Supplementary Materials Collection 1 of 2 for "The Composition of Descriptive Representation"*

John Gerring † Connor T. Jerzak‡ Erzen Öncel §

June 13, 2023

Abstract

How well do governments represent the societies they serve? A key aspect of this question concerns the extent to which leaders reflect the demographic features of the population they represent. To address this important issue in a systematic manner, we propose a unified approach for measuring descriptive representation. We apply this approach to newly collected data describing the ethnic, linguistic, religious, and gender identities of over 50,000 leaders serving in 1,552 political bodies across 156 countries. Strikingly, no country represents social groups in rough proportion to their share of the population. To explain this shortfall, we focus on compositional factors—the size of political bodies as well as the number and relative size of social groups. We investigate these factors using a simple model based on random sampling and the original data described above. Our analyses demonstrate that roughly half of the variability in descriptive representation is attributable to compositional factors.

Keywords: Descriptive representation; Leadership; Political institutions; Social groups; Elites

*See APSR Dataverse doi:10.7910 for Supplementary Materials Collection 2 of 2. See github.com/cjerzak/DescriptiveRepresentationCalculator-software for software computing the expected and observed descriptive representation indices described in the paper.

†Professor, Department of Government, University of Texas at Austin. Email: jgerring@austin.utexas.edu URL: liberalarts.utexas.edu/government/faculty/jg29775

‡Assistant Professor, Department of Government, University of Texas at Austin. Email: connor.jerzak@austin.utexas.edu URL: ConnorJerzak.com

§Assistant Professor, Department of International Relations, Özyeğin University. Email: erzen.oncel@ozyegin.edu.tr URL: ozyegin.edu.tr/en/faculty/erzenoncel

Supplemental Materials I: Scope-Conditions

In the main text, we stipulated that compositional effects arise for all traits except those regarded by selectors as unfit for public office. (Recall that *selectors* are those who appoint or elect persons to public office.) This serves as a principal scope-condition for the theory. In this appendix, we take a closer look at the ways in which preferences—sometimes referred to in the literature as culture or ideology (Paxton and Kunovich 2003; Paxton, Hughes, and Barnes 2020; Norris and Inglehart 2004; Ruedin 2013)—moderate compositional effects.

The preferences of selectors are evidently matters of degree. Consider a trait that is evenly distributed across a population, half of whom are Type A and half Type B. Now consider a range of preferences for that trait across the citizenry, who also serve as selectors (equally weighted in the selection process).

If all selectors favor this trait, preferring candidates of Type A, or disfavor this trait, preferring candidates of Type B, there will be no variation in the chosen representatives, all of which will be of Type A or B. Compositional factors cannot come into play, and the rejected type will receive no representation at all. If, on the other hand, selectors prefer their own type (affinity voting), or are indifferent (ignoring the trait in their selection of candidates), we expect compositional factors to come into play.

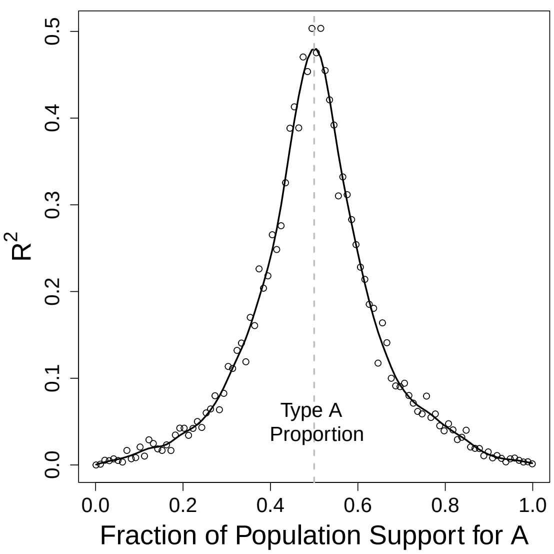

Figure S.I.1: *Here, we see how affinity voting for group A affects the R^2 in predicting observed representation using its expected value under random sample (assuming $R_b = \beta_0 + \beta_1 \cdot \mathbb{E}[R_b] + \epsilon_b$ for body b with β_0 and β_1 coefficients estimated via OLS). In the simulation, body size is randomly varied, generating variation in observed representation to be explained. Type A proportion in the population is fixed at 0.5.*

By enlisting the simulation exercise introduced in Section 1 we can observe these dynamics at work. The y-axis in Figure S.I.1 shows model-fit statistics (R^2) for a regression model where representation scores for political bodies are on the left side and predictions from our random sampling model (which take into account the size of the body and the heterogeneity of the population) are on the right side. The x-axis shows the preferences of selectors for Type A, allowing us to test the impact of varying preferences on compositional effects.

As theorized, R^2 goes to zero when candidates of Type A are unanimously rejected by the selectorate (on the left end of the x-axis) or are unanimously preferred by the selectorate (on the right end of the x-axis). Compositional effects are strongest ($R^2 = 0.5$) in the middle, a situation where each group selects its own type (affinity voting) or where everyone is indifferent (random choice). (Recall that the population is evenly divided between Type A and Type B.)

A crucial caveat concerns the strength of preferences. If weak or indifferent, representation for trait A may be driven by preferences for other traits that are correlated with A. We regard these other traits as confounders insofar as they generate a spurious association between compositional factors and the representation of trait A. For example, if A is correlated with C, and selectors have weak preferences for A and strong preferences for C, then the latter will drive the representation of A and compositional effects for A will be confounded.

S.I.1.1 A Schematic Survey

In reviewing the evidence, let us begin with the simplest example, where a category of persons is legally excluded from holding public office. This described the status of women until recently, and it describes the status of minors, felons, and immigrants in many polities today. Many countries also impose religious qualifications, and at least one (Thailand) imposes explicit educational qualifications.

These exclusions presumably represent the preferences of a restricted selectorate, not the entire population (at least some members of the excluded categories presumably disapprove of their exclusion). In any case, it offers a limiting case of how preferences (among the selectorate) affect the scope-conditions of our theory. Evidently, the size of a political body or the distribution of traits across a population has no bearing on representation where persons with specified traits are formally excluded.

In other instances, where exclusions are informal, the preferences of selectors may nonetheless work against the representation of a group. For example, persons who fall into categories judged as deviant are unlikely to be selected for public office, and may even be discouraged from putting their names forward, thus affecting the availability of candidates.

A third category comprises identities of uncertain political import such as social class and education. Here, extant studies suggest that preferences are mixed. Some studies suggest that selectors may prefer candidates with less education or from working-class backgrounds, all other things equal (Carnes and Lupu 2016, 2023; Gift and Lastra-Anadón 2018; Campbell and Cowley 2014). Other studies suggest there is a preference for more educated candidates (Hainmueller, Hopkins, and Yamamoto 2014; Arceneaux and Vander Wielen 2023). A recent meta-analysis based on conjoint experiments conducted in eleven countries, shows a fairly strong preference for more educated candidates, a preference apparently shared across highly educated and less educated respondents (Simon and Turnbull-Dugarte 2023). In any case, we do not know how salient education-based preferences are, and they may be outweighed by other considerations. Since class and education are correlated with features seen as desirable such as experience, connections, and clout, this may explain why less-educated and working-class representatives are scarce.

An example of this phenomenon can be found in populist leaders such as Donald Trump, who appeals to less-educated American whites but is, himself, highly educated. This suggests that Trump's other attributes outweigh whatever reservations his supporters might have about his plutocratic, Ivy League biography. (One might even speculate that some of them find his exalted status and power attractive.)

Before quitting this discussion, we should point out that it is difficult to determine voter preferences about an attribute like education because the treatment of interest is ambiguous. Survey experiments can manipulate the status of a candidate by describing them as college-educated or high-school-educated. However, the treatment of interest is, arguably, not whether an individual has a title ("BA") next to their name but rather whether educating a person makes them a more attractive candidate. Insofar as higher education shapes students it might make them more compelling candidates after they graduate. The analogous experiment or natural experiment would require comparing the overall attractiveness of candidates who have, and have not, attended university but who were, ex ante, similar. We are not aware of any study of this sort. Evidently, randomizing a title ("BA") is easier than randomizing education.

Suffice it to say, we are in the dark about preferences pertaining to traits like education and social class. However, available evidence suggests that they are not as strong as preferences for ethnicity and gender, and thus may be overridden by other factors.

Now, let us turn to traits that are, by all appearances, *not* chosen in an intentional fashion. For these characteristics most selectors have no preferences; they are indifferent, or nearly so. One might dismiss such traits as irrelevant to politics, and it is surely true that they have generated less interest on the part of academics and the popular press. However, Burden (2007) shows that lawmakers' life experiences affect their views on a vast range of policies that come before them. These intrinsic features presumably also generate spillover effects for other lawmakers. When policies concerning health care arise, legislatures often turn to physicians in their midst with the understanding that they possess unique insights into this policy area. When policies concerning disability rights arise, legislatures may turn to members with special needs for advice. When policies arise that bear differentially upon different age groups, or upon which young and old people have different views, legislatures may turn to members in those age groups for advice. And so forth. Leaders with special traits that distinguish them from other lawmakers may even become national leaders in these policy areas. As such, the incidental traits that leaders bring to their job constitute an important, though perhaps underappreciated, form of descriptive representation.

Since the vast majority of traits exhibited by officeholders are low salience, it follows that compositional factors may be considerably more important than is apparent from an examination of the usual (high-salience) dimensions. However, compositional effects will be strong only if a trait is uncorrelated with valued traits.

Consider *blood type*, a trait that is distributed across a population in a fairly (though not entirely) random fashion and is not regarded as relevant for officeholding. Here, compositional effects should apply.

Now, consider *age* as a feature of representation (Stockemer and Sundström 2018). Since age is generally fairly low-salience, we might expect selection into office to be incidental. However, because age is correlated with other valued traits such as experience, it seems likely that compositional effects for age groups will be muted.

S.I.1.2 Changes Through Time

It follows from what we have said so far that as preferences change, so will compositional effects. As an example, let us consider the interplay between the intentional and incidental features of selection as they appear to have played out in the context of male sexual orientation.[1]

A half-century ago, few men were open about their sexuality if it did not conform to the heterosexual norm. Accordingly, one may suppose that homosexuality was an incidental feature of selection into political bodies, something few were aware of or (if aware) cared to consider. It was irrelevant because suppressed. Accordingly, we can expect strong compositional effects due to random selection. (Of course, one may question whether representation is achieved if the trait in question is entirely sub rosa; but for the moment, let us consider the matter from a purely arithmetic perspective.)

By the late twentieth century, many gay individuals were coming out (voluntarily) or were outed (involuntarily). A latent trait thus became manifest and was highly salient. Because homosexuality

[1] Our account builds on various historical narratives including Adam (2009) and Ayoub (2016).

was viewed as deviant, gay candidates for public office were discriminated against. In this setting, compositional factors became irrelevant; gay individuals were judged ineligible for public office, and few bothered to put themselves forward.

By the twenty-first century, public opinion had shifted dramatically. In many societies, homosexuality was now understood as acceptable. At this point in time, compositional factors should come into play, approximating those we have measured for other generally acceptable traits such as gender, religion, language, and ethnicity.

A final stage in this evolution may be envisioned at some point in the future. At this point, let us conjecture that sexual orientation becomes irrelevant to political representation—a background factor like blood type that carries little political relevance. If so, compositional effects will still exist but for a different reason—because sexuality is a low-salience background characteristic, chosen at random from the population, just as it was in previous eras when the trait was hidden from view.

For our purposes, what is important in this saga is that compositional effects apply to the representation of sexuality at every stage of the journey except the second stage—when the trait was (a) manifest and (b) undesirable.

S.I.1.3 Empirical Tests

Having offered a general theory about the moderating effects of preferences on compositional effects along with several schematic illustrations, let us consider the evidence at hand. Our focus will be on education and age since these aspects of leadership are relatively accessible. (Regrettably, we do not have global data on other characteristics such as sexuality, social class, medical degrees, disability, or blood type.)

Along with gender, religion, language, and ethnicity (Section 2), expert coders for the GLP record the educational attainment and age of leaders. For present purposes, educational attainment is measured as a binary variable—no college/some college. We draw on the Barro and Lee dataset to obtain population-level education values (Barro and Lee 2013).

The age of leaders is classified in 5-year ranges (20-24, 25-29, etc.). This is measured as (a) current age and (b) age upon taking office. Population-level data is obtained from the World Bank's Health Nutrition and Population Statistics (`https://databank.worldbank.org/source/health-nutrition-and-population-statistics`). Ages below 20 are excluded from the construction of the fractionalization index.

Table S.I.1 follows our benchmark specification (Model 3, Table 2). Here, the outcomes of interest are representational indices of age, age upon taking office, and education (full results are posted in Table S.V.3).

We find that body size has the expected sign and is statistically significant in all models. Larger bodies are associated with better representation for age groups and education groups. These results are consistent with theoretical expectations because most of the variability in our sample is *across political bodies* rather than across countries. This provides a great deal more empirical leverage and is not subject to the sort of confounding noted below. Even so, it is worth noting that the result for education is barely significant at standard statistical thresholds. This is not surprising given that there is very little variability in the outcome—nearly all leaders have a college degree, as shown in Figure S.I.3.

By contrast, fractionalization indices for age and education bear a *positive* relationship to representation. This is the opposite of what we found in analyses focused on ethnicity, religion, language, and gender (see Table 4). We regard this result as spurious, a product of selector preferences that happen to correlate with fractionalization.

Bear in mind that countries with higher age fractionalization scores are countries with older populations. (This arises because age pyramids are more evenly distributed across age groups, rendering a higher fractionalization score.) Now consider the fact that older candidates for public office are generally preferred to younger candidates—not because they are older but because they possess other desirable attributes such as experience. In this light, it is not surprising that higher fractionalization would be associated with greater representation. But it is certainly not a causal relationship. It is simply a reflection of a general preference for candidates with desirable traits like experience, who happen to be older than the general population.

A similar phenomenon arises with respect to education. Higher fractionalization scores are generated in countries with more educated populations, where the share of the population with some college education approaches fifty percent (the maximum degree of heterogeneity possible in a two-category measure). Now, assume that candidates with a college degree are generally preferred to candidates with only a secondary school degree. (This is evident when one compares the proportion with some college education among leaders and citizens (see Figure S.I.3).) Again, one can see why a spurious association appears between fractionalization and representation.

The intuitive reason is that, with strong affinity voting, a unit increase in entropy from a baseline state having low entropy means increasing the population share of the preferred group (explaining the positive coefficient in this context). We illustrate this via simulation in Figure S.I.2

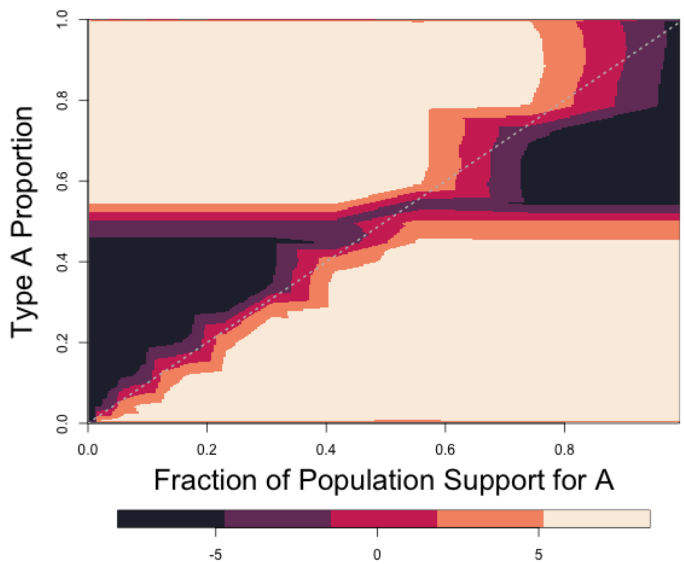

Figure S.I.2: *Here, colors denote the average t-statistic value in relating entropy with the representation index via OLS. The relationship between group entropy and representation can be either positive or negative, depending on the strength of the preference for Type A or B relative to their group population proportions. See* `SizeGetSimResults.R` *in the accompanying Dataverse (doi:10.7910) for replication files for these simulations.*

More generally, we note that the interpretation of a fractionalization index is fraught whenever the underlying categories are *ordered*, as they are for education and age (both of which are constructed as ordinal scales). Ordered categories are much more likely to be correlated with other factors that serve as confounders, especially if preferences are weak, as we have imagined them to be in the case of age and education. By contrast, we face fewer concerns with ethnicity, religion, language, and gender. As nominal categories, they are less likely to be correlated with other traits that are strongly valued (or deplored), and thus less prone to confounding in the present context.

	Age (1)	Age At Taking Office (2)	Education (3)
Body size (log)	0.11 (31.76)*	0.11 (34.11)*	0.01 (2.62)*
Fractionalization	2.96 (6.37)*	2.40 (6.41)*	0.85 (6.61)*
Continuous covariates			
Lexical index	✓	✓	✓
Population (log)	✓	✓	✓
GDP per capita (log)	✓	✓	✓
Gini index	✓	✓	✓
Factor covariates			
Body type	✓	✓	✓
Selection rule	✓	✓	✓
Round	✓	✓	✓
Other statistics			
Countries	144	144	125
Observations	2054	2031	1596
Adjusted R-squared	0.78	0.77	0.44
Dependent variable	Rep. index	Rep. index	Rep. index

Table S.I.1: *Analysis by additional group identity types. Outcome: representation index. Higher values indicate better representation. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes $p < 0.05$. Full model results are given in Table S.V.7.*

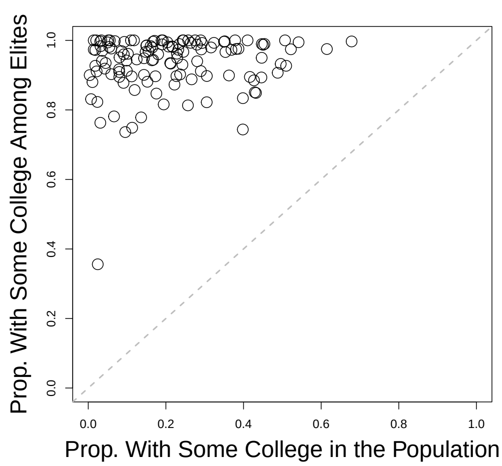

Figure S.I.3: *The proportion of leaders having some college experience is in every country in our sample greater than that in the population. In this sample, the lone country where the college education proportion among elites is less than 0.5 is Mauritania (where the proportion is 0.35).*

Supplemental Materials II: Robustness and Additional Analyses

S.II.1.4 Additional Figures and Descriptive Statistics

	Mean	SD	Median	Min	Max
Fractionalization	0.43	0.21	0.50	0.00	0.96
log(GDP per capita)	9.15	1.16	9.33	5.70	11.95
log(Population)	16.33	1.57	16.17	12.68	21.05
log(Body N)	1.99	1.67	1.79	0.00	7.99
Lexical index	5.09	1.74	6.00	0.00	6.00
Gini coefficient	37.69	8.07	37.10	16.60	81.00

Table S.II.1: *Descriptive statistics table.*

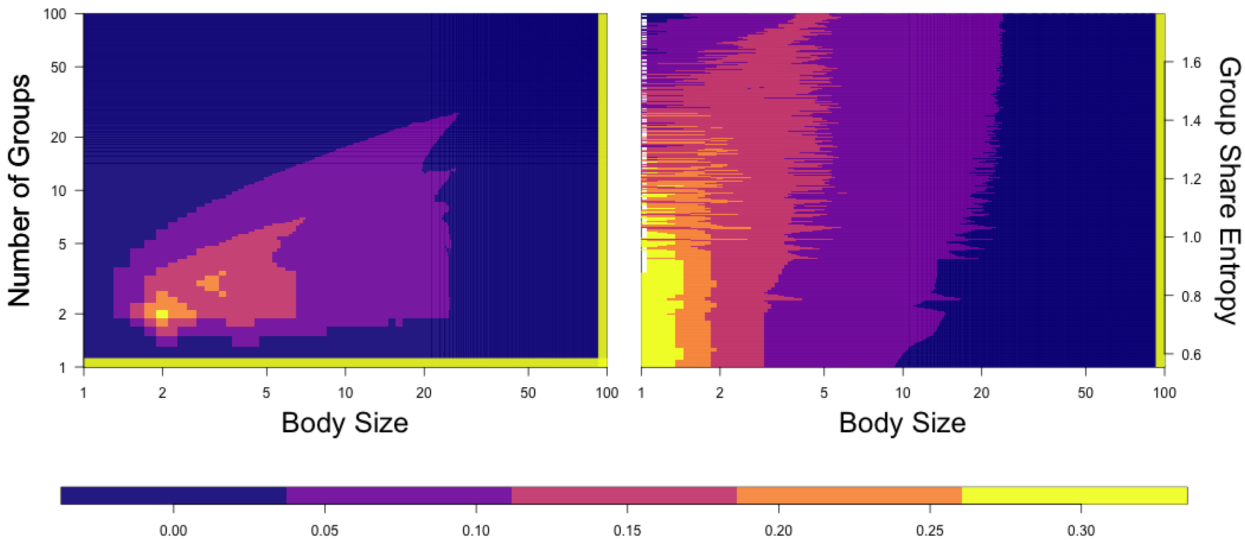

Figure S.II.1: *A depiction of the uncertainty of the random sampling model of descriptive representation. Color values indicate different levels of the residual standard deviation using the expected value of the representation index as outlined in Equation 2 to predict observed representation. That is, colors correspond to* $\sqrt{\mathbb{E}[(R_b - \mathbb{E}[R_b])^2]}$. *In the left panel, we see that increasing the number of groups decreases the residual standard deviation. Increasing the body size generally decreases the residual standard deviation. Entropy has a somewhat more complex relationship with the residual standard deviation. In these figures,* $\mathbb{E}[R_b]$ *is found using our analytical formula;* $\sqrt{\mathbb{E}[(R_b - \mathbb{E}[R_b])^2]}$ *is found using Monte Carlo methods. The simulation design is the same as that used in Figure 2 in the main text.*

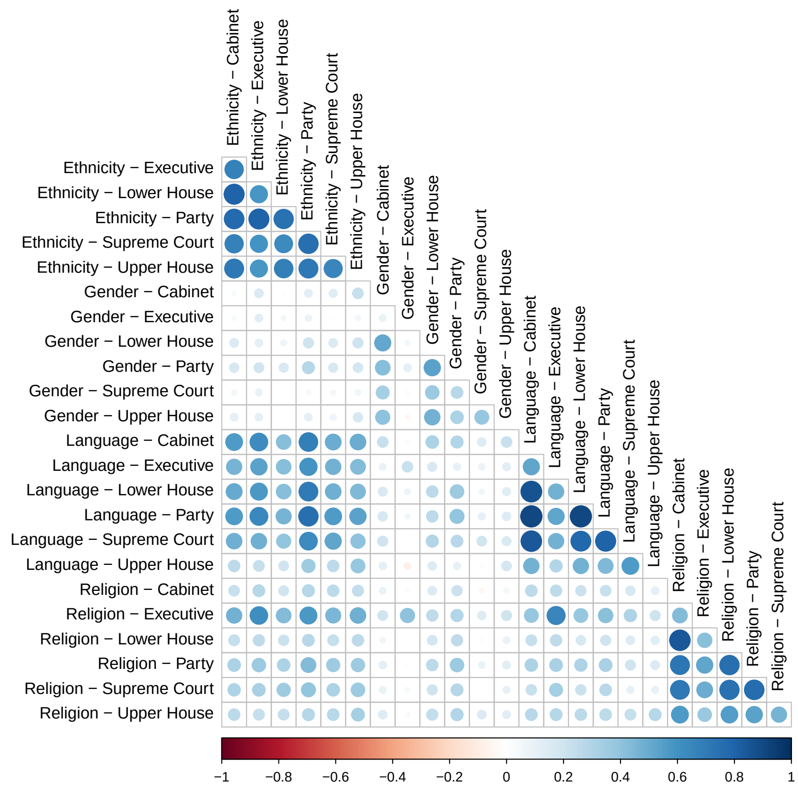

Figure S.II.2: *Correlations between representation indices across group and body. Colors represent correlation values (with negative correlations taking on red and positive correlations blue shades).*

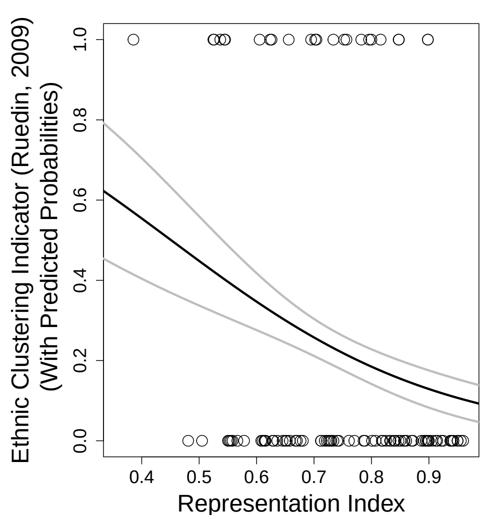

Figure S.II.3: *Comparison of mean country-level ethnic representation with the country-level "heavy ethnic clustering" indicator from Ruedin (2009).*

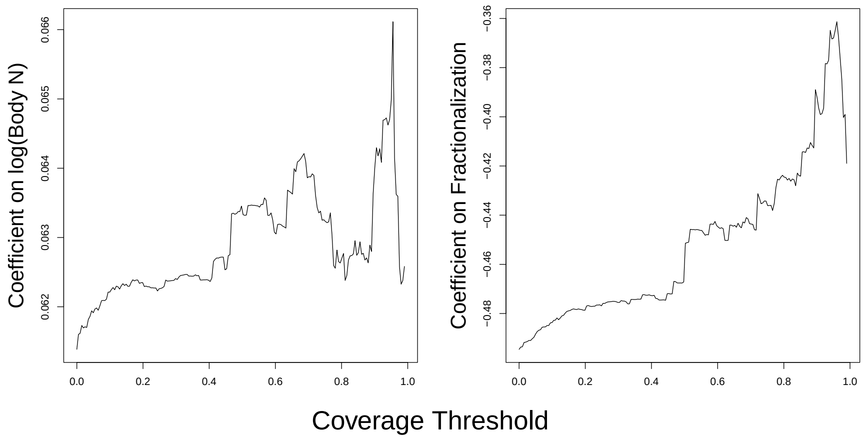

Figure S.II.4: *Sensitivity of main results from Model 4 in Table 2 to choice of coverage threshold.*

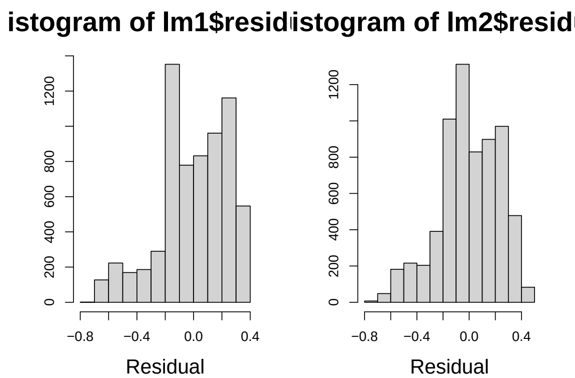

Figure S.II.5: *Residuals using body size (left) and log(body size) (right) to predict the representation index.*

S.II.1.5 Additional Analysis Exploring Representation and Dimensions of Identity

	Model 1	Model 2	Model 3	Model 4
Ethnicity baseline	✓			
Language	0.06 (1.79)			
Religion	-0.03 (-1.10)			
Gender	-0.04 (-1.76)	-0.05 (-2.72)*	0.01 (1.11)	0.01 (1.29)
Body size (log)			0.05 (10.74)*	0.05 (10.80)*
Fractionalization			-0.69 (-19.85)*	-0.69 (-23.83)*
Continuous covariates				
Lexical index			✓	
Population (log)			✓	
GDP per capita (log)			✓	
Gini index			✓	
Factor covariates				
Body type			✓	✓
Gender quota type			✓	✓
Ethnicity quota type			✓	✓
Selection rule			✓	✓
Round			✓	✓
Country				✓
Other statistics				
Countries	156	156	156	156
Observations	6628	6628	6628	6628
Adjusted R-squared	0.02	0.01	0.47	0.47
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.2: *Sub-group analysis. Outcome: representation, measured for each identity—ethnicity, religion, language, and gender. Higher values indicate better representation. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes $p < 0.05$, Full model results are given in Table S.V.9.*

To probe the relationship between representation and different dimensions of identity, regression analyses are conducted in Table S.II.2. Model 1 includes dummies for each dimension of identity, with ethnicity as the excluded category. Not surprisingly, there are no statistically significant differences across these coefficients.

Model 2 drops all identity dummies except gender, which we have reason to believe (based on its unique distribution) may be different. Again, there is no significant difference when contrasted with other identities (now part of the excluded category).

Model 3 adds fractionalization to the specification. Interestingly, the sign for gender flips and the relationship is statistically significant. In other words, once we control for the distribution of identity groups in societies across the world, gender representation is superior to representation along other dimensions. This pattern is robust even when other covariates are added to the specification, as shown in Model 4.

Earlier, we noted that gender is distinct from other identities insofar as the population is split into two relatively equal-sized groups. Entropy is extremely high, giving gender a relatively high score on the fractionalization index. By contrast, many societies are dominated by a single ethnic group (mean dominant group share of 69%), a single religious group (mean dominant group share of 72%), and a single linguistic group (mean dominant group share of 78%). Once we take this factor into account, gender achieves a higher-than-expected score.

To see why this matters, let us imagine a society in which there are strong norm-based objections to the representation of women (comprising 50% of the population) as well as to the representation of a small ethnic group (comprising 5% of the population). As a consequence, both are under-represented. However, the small ethnic group can be under-represented only by five points on our index while women can be under-represented by fifty points on our index. The exclusion of small groups matters less than the exclusion of large groups. (Of course, exclusion is consequential for the excluded group; but it is less consequential for the ideal of representation, considered across all citizens.) Once we build this into the model, by including fractionalization on the right side, gender representation is better than expected. It is nevertheless important to emphasize that our metric of representation is symmetric so that a country that has some bodies over-representing women and some bodies under-representing women could have the same average representation score as a country whose bodies always underrepresent women.

S.II.1.6 Robustness with Randomly Aggregated Ethnicities

	Model 2	Model 3	Model 4	Model 5
Expected representation	0.93 (74.07)*			
Body size (log)		0.04 (7.67)*	0.05 (6.90)*	0.05 (6.61)*
Fractionalization		-0.71 (-22.53)*	-0.48 (-6.72)*	-0.66 (-20.35)*
Continuous covariates				
Lexical index				0.01 (1.56)
Population (log)				0.00 (-0.65)
GDP per capita (log)				0.00 (-0.53)
Gini index				0.00 (0.19)
Factor covariates				
Body type			✓	✓
Ethnicity quota type			✓	✓
Selection rule			✓	✓
Round			✓	✓
Country			✓	
Intercept		✓	✓	✓
Other statistics				
Countries	152	152	152	152
Observations	1756	1756	1756	1756
Adjusted R-squared	0.59	0.63	0.71	0.65
Unit of analysis	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.3: *Main analysis. Outcome: representation index (where 1 = perfect representation) for ethnicity, where ethnicities have been randomly grouped. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes* $p < 0.05$. *In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.10.*

S.II.1.7 Description of Hierarchical Bootstrap Procedure

Goal: Account for country- and leader-level dependencies using a hierarchical bootstrap procedure (Efron 1982; Carpenter, Goldstein, and Rasbash 2003).

Hierarchical Bootstrap Description:

1. For $boot \in \{1, ..., n_{\text{Boot}}\}$:
    (a) Sample countries from the set of countries with replacement
        i. Within each sampled country:
            A. Sample leaders with replacement
            B. Compute discrepancies by group and body
    (b) With bootstrap dataset, fit model and save coefficients
2. Assess statistical significance by examining the 95% bootstrap intervals for each coefficient (using the n_{Boot} realizations) or by computing a bootstrap t-statistic.

S.II.1.8 Results With Uncertain Ethnicities Randomly Drawn

To assess robustness of the results to uncertainties in the coded identities, we rerun the main analysis. In the reanalysis of the data, we randomly perturb the ethnic identities of individuals whose ethnicity was coded, but where coders were not confident in this decision. We randomly draw the ethnicity of these individuals from the pool of ethnicities in the country. We find no significant sensitivity of the results to this perturbation.

	Model 2	Model 3	Model 4	Model 5
Expected representation	0.87 (61.61)*			
Body size (log)		0.06 (11.82)*	0.07 (10.23)*	0.07 (10.01)*
Fractionalization		-0.67 (-19.32)*	-0.33 (-3.16)*	-0.63 (-17.32)*
Continuous covariates				
Lexical index				0.00 (0.55)
Population (log)				-0.01 (-2.57)*
GDP per capita (log)				0.00 (-0.18)
Gini index				0.00 (-0.88)
Factor covariates				
Identity			✓	✓
Body type			✓	✓
Gender quota type			✓	✓
Ethnicity quota type			✓	✓
Selection rule			✓	✓
Round			✓	✓
Country			✓	
Intercept		✓	✓	✓
Other statistics				
Countries	152	152	152	152
Observations	2082	2082	2082	2082
Adjusted R-squared	0.48	0.51	0.59	0.54
Unit of analysis	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.4: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes $p < 0.05$. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.11.*

S.II.1.9 Results With Party Groups in Lower House and >75% Inclusion Criterion

S.II.1.9.1 With Standard Errors Clustered by Country

See main text.

S.II.1.9.2 With Bootstrap Standard Errors

	Model 1	Model 2	Model 3	Model 4	Model 5
Expected representation	0.84 (87.72)*	0.71 (94.10)*			
Body size (log)			0.05 (24.16)*	0.06 (27.97)*	0.06 (24.43)*
Fractionalization			-0.46 (-18.85)*	-0.41 (-13.74)*	-0.45 (-17.11)*
Continuous covariates					
Lexical index					0.01 (2.53)*
Population (log)					-0.01 (-2.48)*
GDP per capita (log)					0.03 (8.33)*
Gini index					0.00 (-3.95)*
Factor covariates					
Identity				✓	✓
Body type				✓	✓
Gender quota type				✓	✓
Ethnicity quota type				✓	✓
Selection rule				✓	✓
Round				✓	✓
Country				✓	
Intercept			✓	✓	✓
Other statistics					
Countries	156	153	153	153	153
Observations	156	13830	13830	13830	13830
Adjusted R-squared	0.60	0.36	0.31	0.66	0.64
Unit of analysis	C	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.5: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes $p < 0.05$. Missing values were imputed in Model 1 to ensure compatibility across country and standard errors calculated via the country-level block bootstrap (with the imputation model re-fit on every bootstrap draw); see main text. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.12.*

S.II.1.10 Results Without Party Groups in Lower House

S.II.1.10.1 With Standard Errors Clustered by Country

	Model 1	Model 2	Model 3	Model 4	Model 5
Expected representation	0.84 (87.72)*	0.83 (123.29)*			
Body size (log)			0.03 (12.00)*	0.02 (4.12)*	0.02 (3.52)*
Fractionalization			-0.61 (-29.49)*	-0.61 (-15.42)*	-0.60 (-24.60)*
Continuous covariates					
Lexical index					0.00 (0.79)
Population (log)					0.00 (-0.46)
GDP per capita (log)					0.01 (2.05)*
Gini index					0.00 (-0.74)
Factor covariates					
Identity				✓	✓
Body type				✓	✓
Gender quota type				✓	✓
Ethnicity quota type				✓	✓
Selection rule				✓	✓
Round				✓	✓
Country				✓	
Intercept			✓	✓	✓
Other statistics					
Countries	156	157	157	157	157
Observations	156	3735	3735	3735	3735
Adjusted R-squared	0.60	0.26	0.42	0.49	0.46
Unit of analysis	C	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.6: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes p* < 0.05*. Missing values were imputed in Model 1 to ensure compatibility across country and standard errors calculated via the country-level block bootstrap (with the imputation model re-fit on every bootstrap draw); see main text. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.13.*

S.II.1.10.2 With Bootstrap Standard Errors

	Model 1	Model 2	Model 3	Model 4	Model 5
Expected representation	0.84 (87.72)*	0.68 (109.10)*			
Body size (log)			0.05 (31.41)*	0.03 (6.85)*	0.03 (6.55)*
Fractionalization			-0.37 (-16.13)*	-0.28 (-12.16)*	-0.31 (-15.05)*
Continuous covariates					
Lexical index					0.00 (0.89)
Population (log)					-0.01 (-2.66)*
GDP per capita (log)					0.04 (11.38)*
Gini index					0.00 (-3.91)*
Factor covariates					
Identity				✓	✓
Body type				✓	✓
Gender quota type				✓	✓
Ethnicity quota type				✓	✓
Selection rule				✓	✓
Round				✓	✓
Country				✓	
Intercept			✓	✓	✓
Other statistics					
Countries	156	149	149	149	149
Observations	156	6399	6399	6399	6399
Adjusted R-squared	0.60	0.29	0.25	0.71	0.69
Unit of analysis	C	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.7: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes* $p < 0.05$. *Missing values were imputed in Model 1 to ensure compatibility across country and standard errors calculated via the country-level block bootstrap (with the imputation model re-fit on every bootstrap draw); see main text. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.14.*

S.II.1.11 Results Without >75% Coverage Condition

S.II.1.11.1 With Standard Errors Clustered by Country

	Model 1	Model 2	Model 3	Model 4	Model 5
Expected representation	0.84 (87.72)*	0.84 (98.11)*			
Body size (log)			0.04 (14.01)*	0.05 (14.85)*	0.05 (13.21)*
Fractionalization			-0.71 (-29.47)*	-0.68 (-18.15)*	-0.70 (-25.58)*
Continuous covariates					
Lexical index					0.01 (2.22)*
Population (log)					0.00 (-1.69)
GDP per capita (log)					0.00 (0.98)
Gini index					0.00 (-1.82)
Factor covariates					
Identity				✓	✓
Body type				✓	✓
Gender quota type				✓	✓
Ethnicity quota type				✓	✓
Selection rule				✓	✓
Round				✓	✓
Country				✓	
Intercept			✓	✓	✓
Other statistics					
Countries	156	157	157	157	157
Observations	156	9438	9438	9438	9438
Adjusted R-squared	0.60	0.39	0.48	0.52	0.49
Unit of analysis	C	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.8: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes* $p < 0.05$. *Missing values were imputed in Model 1 to ensure compatibility across country and standard errors calculated via the country-level block bootstrap (with the imputation model re-fit on every bootstrap draw); see main text. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.8.*

S.II.1.11.2 With Bootstrap Standard Errors

	Model 1	Model 2	Model 3	Model 4	Model 5
Expected representation	0.84 (87.72)*	0.71 (100.42)*			
Body size (log)			0.05 (28.34)*	0.06 (28.89)*	0.06 (24.98)*
Fractionalization			-0.47 (-21.58)*	-0.47 (-17.76)*	-0.50 (-22.19)*
Continuous covariates					
Lexical index					0.01 (2.64)*
Population (log)					0.00 (-1.80)
GDP per capita (log)					0.03 (8.68)*
Gini index					0.00 (-3.96)*
Factor covariates					
Identity				✓	✓
Body type				✓	✓
Gender quota type				✓	✓
Ethnicity quota type				✓	✓
Selection rule				✓	✓
Round				✓	✓
Country				✓	
Intercept			✓	✓	✓
Other statistics					
Countries	156	156	156	156	156
Observations	156	18440	18440	18440	18440
Adjusted R-squared	0.60	0.36	0.33	0.66	0.63
Unit of analysis	C	C-B-G	C-B-G	C-B-G	C-B-G
Dependent variable	Rep. index	Rep. index	Rep. index	Rep. index	Rep. index

Table S.II.9: *Main analysis. Outcome: representation index (where 1 = perfect representation), measured across various identities—ethnicity, religion, language, and gender. Estimator: ordinary least squares, t-statistics in parentheses, standard errors clustered by country. * denotes $p < 0.05$. Missing values were imputed in Model 1 to ensure compatibility across country and standard errors calculated via the country-level block bootstrap (with the imputation model re-fit on every bootstrap draw); see main text. In the unit of analysis row, "C" denotes country, "G" denotes group, and "B" denotes body. Full model results are given in Table S.V.15.*

S.II.1.12 Proof of Equation 2

S.II.1.12.1 Expected Representation Index Using Squared Deviations

We assume that there exists an infinite population from which political leaders are drawn uniformly. Before we derive the expression for the expected absolute deviations of the body group shares from those in the population, we first consider the simpler case where the metric of interest is the sum of squared body-group differences. In this case, we have

$$R_b^{(2)} = 1 - \frac{1}{2} \times \sum_{k=1}^{K} (g_{p_k} - G_{b_k})^2 \tag{1}$$

where g_{p_k} denotes the population group share for group k, G_{b_k} denotes the body group share for group k, and K denotes the total number of groups. We want to find:

$$\mathbb{E}[R_b^{(2)}] = \mathbb{E}\Big[1 - \frac{1}{2} \times \sum_{k=1}^{K} (g_{p_k} - G_{b_k})^2\Big], \tag{2}$$

where the expectation is taken over the uniform sampling process of members of the population to the political body.

Now, because the body size, n_b, is here considered to be fixed, $\mathbf{G}_b \times n_b \sim \text{Multinomial}(\mathbf{g}_b, n_b)$, where $G_{b_k} \times n_b$ represents the counts associated with group k in the body (e.g., the number of female leaders in the lower house). This distributional equality holds because leaders are drawn from the population with probability proportional to their group share, g_{b_k}, and this process is repeated n_b times.

We just need to find $\mathbb{E}[(g_{p_k} - G_{b_k})^2]$, which, by linearity of expectations, will yield the full expression for Equation 2. We see that

$$\begin{aligned}
\mathbb{E}[(g_{p_k} - G_{b_k})^2] &= \mathbb{E}[g_{p_k}^2 - 2g_{p_k}G_{b_k} + G_{b_k}^2] \\
&= g_{p_k}^2 - 2g_{p_k}\mathbb{E}[G_{b_k}] + \mathbb{E}[G_{b_k}^2] \\
&= g_{p_k}^2 - 2g_{p_k}\frac{1}{n_b} \times \mathbb{E}[n_b G_{b_k}] + \mathbb{E}[G_{b_k}^2] \ \text{ (multiply by 1)} \\
&= g_{p_k}^2 - 2g_{p_k}\frac{1}{n_b} \times n_b g_{p_k} + \mathbb{E}[G_{b_k}^2] \ \text{ (by the Multinomial expected value)}
\end{aligned}$$

Using a similar line of reasoning, we see

$$\begin{aligned}
\mathbb{E}[G_{b_k}^2] &= \frac{1}{n_b^2} \times \mathbb{E}[(n_b G_{b_k})^2] \ \text{ (multiply by 1)} \\
&= \frac{1}{n_b^2} \times \Big[n_b(g_{p_k})(1 - (g_{p_k})) + (n_b g_{p_k})^2\Big] \ \text{ (using the Multinomial variance \& fact that } \mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2\text{)}
\end{aligned}$$

Putting everything together,

$$\mathbb{E}[(g_{p_k} - G_{b_k})^2] = g_{p_k}^2 - 2g_{p_k}\frac{1}{n_b} \times n_b g_{p_k} + \frac{1}{n_b^2} \times \Big[n_b(g_{p_k})(1 - (g_{p_k})) + (n_b g_{p_k})^2\Big],$$

After some algebra, we see:

$$\mathbb{E}[(g_{p_k} - G_{b_k})^2] = \frac{g_{p_k}(1 - g_{p_k})}{n_b}. \tag{3}$$

Because $\mathbb{E}[R_b^{(2)}]$ is equivalent to summing over the variance terms for each category draw and because sums are linear operators, we therefore conclude

$$\mathbb{E}[R_b^{(2)}] = 1 - \frac{1}{2} \times \frac{\sum_{k=1}^{K} g_{b_k}(1 - g_{b_k})}{n_b}. \tag{4}$$

This equation is straightforward to interpret. We can conclude from it that the representation index involving the squared body-population increases proportionally with the body size. Moreover, we see that $\mathbb{E}[R_b^{(2)}]$ is smaller when population shares become more uniform. For example, in the binary case, $g_{b_k}(1 - g_{b_k})$ is maximized at $g_{b_k} = 0.5$.

S.II.1.12.2 Expected Representation Index Using Absolute Deviations

When using the Representation Index defined using absolute values (as in the representation index from Equation 1) instead of squared values, we have

$$R_b = 1 - \frac{1}{2} \sum_{k=1}^{K} |g_{p_k} - G_{b_k}|$$

Here, we again use the fact that expectations are linear operators and focus on $\mathbb{E}[|g_{p_k} - G_{b_k}|]$ for some k. We also recall the close relationship between the Multinomial and Binomial distributions. In particular, any single dimension in a Multinomial random vector is itself Binomially distributed (i.e. marginal counts of a Multinomial are Binomial). Therefore, we know that $G_{b_k} \times n_b \sim \text{Binomial}(g_{b_k}, n_b)$. We can here apply the expression for the expected absolute deviation for the Binomial from Kenney and Keeping (1962):

$$\mathbb{E}\left[|n_b G_{b_k} - \mathbb{E}[n_b G_{p_k}]|\right] = \mathbb{E}[|n_b G_{b_k} - n_b g_{p_k}|] = 2\left\{(1 - g_{p_k})^{n_b - \lfloor n_b g_{p_k} \rfloor} \times g_{p_k}^{\lfloor n_b g_{p_k} \rfloor + 1}(\lfloor n_b g_{p_k} \rfloor + 1)\binom{n_b}{\lfloor n_b g_{p_k} \rfloor + 1}\right\}.$$

Because $n_b > 0$, we can take

$$\mathbb{E}[|G_{b_k} - g_{p_k}|] = \frac{2}{n_b}\left\{(1 - g_{p_k})^{n_b - \lfloor n_b g_{p_k} \rfloor} \times g_{p_k}^{\lfloor n_b g_{p_k} \rfloor + 1}(\lfloor n_b g_{p_k} \rfloor + 1)\binom{n_b}{\lfloor n_b g_{p_k} \rfloor + 1}\right\}$$

and thus, by linearity of expectations,

$$\mathbb{E}[R_b] = 1 - \frac{1}{n_b} \sum_{k=1}^{K} \left\{(1 - g_{p_k})^{n_b - \lfloor n_b g_{p_k} \rfloor} \times g_{p_k}^{\lfloor n_b g_{p_k} \rfloor + 1}(\lfloor n_b g_{p_k} \rfloor + 1)\binom{n_b}{\lfloor n_b g_{p_k} \rfloor + 1}\right\}.$$

We illustrate the accuracy of Equation 2 in Figure S.II.6, where we compute the distribution of R_b for the United States under the random sampling assumption. We use the lower house as the body type and ethnicity as the group type in this experiment. The Monte Carlo mean and analytical expectation from Equation 2 cannot be visually distinguished, confirming that the analytical calculations are, indeed, accurate. We also note that the observed representation index is below expected value we obtain under the random sampling assumption, an indication that the political system is much less representative than we would expect under random sampling. Below, we also make available code for computing this expected value.

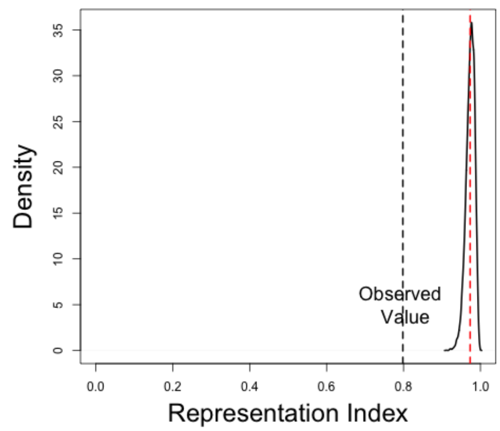

Figure S.II.6: *Validating Equation 2 via Monte Carlo simulations. We use the lower house as the body type and ethnicity as the group type. The observed representation index value is also shown. The Monte Carlo mean and analytical expectation cannot be visually distinguished.*

Alternative representation measures similar to the absolute deviations from proportionality also exist. For example, one might normalize by the number of groups, K. However, this would make interpretation difficult, as the intrinsic scale depends on context. Moreover, the scale would be sensitive to the addition of groups having population and body shares of near 0 because the inclusion/exclusion of small groups, although contributing little to the $|g_{p_k} - G_{b_k}|$ terms, would inflate/deflate the normalization factor, K.

```
expected_representation_index <- function(pop_share, bodyN){
  # function inputs:
  # pop_share -- a vector of population shares that sums to 1 and contains non-zero values
  # bodyN -- a positive integer denoting the body size

  # first, compute each contribution to the sum from k=1 to K on the log scale
  # the log scale is used for computational robustness in dealing with large body sizes
  log_contrib_k <- (bodyN - floor(bodyN*pop_share))*log(1 - pop_share) +
      (floor(bodyN*pop_share)+1)*log(pop_share) +
      log(floor(bodyN*pop_share)+1) +
      lchoose(bodyN,floor(bodyN*pop_share)+1)

  # convert back to original scale
  contrib_k <- exp(  log_contrib_k )

  # compute final score
  return( 1 -   sum( contrib_k / bodyN )  )
}
```

S.II.1.13 Other Additional Analyses

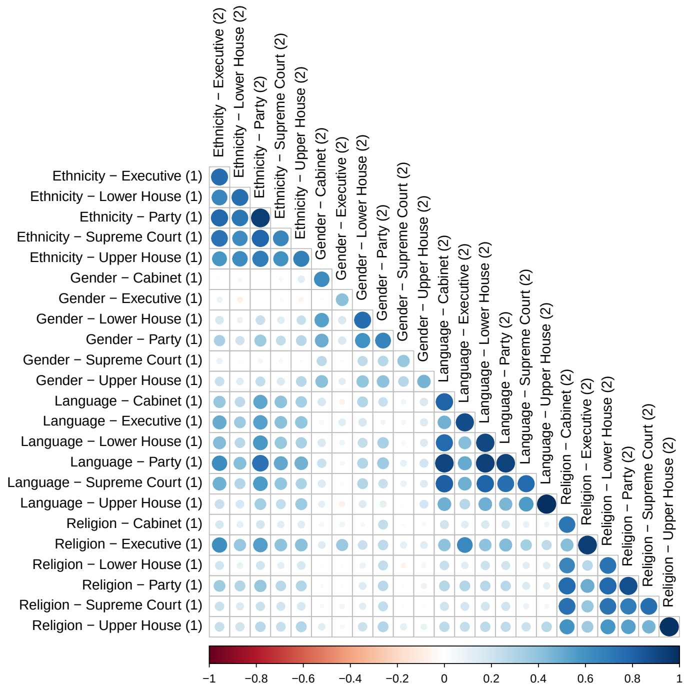

Figure S.II.7: *Observed representation across time between Round 1 (2010-2013) and Round 2 (2017-2019). Correlations between representation indices across group and body between Rounds 1 and 2. Colors represent correlation values (with negative correlations taking on red and positive correlations blue shades). We see the most significant correlations between Round 1 (2010-2013) representation and Round 2 (2017-19) representation for ethnicity. For gender, there is a lower, but still significant, correlation across time for each body. This correlation is strongest for legislative bodies, and weakest for the executive.*

	Lasso Coefficient
Intercept	0.843
Body size (log)	0.036
Body type - Executive	
Body type - Lower house	
Body type - Party	-0.003
Body type - Supreme court	
Body type - Upper house	
Lexical index	0.002
Fractionalization	-0.642
Population (log)	
GDP per capita (log)	0.003
Gini index	
Selection rule - 2	
Selection rule - 3	
Selection rule - 4	
Selection rule - 5	
Selection rule - 6	0.019
Group type - Gender	
Group type - Language	
Group type - Religion	-0.022
Quotas type - 1	
Quotas type - 2	
Quotas type - 3	
Quotas type - 4	
Ethnic quota	
Round	

Table S.II.10: *Main analysis. Outcome: levels of representation (where 1 = perfect representation) estimated to be exactly 0 are listed as blank entries. The regularization penalty was selected via cross-validation, with the selected penalty being the largest value within 1 standard deviation of the cross-validation error minimizer.*

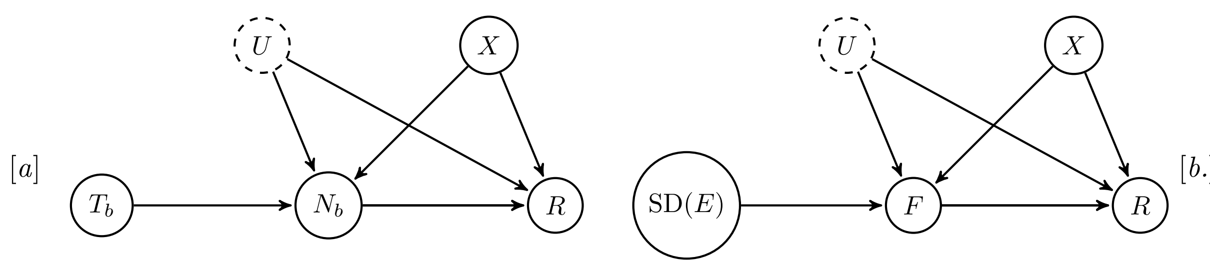

Figure S.II.8: *This figure illustrates the assumptions of the IV analysis. In the left panel, we illustrate the assumption that, among elected bodies, the body type (T_b) only affects the representation index (R) through the effect of body type on body size (N_b), using the fact that some bodies tend to be bigger than others. In the right panel, we illustrate the assumption that variability in elevation (SD(E)) affects the representation score (R) only through its effect on ethnic fractionalization (F), for example, because more variable geographies tend to have a greater number and dispersion of ethnic groups). In both figures, U refers to unobserved confounders and X refers to observed confounders.*

References

Adam, Barry D. 2009. *The Global Emergence of Gay and Lesbian Politics: National Imprints of a Worldwide Movement.* Temple University Press.

Arceneaux, Kevin and Ryan J Vander Wielen. 2023. "Do Voters Prefer Educated Candidates? How Candidate Education Influences Vote Choice in Congressional Elections". *Electoral Studies* 82: 102596.

Ayoub, Phillip M. 2016. *When States Come Out: Europe's Sexual Minorities and the Politics of Visibility.* Cambridge Studies in Contentious Politics. Cambridge University Press.

Barro, Robert J and Jong Wha Lee. 2013. "A New Data Set of Educational Attainment in the World, 1950–2010". *Journal of Development Economics* 104: 184–198.

Burden, Barry C. 2007. *Personal Roots of Representation.* Princeton University Press.

Campbell, Rosie and Philip Cowley. 2014. "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment". *Political Studies* 62 (4): 745–765.

Carnes, Nicholas and Noam Lupu. 2016. "Do Voters Dislike Working-class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class". *American Political Science Review* 110 (4): 832–844.

Carnes, Nicholas and Noam Lupu. 2023. "The Economic Backgrounds of Politicians". *Annual Review of Political Science* 26.

Carpenter, James R, Harvey Goldstein, and Jon Rasbash. 2003. "A Novel Bootstrap Procedure for Assessing the Relationship between Class Size and Achievement". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52 (4): 431–443.

Efron, Bradley. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM.

Gift, Thomas and Carlos X Lastra-Anadón. 2018. "How Voters Assess Elite-educated Politicians: A Survey Experiment". *Electoral Studies* 56: 136–149.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments". *Political Analysis* 22 (1): 1–30.

Kenney, J. F. and E. S. Keeping. 1962. "Mean Absolute Deviation". In *Mathematics of Statistics*, pp. 76–77. Princeton, NJ: Van Nostrand.

Norris, Pippa and Ronald Inglehart. 2004. *Cultural Barriers to Women's Leadership: A Worldwide Comparison.* Citeseer.

Paxton, Pamela, Melanie M. Hughes, and Tiffany D. Barnes. 2020. *Women, Politics, and Power: A Global Perspective, 4th Edition.* Lanham, MD: Rowman Littlefield Press.

Paxton, Pamela and Sheri Kunovich. 2003. "Women's Political Representation: The Importance of Ideology". *Social Forces* 82 (1): 87–113.

Ruedin, Didier. 2009. "Ethnic Group Representation: A Cross-National Comparison". *Journal of Legislative Studies* 15 (4): 335–54.

Ruedin, Didier. 2013. *Why Aren't They There?: The Political Representation of Women, Ethnic Groups and Issue Positions in Legislatures. Colchester*. ECPR Press.

Simon, Elizabeth and Stuart J Turnbull-Dugarte. 2023. "Education as Identity? A Meta-analysis of Education-based In-group Preferences in Candidate Choice Experiments".

Stockemer, Daniel and Aksel Sundström. 2018. "Age Representation in Parliaments: Can Institutions Pave the Way for the Young?". *European Political Science Review* 10 (3): 467–490.