and Herzegovina (2018).

Note: The top row shows the overall demographic shares of each group. All numbers in the other rows are fractions indicating the relative size of each group's subgroups, given territory, religion, language, and phenotype. For example, Bosniaks have an overall demographic share of 50% and 95% of them live in the Federacija.

References

Afrobarometer Data (1999) *Afrobarometer Rounds 1-6* (http://www.afrobarometer.org).

Bühlmann, Marc & Miriam Hänni (2012) Creating a 'We-feeling' in Heterogeneous Nations: A Cross-country Investigation of the Impact of Inclusion and Autonomy on the National Pride of Ethnic Minorities. *Comparative Sociology* 11(3): 325–353.

Claassen, Christopher (2019) Estimating Smooth Country–Year Panels of Public Opinion. *Political Analysis* 27(1): 1–20.

Elkins, Zachary & John Sides (2007) Can institutions build unity in multiethnic states? *American Political Science Review* 101(04): 693–708.

EVS (2020) European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). ZA4804 Datenfile Version 3.1.0. GESIS Datenarchiv, Köln. (https://doi.org/10.4232/1.13486).

Hänni, Miriam (2017) Responsiveness – To Whom? Why the Primacy of the Median Voter Alienates Minorities. *Political Studies* 65(3): 665–684.

Harmel, Robert & Yao-Yuan Yeh (2015) China's Age Cohorts: Differences in Political Attitudes and Behavior*: China's Age Cohorts. *Social Science Quarterly* 96(1): 214–234.

Inglehart, Ronald, C Haerpfer, A Moreno, C Welzel, K Kizilova, J Diez-Medrano, M Lagos, Pippa Norris, E Ponarin & B Puranen (2014) *World Values Survey: All Rounds - Country-Pooled Datafile Version.* Madrid: JD Systems Institute (http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp).

Inoguchi, Takashi & Seiji Fujii (2008) The AsiaBarometer: Its Aim, Its Scope and Its Development. In: Valerie Møller (ed.) *Barometers of Quality of Life around the Globe: How Are We Doing?* s.l.: Springer, 187–232.

ISSP Research Group (2010) *International Social Survey Programme: ISSP - National Identity I-III*. Data file Version 1.0.0. Cologne: GESIS Data Archive.

Juon, Andreas (2020) Minorities overlooked: Group-based power-sharing and the exclusion-amid-inclusion dilemma. *International Political Science Review* 41(1): 89–107.

Juon, Andreas (2023) Inclusion, Recognition, and Inter-Group Comparisons: The Effects of Power-Sharing Institutions on Grievances. *Journal of Conflict Resolution* (January) (http://journals.sagepub.com/doi/10.1177/00220027231153583).

Lindberg, Staffan I, Nils Düpont, Masaaki Higashijima, Yaman Berker Kavasoglu, Kyle L Marquardt, Michael Bernhard, Holger Döring, Allen Hicken, Melis Laebens, Juraj Medzihorsky, Anja Neundorf, Ora John Reuter, Saskia P Ruth-Lovell, Ketih R Weghorst, Nina Wiesehomeier, Joseph Wright, Nazifa Alizada, Paul Bederke, Lisa Gastaldi, Sandra Grahn, Garry Hindle, Nina Ilchenko, Johannes von Römer, Steven Wilson, Daniel Pemstein & Brigitte Seim (2022) *Varieties of Party Identity and Organization (V-Party) Dataset V2.* (https://doi.org/10.23696/vpartydsv2).

Norwegian Centre for Research Data, Norway (2002) ESS Rounds 1-9 (https://www.europeansocialsurvey.org/).

Robinson, Amanda Lea (2014) National Versus Ethnic Identification in Africa: Modernization, Colonial Legacy, and the Origins of territorial Nationalism. *World Politics* 66(4): 709–746.

Rose, R (2010a) *New Baltic Barometer I-VI Trend Dataset, 1993-2004*. Colchester, Essex: UK Data Service (http://doi.org/10.5255/UKDA-SN-6510-1).

Rose, R (2010b) *New Europe Barometer I-XV Trend Dataset, 1991-2007*. 2nd Edition. Colchester, Essex: UK Data Service (http://doi.org/10.5255/UKDA-SN-5241-1).

Rose, R (2010c) *New Russia Barometer I-XVIII Trend Dataset, 1992-2009*. Colchester, Essex: UK Data Service.

Vogt, M, NC Bormann, S Rüegger, LE Cederman, P Hunziker & L Girardin (2015) Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Data Set Family. *Journal of Conflict Resolution* 59(7): 1327–1342.