The Global Legislators Database

The Personal Backgrounds of National Legislators in the World's Democracies

Nicholas Carnes* Joshua Ferrer† Miriam Golden‡ Esme Lillywhite§
Noam Lupu¶ Eugenia Nazrullaeva‖

December 9, 2022; Version 2.2; word count 3,800

*Professor, Sanford School of Public Policy, Duke University, 201 Science Drive, SB 227, Box 90245, Durham NC 27708, +1-919-438-0109, nicholas.carnes@duke.edu

†Ph.D. candidate, Department of Political Science, 4289 Bunche Hall, University of California at Los Angeles, Los Angeles, CA 90095, +1-310-825-4331, joshuaferrer@g.ucla.edu

‡Corresponding author, Peter Mair Chair in Comparative Politics, European University Institute, 9 via della Badia dei Roccettini, 50014 Fiesole FI, Italy, +39-366-604-7310, miriam.golden@eui.eu

§Ph.D. candidate, Institute for Inspiring Children's Futures, University of Strathclyde, Glasgow, Scotland G1 1XQ UK, +44-730-503-9880, esme.lillywhite@strath.ac.uk

¶Associate Professor, Department of Political Science, Vanderbilt University, Commons Center, PMB 0505, 230 Appleton Place, Nashville, TN 37203, +1-615-875-9181, noam.lupu@vanderbilt.edu

‖Post-doctoral fellow, School of Public Policy, London School of Economics and Political Science, Houghton Street, London WC2A 2AE UK, +44-730-503-9880, e.nazrullaeva@lse.ac.uk

Abstract

This note describes the Global Legislators Database (GLD), a new crossnational dataset on characteristics — political party, gender, age, education, and occupational background — of the roughly 20,000 lawmakers in the world's democracies. The database includes 97 democracies (of 103) with populations over 300,000, with information about the 99.9 percent of legislators who held office in each country's lower chamber or unicameral legislature during one legislative session in 2016 or 2017. The GLD is the largest individual-level biographical database on national legislatures ever assembled, and it has a wide range of potential applications. In this note, we show that the GLD's estimates of characteristics such as female representation are strongly validated by alternative estimates; we preview one potential application by conducting tests of hypotheses about gender, education, and occupationally-based gaps in reelection rates; and we discuss other possible uses for this one-of-a-kind resource for studying representation in the world's democracies. [150 words]

Research on numerical or descriptive representation of social groups in elected political institutions has surged in political science (e.g., Carnes and Lupu, 2023; Gulzar, 2021; Wängnerud, 2009; Krcmaric et al., 2020). But data availability remains a fundamental challenge for scholars of representation. Off-the-shelf data on the characteristics of members of governing bodies is rarely available. For many times, places, and institutions, scholars have not collected usable data on politicians' gender, race, education, age, partisanship, and occupation.

In this note, we describe a new crossnational database — the Global Legislators Database (GLD) — that provides the most detailed and comprehensive information to date about the backgrounds and demographics of national legislators in the world's democracies.[1] The dataset focuses on the lower or unicameral chamber in the national legislatures of 97 democracies, almost all of the world's 103 democracies with more than 300,000 residents.[2] It includes information about 19,704 lawmakers and represents the culmination of five years of work and the international collaboration of four principal investigators and 28 undergraduate and graduate research assistants.

For each country, we have assembled information about the members of one national legislature that sat between 2016 and 2017.[3] For each member who held office, the GLD includes political party affiliation, gender, education, age, and previous occupation. Occupation is also classified using an internationally-standardized coding system, which allows researchers to use the dataset without first coding the raw text of occupational descriptions.

The Global Legislators Database can be used to answer a wide range of questions. It can answer questions about whether the social class makeup of national legislatures matches the makeup of the country's labor force, or whether the makeup of parliaments varies with country-level characteristics like the level of economic development, the concentration of political parties, or regime characteristics. Because the dataset is constructed at the level of individual parliamentarians, it can also be used to study the characteristics of political parties (e.g., Do political parties on the ideo-

[1] We refer to the individuals comprising the dataset as legislators or representatives.

[2] The list of countries is available in Appendix A in the Supplementary Information.

[3] The legislatures we study were elected between 2011 and 2017, with most having been elected between 2013 and 2016.

logical right elect fewer women?) or individual legislators (e.g., Do lawmakers with more formal education vote differently than their less educated counterparts?), or other divisions (e.g. Are older legislators more frequent in countries in Europe compared to Latin America?).

This note summarizes the key features of the Global Legislators Database, presents the results of brief validation exercises, and highlights just one of many potential applications of the dataset with an analysis that combines an external dataset on reelection rates with the new biographical data we have assembled. [4]

The Global Legislators Database

We began constructing the GLD by identifying the 103 countries that Freedom House defined as democracies in 2016 that had populations over 300,000. Over the course of the data collection process, it became clear that we could not obtain reliable data for six of these countries, so the final GLD consists of the remaining 97 democracies. If a country had more than one legislative session during the years 2016 and 2017, we selected one at random.

The variables included in the Global Legislators Database are legislator name, date of birth, gender, political party affiliation when elected, last occupation prior to being elected to public office, and level of education attained prior to the current term when the individual was elected. Occupations are coded into 128 minor groups using the International Labor Organization's (ILO) International Standard Classification of Occupations (ISCO-08), the most widely-accepted international occupational classification system.

The dataset also captures contextual data for each country, including year of the election that produced the legislature, years when the legislature sat, legislature number, total number of legislators in the chamber, number of those legislators included in our dataset, and the date that our team performed its final verification of the data for that country (to provide a data collection date). The GLD also includes extensive sourcing information for each country and for many individual

[4] To give readers a clear idea of the work behind the construction of the GLD, we make the codebook available as Appendix B in the Supplementary Information.

legislators, making the dataset as transparent and reproducible as possible. Sourcing and coding decisions are documented in the GLD codebook (Appendix B in the SI).

One of our primary goals was to create the most exhaustive dataset possible and thus to eliminate missingness. The legislators included in the GLD were checked against official parliamentary lists. In the few countries for which we could not locate a canonical list of elected legislators for the selected legislature, we triangulated against other sources and consulted domestic election authorities. As a result, there are only 10 countries with discrepancies between the number of legislators recorded in the GLD and the number of seats in parliament, totaling 26 missing legislators out of 19,730 — a successful inclusion rate of 99.9 percent.

In addition to conducting online searches, research assistants contacted parliamentary offices, individual legislators, and country experts. Thanks to these efforts, we have information on date of birth for 90.6 percent, gender for 99.5 percent, occupational data for 93.6 percent, and educational data for 90.1 percent of all included legislators.

The GLD represents a significant improvement in crossnational data on the personal backgrounds of politicians. Most datasets that include biographical information about politicians focus on heads of state (Baturo, 2016; Brambor et al., 2014; Goemans et al., 2009; Ellis et al., 2015) or cabinet members (Alexiadou, 2016; Best and Edinger, 2005; Braun and Raddatz, 2010; Ennser-Jedenastik et al., 2022). Of the few that collect data on legislators, some include only those elected in selected OECD democracies (Best and Edinger, 2005; Dowding and Dumont, 2009; Faccio, 2006, 2010; Göbel and Munzert, 2022) while others include more expansive lists of countries but only subsets of lawmakers within them (Nelson, 2014). Other efforts, such as the Global Data on National Parliaments (PARLINE), available through the Inter-Parliamentary Union, provide nationally-aggregated party, gender, and other demographic features of legislators, but not individual-level data (see also Ruedin (2009)). We know of no other dataset that provides virtually complete individual-level biographical data on lawmakers for such a large sample of democracies.

The most comprehensive previous collection to date has been the Global Leadership Project (GLP) (Gerring et al., 2019). The GLP consists of legislator-level data spanning 162 countries, and

was assembled using a questionnaire directed at country experts and collected in two waves, 2010–2013 and 2017–2018. Because we consulted authoritative sources—first and foremost, parliamentary websites—our dataset improves on the GLP collection by containing more precise information on legislators in electoral democracies. The GLD boasts a more accurate count of legislators for each country, minimizes missing data, includes more sourcing information, and is more thoroughly cleaned and standardized across countries, allowing for easy cross-country comparisons. However, it is a smaller dataset at the country level, containing information on 97 democracies and excluding non-democratic countries.

One improvement we make relative to the GLP is how we handle occupational data, a notoriously thorny variable in the study of elites. Whereas survey-based datsets like the GLP often ask legislators to classify themselves into broad occupational categorizations (the GLP uses a 20-category scheme), the GLD includes both raw occupational descriptions collected from official websites and three-digit occupational codes from ISCO-08 (the lawmakers we study fall into 128 unique ISCO categories). Developed under the auspices of the International Labor Organization, ISCO provides very specific occupational information. We initially recorded detailed information about professions (e.g., "industrial engineer," "accountant," "solicitor," and "sales manager of construction materials business"). We then mapped these onto three-digit ISCO codes using the University of Warwick's Computer Assisted Structured Coding Tool (CASCOT); the resulting output was manually reviewed and corrected to further ensure consistency. The resulting data allows researchers to immediately ascertain the economic and social class of legislators without text-based analysis or imputation, whereas the open-ended descriptions allow validation of our team's coding and easily permit alternative coding schemes.

Consistently coding occupations and educational attainments was the most challenging aspect of collecting the dataset. Information about occupations was often difficult to find in public sources. We asked coders to record each legislator's full-time paid, unelected job prior to their first full-time paid elected public office, with political patronage positions and political appointments excluded except where no other occupational data was available. This information was often difficult to find

for several reasons, chief among them that legislators' official websites and other public-facing sites tend to provide occupational information using categories ("business owner," "educator") that are much broader than the three-digit ISCO-08 categories. Coders made heroic efforts to obtain precise information that could be mapped onto ISCO-08 codes.

Educational attainment was also challenging to collect. For each legislator, we attempted to ascertain the final degree completed before the legislator was elected into the office for the parliamentary term in question. It is not always easy to verify an individual's educational attainment from public information, nor is it obvious how to reconcile different degrees across countries. We relied on the International Standard Classification of Education to reconcile differences.

To ensure uniformity in coding, we provided research assistants detailed instructions. In addition, the supervising P.I. met weekly with the team of assistants to review coding questions and discrepancies as they arose. The coding instructions that emerged out of the process of year-long weekly meetings are included in the GLD codebook. Thanks to the ongoing regular interactions among coders, we believe the GLD provides consistency in coding across countries.

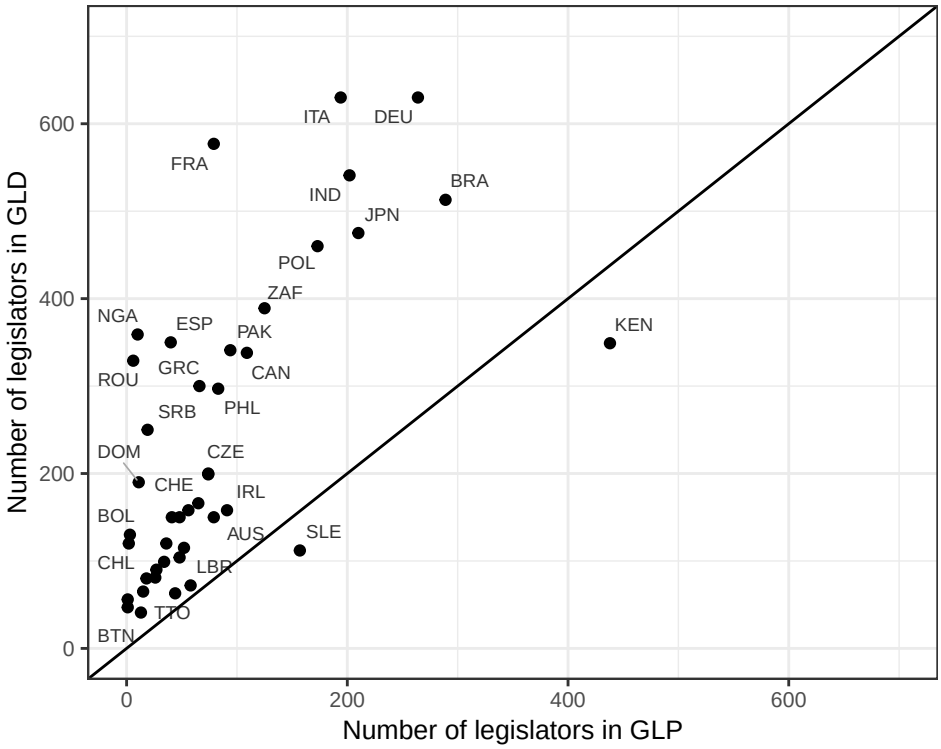
Validity Checks

After compiling the GLD, we conducted multiple validity checks. For many of the traits we study, there are no comparable large-scale cross-national datasets available for validation. Here we focus on two traits that we can compare to existing datasets and for which there is little room for interpretation when coding: the numbers of legislators in the lower chamber or unicameral legislature (to check that the GLD improves on the sample sizes in other individual-level datasets like the Global Leadership Project) and the proportion of women in each legislature (from the popular Varieties of Democracy (V-Dem) project, to check that our biographical data produced estimates comparable to those in a reliable source).

In the individual-level GLP dataset, information is available for lower-chamber legislators in 80 countries that overlap with those in our GLD dataset. Figure 1 plots the numbers of legislators

in the GLP sample and in our dataset for these 80 countries. Countries for which the two datasets have identical numbers of legislators appear on the 45-degree line; countries where our dataset includes more legislators are above it. It is easy to see that the the GLD dataset includes more legislators than the GLP in all but two countries (Kenya and Sierra Leone). In many cases, the differences are substantial. For instance, the GLD contains information on 630 of 631 members of the German lower house whereas the GLP dataset reports data on only 264 members. Likewise, the GLD includes information about all 577 members of the French lower house, while the GLP dataset misses 86.3 percent of French lawmakers.

Figure 1: Numbers of lawmakers included in the GLD and the GLP for 80 countries



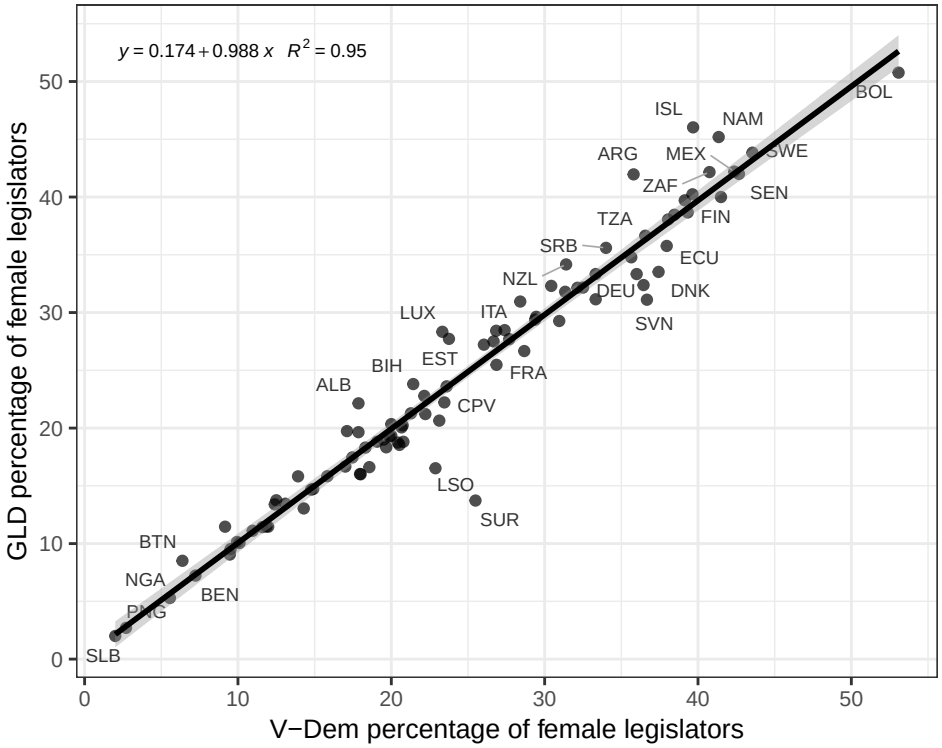
Of course, the two datasets have different aims. The GLP, for instance, contains far more countries than our dataset, including many non-democracies. But this simple exercise illustrates how our dataset contributes depth and completeness that usefully complements the breadth of datasets like the GLP.

Whereas the GLP is the only other dataset to provide individual-level information on legislators in so many countries around the world, some datasets provide country-level information that can

further validate the GLD. One of the most commonly used is the Varieties of Democracy (V-Dem) Project, which collects data from experts on 202 countries (Coppedge et al., 2022). Most V-Dem variables are not comparable to the individual-level data we collected in the GLD, but V-Dem and the GLD share one variable in common: women's representation.

Figure 2 plots the relationship between women's representation from the country-level V-Dem dataset (Coppedge et al., 2022, 159) and country estimates from our individual-level dataset. As the figure illustrates, the data from the two sources are almost identical. A simple bivariate regression produces an R^2 of 0.95 and an intercept of less than one percentage point. (In other words, the two datasets' estimates differ from one another by less than one fifth of one percentage point on average.) The 45-degree line is not visible in Figure 2 because the best fit line sits directly on top of it.

Figure 2: Shares of women lawmakers in the GLD and V-Dem dataset for 93 countries



This finding provides reassuring validation of our data collection effort. V-Dem is a well-validated, widely-used, expert-driven dataset. The data we have collected, using entirely different methods and sources, comes close to replicating the V-Dem estimates of women's representation.

Application: Reelection Rates and Politician Characteristics

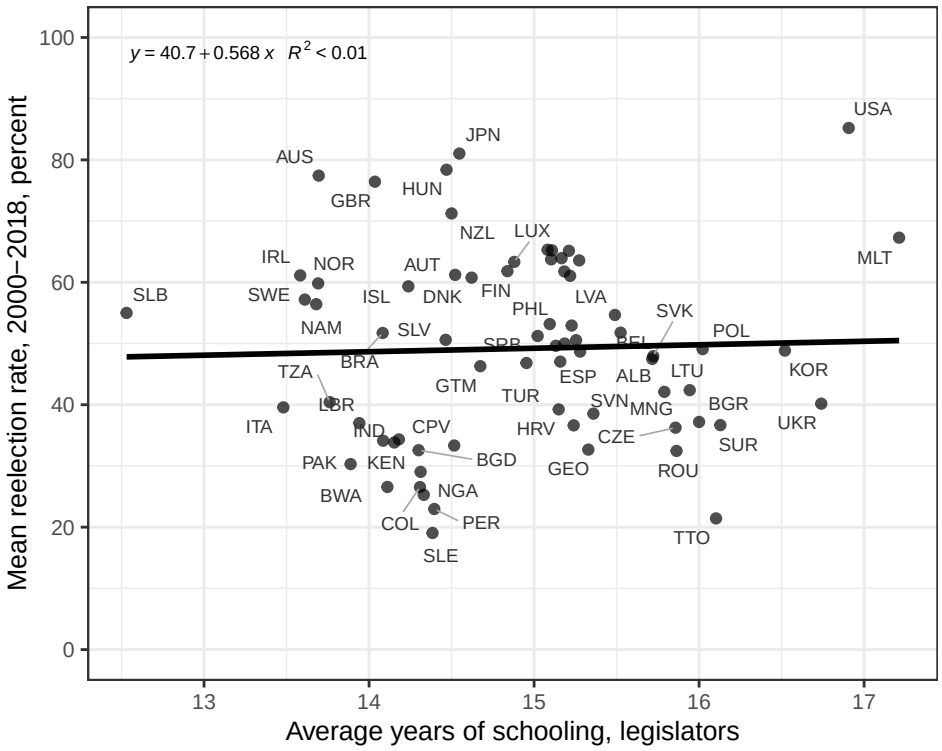
What can researchers do with the GLD dataset? There are numerous potential applications that widen the field of questions scholars can now answer with precise and complete data. Because the dataset is individual-level, it can be merged with other individual-, country-, or party-level datasets. It can be used to answer questions about the causes or consequences of the numerical or descriptive representation of women, representatives with more and less formal education, older and younger legislators, and those from different occupational backgrounds. It can be used to provide controls for studies focused primarily on other variables but for which representational issues might be potential confounds. And, most fundamentally, it can answer simple descriptive questions about which scholars have previously only been able to speculate.

As an illustration of how researchers might use the GLD, we matched GLD data to data available from the Reelection Rates in Democracies Around the World Dataset, compiled by Miriam Golden and Eugenia Nazrullaeva (Golden and Nazrullaeva, 2022). The reelection rates data allowed us to determine whether each legislator in the GLD had also held office in the immediately preceding legislative term for legislator-level data in the 67 country-level observations that matched the GLD.[5]

Using the new merged dataset, we first ask if reelection rates are higher in countries where lawmakers have more formal education. Many scholars equate educational attainment with skill or ability (e.g., Bovens and Wille, 2017; Besley et al., 2011; Besley and Reynal-Querol, 2011; Hallerberg and Wehner, 2013), so we might hypothesize that countries where lawmakers have more formal education would have less legislative turnover. In Figure 3, however, a simple scatterplot of the country-level relationship between schooling and reelection rates does not support this idea. The data show an extremely modest positive association. Of course this simple analysis is not causally-identified, but the figure does not provide any prima facie corroboration of the hypothesis (see also Carnes and Lupu, 2016).

[5]For some countries, the legislature in the GLD dataset is the first legislature available in the reelection datatset, i.e., we have data for the following legislative term but not the previous. We exclude these countries from the analysis of legislators' reelection rates since we cannot code the incumbency status of legislators.

Figure 3: Reelection rates of lawmakers by years of education for 70 countries

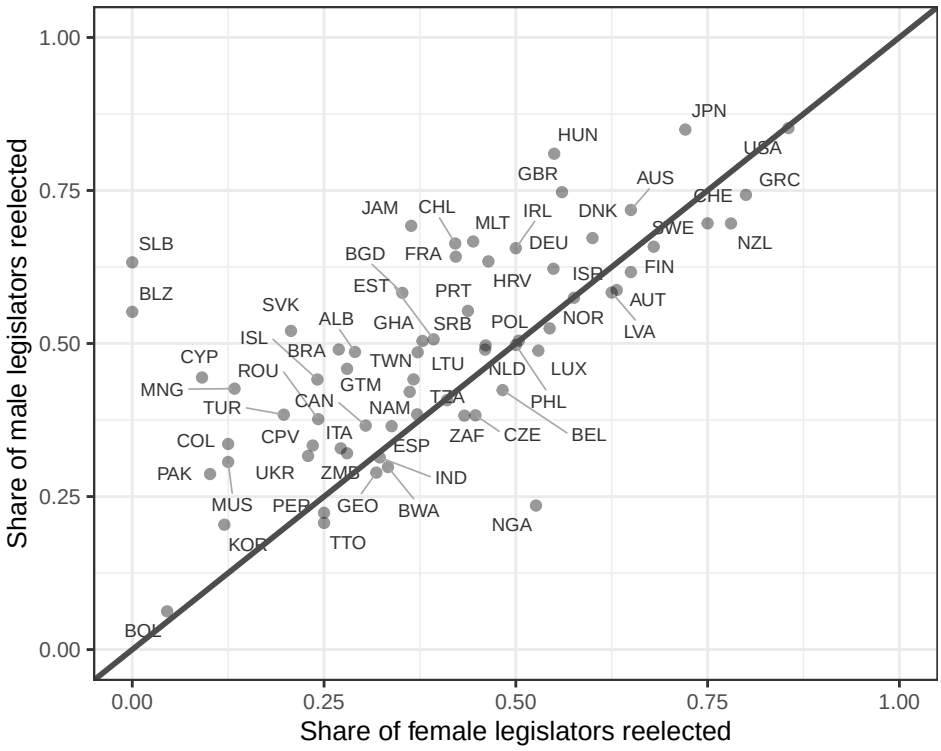


We next ask whether men and women in legislatures are reelected at different rates. In Figure 4, we present data that shows clear evidence of a difference. The figure plots the average reelection rate for men (vertical axis) against the average reelection rate for women (horizontal axis); thus, in countries above the 45-degree line, men are reelected at higher rates than women. Most countries are above the 45-degree line, showing that men experience a relatively consistent advantage in securing reelection. The phenomenon is not universal; in 22 countries, women are reelected at higher rates than men. Why this might be remains an open question.

Researchers are increasingly interested in explaining why there are so few working-class politicians in national public office around the world (Carnes and Lupu, 2023). One possible contributing factor could be that even if they do get elected, working-class politicians find it harder to get re-elected. In Figure 5, we plot reelection rates among legislators from working-class occupations[6]

[6]We define working class as individuals who last worked in manual labor, service industry, clerical, or labor union jobs, which comprise ISCO-08 categories 4 through 9. We remove farm owners and managers and police officers and we include labor union employees not covered in other categories.

Figure 4: Reelection rates of male and female lawmakers for 68 countries

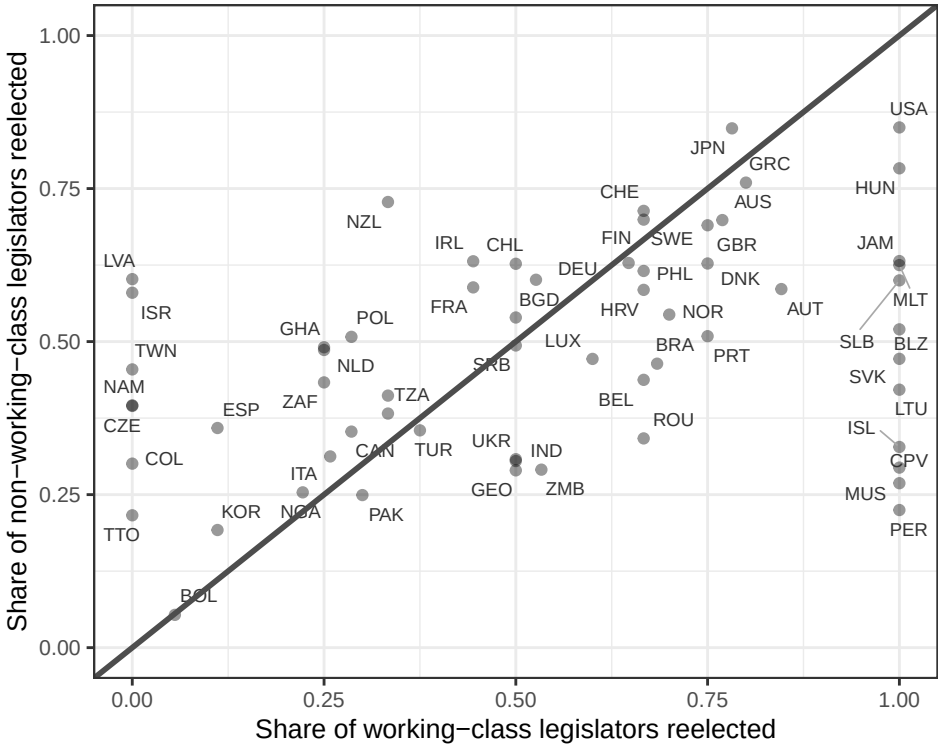


against reelection rates of legislators who did not hold working-class jobs prior to their first election into public office.

The data displayed in Figure 5 offer no support for the hypothesis that lower reelection rates help explain the shortage of working-class politicians. Reelection rates for working-class incumbents are far more varied than for their non-working class counterparts, perhaps because in most legislatures there are only a handful of working-class incumbents and in many there are none at all. Even for the few working class members of lower/unitary chambers, no clear patterns emerge. Countries are about as likely to be above the 45-degree line (non-working-class legislators are reelected more often than working-class legislators) as they are to be below it (working-class legislators are reelected more often than non-working-class legislators).

Taken together, these simple comparisons suggest that, among underrepresented groups, women face unique hurdles in securing reelection. We find no consistent crossnational evidence that lawmakers from working-class jobs and lawmakers with less formal education fare worse in future races once they gain initial entry into elected office. But in most countries, women who

Figure 5: Reelection rates among lawmakers from working-class and non-working-class occupations for 56 countries



make it into the national legislature still face disadvantages when they seek reelection.

Opportunities for Research

The figures we have presented are only illustrative and correlational rather than causal. However, they highlight just a few potential uses of the Global Legislators Database and exemplify how much we can learn even from simple descriptive analyses.

When the GLD is released, we hope that other researchers will take advantage of it. The GLD data can be aggregated to the political party or country level, or used to answer individual-level questions about legislators. It has the potential to shed light on questions about the causes and consequences of the numerical or descriptive representation of social groups defined by gender identity, educational level, age, and past occupation.

We hope other researchers will use the GLD to take up these and other questions. And we hope

future researchers will help us collect the next wave of the GLD.

References

Alexiadou, Despina. 2016. *Ideologues, Partisans, and Loyalists: Ministers and Policymaking in Parliamentary Cabinets*. Oxford University Press.

Baturo, Alexander. 2016. "Cursus Honorum: Personal Background, Careers and Experience of Political Leaders in Democracy and Dictatorship—New Data and Analyses." *Politics and Governance* 4(2): 138-157.

Besley, Timothy, Jose G. Montalvo, and Marta Reynal-Querol. 2011. "Do Educated Leaders Matter?" *The Economic Journal* 121(554): F205–227.

Besley, Timothy, and Marta Reynal-Querol. 2011. "Do Democracies Select More Educated Leaders?" *American Political Science Review* 105(3): 552–66.

Best, Heinrich, and Michael Edinger. 2005. "Converging Representative Elites in Europe? An Introduction to the EurElite Project." *Sociologický Časopis / Czech Sociological Review* 41(3): 499–510.

Bovens, Mark, and Anchrit Wille. 2017. *Diploma Democracy: The Rise of Political Meritocracy*. Oxford: 0xford University Press.

Brambor, Thomas, Johannes Lindvall, and Annika Stjernquist. 2014. "The Ideology of Heads of Government (HOG), 1870–2012: Codebook, Version 1.2.".

Braun, Matías, and Claudio Raddatz. 2010. "Banking on Politics." *World Bank Economic Review* 24(2): 234–279.

Carnes, Nicholas, and Noam Lupu. 2016. "What Good is a College Degree? Education and Leader Quality Reconsidered." *Journal of Politics* 78(1): 35–49.

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

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Nazifa Alizada, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lina Gastaldi, Haakon Gjerløw, Adam Glynn, Sandra Grahn, Allen Hicken, Nina Ilchenko, Katrin Kinzelbach, Joshua Krusell, Kyle L. Marquardt, Kelly M. McMann, Valeriya Mechkova, Juraj Medzihorsky, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Oskar Ryden, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Yiting Wang, Tore Wig, Steven Wilson, and Daniel. Ziblatt. 2022. V-Dem Codebook v12. Working paper V-Dem Institute, University of Gothenburg.

Dowding, Keith and Patrick Dumont, eds. 2009. *The Selection of Ministers in Europe: Hiring and Firing*. Routledge.

Ellis, Cali, Michael Horowitz, and Allan Stam. 2015. "Introducing the LEAD data set." *International Interactions* 41: 718-741.

Ennser-Jedenastik, Laurenz, Florence Ecormier-Nocca, Oliver Huwyler, Jeanne Marlier, and Manu Wagner. 2022. "The 'De-party-politicization' of Europe's Political Elites: How the Rise of Technocrats and Political Outsiders Transforms Representative Democracy'): Codebook, Version 0.1.".

Faccio, Mara. 2006. "Politically Connected Firms." *American Economic Review* 96(1): 369–386.

Faccio, Mara. 2010. "Differences between Politically Connected and Nonconnected Firms: A Cross-Country Analysis." *Financial Management* 39(3): 905–928.

Gerring, John, Erzen Oncel, Kevin Morrison, and Daniel Pemstein. 2019. "Who Rules the World? A Portrait of the Global Leadership Class." *Perspectives on Politics* 17(4): 1079–1097.

Göbel, Sascha, and Simon Munzert. 2022. "The Comparative Legislators Database." *British Journal of Political Science* 52(3): 1398–1408.

Goemans, Henk E., Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46(2): 269-283.

Golden, Miriam, and Eugenia Nazrullaeva. 2022. "Reelection Rates in Democracies Around the World Dataset.".
URL: *https://doi.org/10.7910/DVN/LKN32V*

Gulzar, Saad. 2021. "Who Enters Politics and Why?" *Annual Review of Political Science*(24): 253–75.

Hallerberg, Mark, and Joachim Wehner. 2013. "The Educational Competence of Economic Policymakers in the EU." *Global Policy* 3(s1): 9–15.

Krcmaric, Daniel, Stephen C. Nelson, and Andrew Roberts. 2020. "Studying Leaders and Elites: The Personal Biography Approach." *Annual Review of Political Science*(23): 133–51.

Nelson, Stephen C. 2014. "Playing Favorites: How Shared Beliefs Shape the IMF's Lending Decisions." *International Organization* 68(2): 297–328.

Ruedin, Didier. 2009. "Ethnic Group Representation in a Cross-National Comparison." *The Journal of Legislative Studies* 15(4): 335–354.

Wängnerud, Lena. 2009. "Women in Parliaments: Descriptive and Substantive Representation." *Annual Review of Political Science*(12): 51–69.

Appendix A: Countries Included

Table A-1: Summary Information on Countries Included

Country	Country Code	Election Year	Period	Leg No	MPs
Albania	ALB	2013	2013-2017	30	140
Argentina	ARG	2015	2015-2019	–	255
Australia	AUS	2016	2016-2019	45	150
Austria	AUT	2013	2013-2017	25	183
Bahamas	BHS	2012	2012-2017	13	38
Bangladesh	BGD	2014	2014-2018	10	349
Belgium	BEL	2014	2014-2019	54	150
Belize	BLZ	2015	2015-2020	–	31
Benin	BEN	2015	2015-2019	–	83
Bhutan	BTN	2013	2013-2018	2	47
Bolivia	BOL	2014	2014-2019	–	130
Bosnia Herzegovina	BIH	2014	2014-2018	7	42
Botswana	BWA	2014	2014-2019	11	63
Brazil	BRA	2014	2014-2018	55	513
Bulgaria	BGR	2014	2014-2017	43	240
Canada	CAN	2015	2015-2019	42	338
Cape Verde	CPV	2016	2016-2021	–	72
Chile	CHL	2013	2013-2017	53	120
Colombia	COL	2014	2014-2018	–	166
Costa Rica	CRI	2014	2014-2018	–	57
Cote d’Ivoire	CIV	2011	2011-2016	–	253
Croatia	HRV	2016	2016-2020	–	151

Table A-1: Summary Information on Countries Included

(continued)

Country	Country Code	Election Year	Period	Leg No	MPs
Cyprus	CYP	2016	2016-2021	–	56
Czech Republic	CZE	2013	2013-2017	7	200
Denmark	DNK	2015	2015-2019	–	179
Dominican Republic	DOM	2016	2016-2020	–	190
East Timor	TLS	2012	2012-2017	3	65
Ecuador	ECU	2017	2017-2021	–	137
El Salvador	SLV	2015	2015-2018	11	84
Estonia	EST	2015	2015-2019	13	101
Fiji	FJI	2014	2014-2018	6	50
Finland	FIN	2015	2015-2019	–	200
France	FRA	2012	2012-2017	14	577
Georgia	GEO	2016	2016-2020	9	150
Germany	DEU	2013	2013-2017	18	630
Ghana	GHA	2016	2016-2020	7	275
Greece	GRC	2015	2015-2019	–	300
Guatemala	GTM	2015	2015-2019	–	158
Guyana	GUY	2015	2015-2020	11	65
Hungary	HUN	2014	2014-2018	–	199
Iceland	ISL	2013	2013-2016	–	63
India	IND	2014	2014-2019	16	541
Ireland	IRL	2016	2016-2020	32	158
Israel	ISR	2015	2015-2019	20	120
Italy	ITA	2013	2013-2018	17	630

Table A-1: Summary Information on Countries Included

(continued)

Country	Country Code	Election Year	Period	Leg No	MPs
Jamaica	JAM	2016	2016-2020	12	63
Japan	JPN	2014	2014-2017	47	475
Kenya	KEN	2013	2013-2017	11	349
Kosovo	–	2014	2014-2017	6	120
Latvia	LVA	2014	2014-2018	–	100
Lesotho	LSO	2017	2017-2022	10	115
Liberia	LBR	2011	2011-2017	53	72
Lithuania	LTU	2016	2016-2020	12	141
Luxembourg	LUX	2013	2013-2018	–	60
Madagascar	MDG	2013	2013-2019	–	155
Malta	MLT	2013	2013-2017	12	69
Mauritius	MUS	2014	2014-2019	9	70
Mexico	MEX	2015	2015-2018	63	500
Moldova	MDA	2014	2014-2019	9	101
Mongolia	MNG	2016	2016-2020	5	76
Montenegro	MNE	2016	2016-2020	–	81
Namibia	NAM	2014	2014-2019	6	104
Netherlands	NLD	2017	2017-2021	–	150
New Zealand	NZL	2014	2014-2017	51	120
Nigeria	NGA	2015	2015-2019	8	359
Norway	NOR	2013	2013-2017	164	169
Pakistan	PAK	2013	2013-2018	14	341
Panama	PAN	2014	2014-2019	–	71

Table A-1: Summary Information on Countries Included

(continued)

Country	Country Code	Election Year	Period	Leg No	MPs
Papua New Guinea	PNG	2012	2012-2017	9	111
Paraguay	PRY	2013	2013-2018	–	80
Peru	PER	2016	2016-2020	–	130
Philippines	PHL	2016	2016-2019	17	297
Poland	POL	2015	2015-2019	8	460
Portugal	PRT	2015	2015-2019	13	230
Romania	ROU	2016	2016-2020	8	329
Senegal	SEN	2012	2012-2017	12	150
Serbia	SRB	2016	2016-2020	–	250
Sierra Leone	SLE	2012	2012-2018	–	112
Slovak Republic	SVK	2016	2016-2020	–	150
Slovenia	SVN	2014	2014-2018	7	90
Solomon Islands	SLB	2014	2014-2019	10	50
South Africa	ZAF	2014	2014-2019	26	389
South Korea	KOR	2016	2016-2020	20	300
Spain	ESP	2016	2016-2019	12	350
Suriname	SUR	2015	2015-2020	7	51
Sweden	SWE	2014	2014-2018	–	349
Switzerland	CHE	2015	2015-2019	50	199
Taiwan	TWN	2016	2016-2020	9	113
Tanzania	TZA	2015	2015-2020	10	382
Trinidad and Tobago	TTO	2015	2015-2020	11	41
Tunisia	TUN	2014	2014-2019	1	217

Table A-1: Summary Information on Countries Included *(continued)*

Country	Country Code	Election Year	Period	Leg No	MPs
Turkey	TUR	2015	2015-2018	26	550
Ukraine	UKR	2014	2014-2019	8	418
United Kingdom	GBR	2015	2015-2017	56	650
United States	USA	2016	2016-2018	115	441
Uruguay	URY	2014	2014-2019	48	99
Zambia	ZMB	2016	2016-2021	12	156

Appendix B: GLD Codebook

Global Legislator Database (GLD) Codebook

2022-12-09; Version 1.2

Description

Cross-sectional data of biographical characteristics of legislators in 97 democracies across the world. Primary characteristics collected include date of birth, gender, party, occupation, and education. Data is for the lower house or unicameral legislature, and is captured for a single legislature in each country that was elected between 2011 and 2017.

Principal Investigators

Nick Carnes, Duke University

Miriam Golden, European University Institute and University of California at Los Angeles (emerita)

Noam Lupu, Vanderbilt University

Eugenia Nazrullaeva, University of Glasgow

Contact

Nick Carnes: email: `nicholas.carnes@duke.edu`; homepage https://people.duke.edu/~nwc8/

Miriam Golden: email `golden@ucla.edu`; homepage https://www.miriamgolden.com/

Noam Lupu: email `noam.lupu@vanderbilt.edu`; homepage https://www.noamlupu.com/

GLD web-page: https://doi.org/10.7910/DVN/U1ZNVT

Codebook Authors

Joshua Ferrer, University of California at Los Angeles

Miriam Golden, European University Institute and University of California at Los Angeles (emerita)

Project Managers

Esme Lillywhite, SciencesPo

Maggie Dechert, Vanderbilt University

Programming Manager

Joshua Ferrer, University of California at Los Angeles

How to Cite this Dataset

Carnes, Nick; Golden, Miriam; Lupu, Noam; and Nazrullaeva, Eugenia. 2021. "Global Legislator Database," https://doi.org/10.7910/DVN/U1ZNVT, Harvard Dataverse, V1.

Contents

List of Tables

Acknowledgments

In addition to our thanks to Maggie Dechert for project management, we acknowledge Arjun Aurora, David Castro, Caelan Collins, Mary Coyne, Sridhar Dhara, Caleb Harper, Mo Se Kim, Melissa Lee, Amanda Llewellen, Alexandra Lynn, Morgan Marquez, Sigrid Ohnesorge, Nicholas Rudd, Ben Seitelman, Tanya Wadhawan, Chris Westfall, Anirudh Yedula, and especially Eui Young (Emily) Noh for their invaluable research assistance in the first phase of data collection. We also thank Esme Lillywhite for project management in the second phase, as well as Jonne Kamphorst for stepping in to cover for Esme, Joshua Ferrer for programming management, Ivan Formichev for additional coding, and for their extraordinary efforts in collecting and verifying the dataset, Alejandro Tirado Castro, Armel François, Mamoor Ali Khan, Eui Young (Emily) Noh, Karan Sanghavi, and Giovani Bastiani Roggia.

For access to the CASCOT software, we thank Professor Peter Elias and the Warwick Institute for Employment Research.

Government and other offices and individuals who provided country-specific data are acknowledged by country.

Funding

We received funding from the Academic Senate of the University of California at Los Angeles, the European University Institute, the Sanford School of Public Policy at Duke University, Vanderbilt University, a Vanderbilt University Office of Equity, Diversity, and Inclusion Seed Grant, and a Research Fund Award from the London School of Economics and Political Science, Department of Government.

Introduction

Aim of the research project

The aim of this project was to collect a cross-sectional of biographical characteristics for national legislators in every electoral democracy in the world. In total, we collected data on 97 countries, totaling 19,704 legislators. In most countries, the parliament captured is for the election taking place between 2013 and 2016. The full range of captured elections is 2011 to 2017, corresponding with parliaments in session between 2011 to 2022. Each country has data on one parliament that was elected and seated during this time span.

Variables collected

The main variables collected were `name`, `dob`, `gender`, `party`, `last_occupation` (prior to full-time elected office), and level of `education` attained. Occupations were coded into ILO ISCO-08 codes (`ISCO08`), the most widely accepted occupation classification system, using the University of Warwick's Computer Assisted Structured Coding Tool (CASCOT). `ISCO08` codes were manually entered for occupations that CASCOT was unable to code. Contextual data for each country includes the `year_of_election` and `parliamentary_period` captured, the total number of legislators in the parliament (`total_mps`), the total number of legislators captured in the dataset (`total_mps_in_data`), and the final date of data verification (`date_verified`). Extensive sourcing information was captured for each country and legislator as well.

Release Notes

- Version 1

General instructions

Country selection criteria

We initially aimed to include all 103 electoral democracies with populations over 300,000, as defined by Freedom House for 2016 (https://freedomhouse.org/sites/default/files/FH_FITW_Report_2016.pdf). Comoros, Indonesia, Malawi, Nepal, Niger, and Sri Lanka were not included in the final dataset because very little occupation or education data could be located for them. This left us with 97 countries in the final dataset.

Countries included

Table 1: Summary of Countries Included

Country	Year of Election	Parliamentary Period	Legislature	MPs
Albania	2013	2013-2017	30	140
Argentina	2015	2015-2019	NA	255
Australia	2016	2016-2019	45	150
Austria	2013	2013-2017	25	183
Bahamas	2012	2012-2017	13	38
Bangladesh	2014	2014-2018	10	349
Belgium	2014	2014-2019	54	150
Belize	2015	2015-2020	NA	31
Benin	2015	2015-2019	NA	83
Bhutan	2013	2013-2018	2	47
Bolivia	2014	2014-2019	NA	130
Bosnia Herzegovina	2014	2014-2018	7	42
Botswana	2014	2014-2019	11	63
Brazil	2014	2014-2018	55	513
Bulgaria	2014	2014-2017	43	240

Table 1: Summary of Countries Included *(continued)*

Country	Year of Election	Parliamentary Period	Legislature	MPs
Canada	2015	2015-2019	42	338
Cape Verde	2016	2016-2021	NA	72
Chile	2013	2013-2017	53	120
Colombia	2014	2014-2018	NA	166
Costa Rica	2014	2014-2018	NA	57
Cote d'Ivoire	2011	2011-2016	NA	253
Croatia	2016	2016-2020	NA	151
Cyprus	2016	2016-2021	NA	56
Czech Republic	2013	2013-2017	7	200
Denmark	2015	2015-2019	NA	179
Dominican Republic	2016	2016-2020	NA	190
East Timor	2012	2012-2017	3	65
Ecuador	2017	2017-2021	NA	137
El Salvador	2015	2015-2018	11	84
Estonia	2015	2015-2019	13	101
Fiji	2014	2014-2018	6	50
Finland	2015	2015-2019	NA	200
France	2012	2012-2017	14	577
Georgia	2016	2016-2020	9	150
Germany	2013	2013-2017	18	630
Ghana	2016	2016-2020	7	275
Greece	2015	2015-2019	NA	300

Table 1: Summary of Countries Included *(continued)*

Country	Year of Election	Parliamentary Period	Legislature	MPs
Guatemala	2015	2015-2019	NA	158
Guyana	2015	2015-2020	11	65
Hungary	2014	2014-2018	NA	199
Iceland	2013	2013-2016	NA	63
India	2014	2014-2019	16	541
Ireland	2016	2016-2020	32	158
Israel	2015	2015-2019	20	120
Italy	2013	2013-2018	17	630
Jamaica	2016	2016-2020	12	63
Japan	2014	2014-2017	47	475
Kenya	2013	2013-2017	11	349
Kosovo	2014	2014-2017	6	120
Latvia	2014	2014-2018	NA	100
Lesotho	2017	2017-2022	10	115
Liberia	2011	2011-2017	53	72
Lithuania	2016	2016-2020	12	141
Luxembourg	2013	2013-2018	NA	60
Madagascar	2013	2013-2019	NA	155
Malta	2013	2013-2017	12	69
Mauritius	2014	2014-2019	9	70
Mexico	2015	2015-2018	63	500
Moldova	2014	2014-2019	9	101
Mongolia	2016	2016-2020	5	76

Table 1: Summary of Countries Included *(continued)*

Country	Year of Election	Parliamentary Period	Legislature	MPs
Montenegro	2016	2016-2020	NA	81
Namibia	2014	2014-2019	6	104
Netherlands	2017	2017-2021	NA	150
New Zealand	2014	2014-2017	51	120
Nigeria	2015	2015-2019	8	359
Norway	2013	2013-2017	164	169
Pakistan	2013	2013-2018	14	341
Panama	2014	2014-2019	NA	71
Papua New Guinea	2012	2012-2017	9	111
Paraguay	2013	2013-2018	NA	80
Peru	2016	2016-2020	NA	130
Philippines	2016	2016-2019	17	297
Poland	2015	2015-2019	8	460
Portugal	2015	2015-2019	13	230
Romania	2016	2016-2020	8	329
Senegal	2012	2012-2017	12	150
Serbia	2016	2016-2020	NA	250
Sierra Leone	2012	2012-2018	NA	112
Slovak Republic	2016	2016-2020	NA	150
Slovenia	2014	2014-2018	7	90
Solomon Islands	2014	2014-2019	10	50
South Africa	2014	2014-2019	26	389

Table 1: Summary of Countries Included *(continued)*

Country	Year of Election	Parliamentary Period	Legislature	MPs
South Korea	2016	2016-2020	20	300
Spain	2016	2016-2019	12	350
Suriname	2015	2015-2020	7	51
Sweden	2014	2014-2018	NA	349
Switzerland	2015	2015-2019	50	199
Taiwan	2016	2016-2020	9	113
Tanzania	2015	2015-2020	10	382
Trinidad and Tobago	2015	2015-2020	11	41
Tunisia	2014	2014-2019	1	217
Turkey	2015	2015-2018	26	550
Ukraine	2014	2014-2019	8	418
United Kingdom	2015	2015-2017	56	650
United States	2016	2016-2018	115	441
Uruguay	2014	2014-2019	48	99
Zambia	2016	2016-2021	12	156

Legislator selection criteria

We include legislators who were elected at the general election to their country's national parliament or, in countries with bicameral legislatures, to the lower chamber. In countries with both national and devolved regional governments, only data on the national legislature is captured. In some countries, non-voting delegates are also included.

The dataset does not include legislators who were elected, appointed, or otherwise acceded

to office after the general election. All mid-cycle and replacement legislators are thereby excluded. We also do not capture information on substitute or deputy legislators. Elected MPs who were never seated are included but MPs who replaced an elected legislator after the main election but before the start of the parliamentary term are excluded.

For each country, research assistants attempted to locate the most accurate list of members of the target parliament. This typically involved triangulating data from the country's parliamentary website, the country's electoral or judicial commission, and Wikipedia (sometimes in the country language). Research assistants also used the PARLINE database on national parliaments (http://archive.ipu.org/parline-e/parlinesearch.asp).

General rules for coding missing data

Missing legislators

In a some countries, we failed to locate a canonical list of elected legislators. This occasionally resulted in small discrepancies between the officially-recorded size of the legislature and the number of MPs recorded in the dataset. These discrepancies are listed in Table 2.

Some discrepancies are due to vacant seats in which no person was put forward for election in certain areas. In Ukraine, several dozen vacancies were never filled so `total_mps` and `total_mps_in_data` reflect the number of filled seats rather than the official size of the legislature (418 rather than 450). In Sierra Leone, three initial vacancies were created due to legal reasons but filled days after the initial election. These three seats are included in the dataset and with `total_mps`. Similarly, four initially vacant seats were soon filled in Madagascar and are included in the dataset and with `total_mps`. Lesotho and Pakistan also include initial vacancies that were soon filled.

We also conducted post-data collection checks for potential duplicates and eliminated these where identified.

Table 2: Summary of Missing MPs

Country	Parliamentary Period	Total MPs	MPs in data	Missing
Argentina	2015-2019	257	255	2
Bangladesh	2014-2018	350	349	1
Cote d'Ivoire	2011-2016	255	253	2
India	2014-2019	545	541	4
Lesotho	2017-2022	117	115	2
Liberia	2011-2017	73	72	1
Nigeria	2015-2019	360	359	1
Pakistan	2013-2018	342	341	1
South Africa	2014-2019	400	389	11
Switzerland	2015-2019	200	199	1

Missing occupation and education data

While we made efforts to reduce missingness as much as possible, we took additional measures for countries with high levels of missingness with the aim of reducing missing occupation and education entries below 5 percent. Once we achieved a missingness of 5 percent or less on education and occupation, we ceased data collection. Beyond general online searches, we attempted three additional steps: contacting parliaments, contacting other researchers, and contacting MPs directly.

Contacting parliament: Research assistants contacted the parliamentary secretary/clerk/librarian or equivalent office by email and phone. We used a standardized template, tailored to the specific country. Multiple follow-ups were attempted as well. This method frequently resulted in successful outreach and useful data.

Contacting other researchers: Attempts were also made to directly contact other sources, including the country's electoral commissions, NGOs, and academics who have undertaken similar research. Records of all email correspondence and external datasets collected were retained.

Contacting MPs: As a last resort, research assistants directly contacted MPs via email and social media accounts using a template letter. We rarely received a response from these inquiries.

When countries were considered finished

If the efforts outlined above were undertaken but a country still had missingness above the 5 percent threshold, then we considered data collection for that country finished. Very little useful data could be collected from Comoros, Indonesia, Malawi, Nepal, Niger, and Sri Lanka. We therefore decided to exclude these countries from the dataset.

Table 3: Summary of Missing Data

Country	DOB	%	Gender	%	Party	%	Occupation	%	Education	%
Albania	13	9	0	0	0	0	5	4	6	4
Argentina	0	0	0	0	0	0	15	6	29	11
Australia	0	0	0	0	0	0	1	1	9	6
Austria	0	0	0	0	0	0	1	1	1	1
Bahamas	8	21	0	0	0	0	0	0	0	0
Bangladesh	11	3	0	0	1	0	4	1	16	5
Belgium	2	1	0	0	9	6	3	2	7	5
Belize	11	35	0	0	0	0	5	16	10	32
Benin	77	93	0	0	0	0	58	70	54	65
Bhutan	29	62	0	0	0	0	0	0	1	2

Table 3: Summary of Missing Data *(continued)*

Country	DOB	%	Gender	%	Party	%	Occupation	%	Education	%
Bolivia	0	0	0	0	0	0	0	0	41	32
Bosnia Herzegovina	1	2	0	0	0	0	0	0	0	0
Botswana	52	83	0	0	0	0	17	27	9	14
Brazil	0	0	0	0	0	0	9	2	5	1
Bulgaria	3	1	0	0	0	0	20	8	211	88
Canada	11	3	0	0	0	0	1	0	42	12
Cape Verde	47	65	1	1	0	0	3	4	10	14
Chile	0	0	0	0	0	0	0	0	0	0
Colombia	27	16	0	0	0	0	2	1	4	2
Costa Rica	0	0	0	0	0	0	0	0	1	2
Cote d'Ivoire	230	91	0	0	0	0	3	1	253	100
Croatia	0	0	0	0	0	0	6	4	1	1
Cyprus	1	2	0	0	0	0	0	0	2	4
Czech Republic	0	0	0	0	0	0	3	2	4	2
Denmark	0	0	0	0	0	0	2	1	2	1
Dominican Republic	12	6	0	0	0	0	7	4	10	5
East Timor	0	0	1	2	0	0	12	18	7	11
Ecuador	57	42	0	0	0	0	6	4	9	7
El Salvador	0	0	0	0	0	0	0	0	0	0
Estonia	0	0	0	0	0	0	1	1	0	0
Fiji	35	70	0	0	0	0	8	16	19	38
Finland	0	0	0	0	0	0	0	0	8	4

Table 3: Summary of Missing Data *(continued)*

Country	DOB	%	Gender	%	Party	%	Occupation	%	Education	%
France	1	0	0	0	0	0	9	2	54	9
Georgia	0	0	0	0	0	0	3	2	1	1
Germany	0	0	0	0	0	0	13	2	7	1
Ghana	0	0	0	0	0	0	0	0	0	0
Greece	12	4	0	0	0	0	12	4	11	4
Guatemala	0	0	0	0	1	1	24	15	26	16
Guyana	45	69	0	0	0	0	7	11	14	22
Hungary	0	0	0	0	0	0	9	5	3	2
Iceland	0	0	0	0	0	0	0	0	0	0
India	0	0	1	0	0	0	14	3	2	0
Ireland	1	1	0	0	0	0	8	5	19	12
Israel	0	0	0	0	0	0	0	0	1	1
Italy	1	0	0	0	0	0	10	2	16	3
Jamaica	42	67	0	0	0	0	5	8	14	22
Japan	2	0	0	0	0	0	11	2	0	0
Kenya	25	7	19	5	1	0	18	5	15	4
Kosovo	6	5	0	0	0	0	15	12	6	5
Latvia	0	0	0	0	0	0	1	1	0	0
Lesotho	97	84	64	56	1	1	87	76	86	75
Liberia	66	92	0	0	0	0	0	0	4	6
Lithuania	0	0	0	0	0	0	0	0	0	0
Luxembourg	1	2	0	0	0	0	2	3	10	17
Madagascar	9	6	0	0	0	0	123	79	141	91

Table 3: Summary of Missing Data *(continued)*

Country	DOB	%	Gender	%	Party	%	Occupation	%	Education	%
Malta	16	23	0	0	0	0	4	6	3	4
Mauritius	54	77	0	0	4	6	1	1	6	9
Mexico	84	17	0	0	0	0	15	3	4	1
Moldova	2	2	0	0	0	0	5	5	7	7
Mongolia	60	79	0	0	0	0	1	1	0	0
Montenegro	9	11	3	4	0	0	4	5	5	6
Namibia	6	6	0	0	0	0	6	6	4	4
Netherlands	0	0	0	0	0	0	0	0	3	2
New Zealand	2	2	0	0	0	0	3	2	4	3
Nigeria	34	9	0	0	0	0	11	3	9	3
Norway	0	0	0	0	0	0	13	8	7	4
Pakistan	3	1	0	0	0	0	2	1	1	0
Panama	30	42	0	0	0	0	8	11	14	20
Papua New Guinea	37	33	0	0	0	0	14	13	11	10
Paraguay	19	24	0	0	0	0	12	15	16	20
Peru	0	0	0	0	0	0	0	0	1	1
Philippines	122	41	0	0	0	0	73	25	118	40
Poland	0	0	0	0	0	0	2	0	0	0
Portugal	0	0	0	0	0	0	0	0	1	0
Romania	0	0	0	0	0	0	13	4	7	2
Senegal	82	55	1	1	0	0	149	99	98	65
Serbia	0	0	0	0	0	0	9	4	34	14

Table 3: Summary of Missing Data *(continued)*

Country	DOB	%	Gender	%	Party	%	Occupation	%	Education	%
Sierra Leone	0	0	0	0	0	0	2	2	34	30
Slovak Republic	0	0	0	0	0	0	7	5	6	4
Slovenia	0	0	0	0	0	0	1	1	1	1
Solomon Islands	0	0	0	0	0	0	3	6	1	2
South Africa	221	57	0	0	0	0	227	58	156	40
South Korea	0	0	0	0	0	0	10	3	0	0
Spain	0	0	0	0	0	0	1	0	4	1
Suriname	20	39	0	0	0	0	17	33	20	39
Sweden	0	0	0	0	0	0	29	8	46	13
Switzerland	0	0	0	0	0	0	1	1	19	10
Taiwan	0	0	0	0	0	0	1	1	0	0
Tanzania	20	5	0	0	0	0	17	4	13	3
Trinidad and Tobago	33	80	0	0	0	0	3	7	2	5
Tunisia	3	1	2	1	0	0	4	2	67	31
Turkey	2	0	0	0	0	0	4	1	0	0
Ukraine	0	0	0	0	0	0	6	1	2	0
United Kingdom	0	0	0	0	0	0	5	1	5	1
United States	0	0	0	0	0	0	0	0	0	0
Uruguay	61	62	0	0	0	0	8	8	17	17
Zambia	0	0	0	0	0	0	0	0	1	1

Data collection procedures

General procedures

Data was collected between 2016 and 2021.

We collected data in two phases. Nick Carnes and Noam Lupu oversaw the first phase of the project, which took place in 2016–2017. Miriam Golden oversaw the second phase of data collection, which took place during 2020–2021.

First phase procedures

Noam Lupu and Nick Carnes, along with a team of 18 research assistants, collected information on legislators in 103 democracies between 2016 and 2017. The data collection was supervised by Carnes and Lupu along with Maggie Dechert, a Ph.D. student at Vanderbilt University. Country information was collected by individual research assistants. Emily Noh, another Ph.D. student at Vanderbilt, standardized all individual files, ensuring their columns and titles were the same, all blanks were recorded similarly, and only plain English characters were used. Noh also verified and filled in missing observations in the OECD countries.

Second phase procedures

The Lupu-Carnes data was incomplete when that phase of data collection was ceased. In addition, many sources had not been fully recorded. Over an 18 month period spanning 2020–2021, a team of nine research assistants supervised by Miriam Golden with the assistance of project director Esme Lillywhite verified all existing data from the first phase and added data where missing. The researchers searched online and contacted parliamentary offices, MPs, NGOs, and academics with requests for missing data.

Twenty-five countries in the first phase of collection were carefully verified and thus had a high accuracy rate. These countries are: Australia, Austria, Benin, Cape Verde, Czech Republic, Estonia, Finland, France, Greece, Italy, Japan, Latvia, Liberia, Luxembourg, Mongolia, Netherlands, New Zealand, Norway, Panama, Poland, Slovak Republic, Slovenia, Spain,

Switzerland, and Turkey. In the second phase, less time was spent on verifying these countries than on other countries from the first phase of data collection. Therefore, occupational and educational entries for these 25 countries may less closely follow the standard criteria, especially in regards to the rules for coding `last_occupation`. We undertook more thorough data verification for countries in the list which were found to have a high error rate. This was the case in particular for Austria and Czech Republic.

Research assistants cleaned files for each individual country using coding instructions detailed in the Sections on Sourcing and on List of Variables Coded. Individual standardized files were then prepared for R processing. The column order, names, and number were standardized. Edits were also made to DOBs, duplicate entries, and some occupational, educational, and party data. Source columns and notes outside of the `notes` column were deleted, dots were added to any blank spaces, and `education` was checked to ensure that the correct labels were being used.

Sourcing

We attempted to include specific sourcing information for each data point. Sourcing for specific variables is reflected in the following columns: `name_source`, `party_source`, `dob_source`, `education_source`, and `occupation_source.` Two additional columns house general source data: `source` and `source_2.` Country-specific source data is also listed in Section . `date_verified` marks the time when the collection of these sources concluded, and should be used when attempting to locate the original sources. Data for most countries was collected over a 1–2 month span.

Parliamentary websites were the preferred principal source of data. Additional sources regularly consulted include legislators' personal websites and social media pages, party websites, news and journalism sites, electoral commissions, non-profit organizations, Wikipedia (both English-language and country-specific), and other wiki pages. In addition, the Internet

Archive (Way Back Machine — https://archive.org/web/) was used to capture archived data that had been removed from websites.

We judged the information on some websites as unreliable and thus avoided them when possible. These include EveryPolitician.org, peoplepill.com, and ductum.

Where information could not be acquired online, we attempted to collect data through personal communication with each country's legislative parliament, parties, and judicial branch, and through other scholar's efforts. Information derived from external sources is recorded in the Section on Country Specific Data Collection and Coding Procedures.

List of variables coded

`unique_id`

Each legislator was assigned a randomly-generated unique identification number between 1 and 26,000. Note that legislators from the same country do not always have unique id numbers close to one another.

`country_name`

The country name of the legislature. Only English characters are used, with no accent marks.

`country_code`

Three-digit numeric codes from the United Nations Statistics Division (UNSD) used by the World Bank, equivalent to the International Standards Organization (ISO) three-digit alphabetic codes (see https://wits.worldbank.org/wits/wits/witshelp/content/codes/country_codes.htm and https://wits.worldbank.org/countryprofile/metadata/en/country/all).

`name`

First and last name combined. Only English characters are used, with no accent marks.

first_name

First or given name. Only English characters are used, with no accent marks.

last_name

Last, surname, or family name. Only English characters are used, with no accent marks.

dob

Date of birth, formatted as YYYY-MM-DD. When only the year is known, DOB is written as YYYY-01-01.

gender

Gender coded as a male/female binary.

party

The party the legislator was elected with, recorded using only plain English characters. We verified the party composition of each legislature with multiple sources, where possible. Legislators who are not official members of parties but who were elected through a party list are listed as the party they were elected with.

last_occupation

`last_occupation` captures each legislator's last full-time paid unelected job held prior to their *first* full-time paid elected office. Our goal was to capture each legislator's social class prior to any rise in social standing that resulted from a political career. Full-time paid elected positions could include trade union officials, mayors, city councilors, and party officials. For individuals who cycle in and out of elected office — for instance, being put in party or civil service positions between holding elected office — the rule regarding full-time paid elected

positions still holds. `last_occupation` reflects the last job prior to the legislator's very first full-time elected position.

Additionally, we sought to avoid coding as `last_occupation` political patronage positions or political appointments unless no other occupational data was available. Unpaid political, party, or union jobs and elected part-time political jobs were not coded. Similarly, political jobs held concurrently with non-political jobs were not coded. On the other hand, party and union jobs were coded when they did not entail full-time elected positions. University rectors and pastors were also coded as `last_occupation`.

For example, "campaign manager" is coded since it is typically unelected, even though it is full-time and political in nature. "Party treasurer" is typically not coded, and an appointed cabinet position is also not coded. If the treasurer position is the legislator's first such elected office, their job held immediately prior to the election would be coded. If the appointed cabinet position immediately preceded the legislator's first full-time elected office, then the job held prior to the political appointment would be entered for `last_occupation`.

It was often difficult to discern whether a political position is elected and full-time (for instance, a high ranking position in a union or certain party leadership positions). In this case, research assistants used contextual and country-specific information. Where confusion remained, coding decisions are explained in the `notes` column. In cases where no other occupational data was captured, political occupations were entered for `last_occupation`.

If legislators simultaneously held multiple full-time jobs prior to their first full-time elected political position, then each was listed. If legislators simultaneously held two jobs, then the one the legislator devoted more time too was coded. In most cases, multiple occupations are listed because the chronology or time division was unclear.

Research assistants were instructed to gather as specific occupational data as possible for the purposes of coding occupations into `ISCO08` codes. For instance, rather than simply entering "businessman", research assistants were encouraged to identify the sector of industry and the

exact job title or management level each legislator held. In many cases, however, only vague occupational titles were available.

If someone entered office straight after education, they are coded as a "student". Ph.D. students that went straight into elected office are also coded as students. Legislators were coded as students if gaps in their job history between finishing educational studies and being elected were three years or less.

Where a distinction could not be made between primary, secondary, and high school teachers, educators were simply coded as "teacher (ambiguous)".

In many countries, legislators' education was listed as their occupation on parliamentary documents. Research assistants attempted to verify whether legislators actually held the occupation listed. A record was left in `notes` for particularly confusing cases. Decisions that affected multiple entries is listed in the Section on Country Specific Data Collection and Coding Procedures.

Examples of tricky `last_occupation` cases Student → internal party elected positions → public office: **Code student.**

Student → internal party unelected positions → public office: **Code unelected party position.**

Student → internal party elected positions that is not payed → non-political jobs → public office: **code non-political job.**

Student → internal party elected positions that is full-time (e.g. head of party) → non-political jobs → public office: **code student**

Student → non-political jobs → internal party elected positions that is full-time → public office: **code non-political jobs.**

Student → non-political jobs + internal party elected positions (concurrent) → public office:

code non-political jobs.

Student + social/party activism (concurrent) → public office: **code social/activism.**

Student + non-political job (concurrent) → public office: **code non-political jobs.**

Student + non-political job (concurrent) → unemployed → public office: **code unemployed** (unless very short, then code non-political job).

`education`

`education` captures the highest degree completed before the legislator was elected into office for the coded parliamentary term (as recorded in `year_of_election`). Educational qualifications attained while holding elected office were counted so long as they were awarded prior to the parliamentary term of record. Partially completed degrees are not counted. If a legislator has multiple qualifications, only the highest degree is listed. Research assistants familiarized themselves with each country's education system in order to properly standardize qualifications. Legislator-specific coding decisions are described in the `notes` column and country-specific coding decisions are found in the Section on Country Specific Data Collection and Coding Procedures.

The following international standardized educational categories are used: uneducated, primary, secondary, short-cycle tertiary, bachelors, masters, and Ph.D. Law degrees were distinguished as LLB (equivalent to bachelors), LLM (equivalent to masters), and J.D. (equivalent to Ph.D.). Medical degrees (M.D., equivalent to Ph.D) were also distinguished.

Diplomas, certificates, and "tertiary" degrees are generally counted as short-cycle tertiary. Honorary degrees are not counted. Rather than distinguishing between lower and upper secondary, only primary and secondary degrees are distinguished. Legislators who attain lower secondary qualifications are coded as "primary". This is to attain better accuracy, as information is generally not reliable when distinguishing between lower and upper secondary categories.

Due to a lack of clear-cut dates for education degrees in many countries, there is some inconsistency concerning which degrees are finished prior to each legislator's election to the parliament captured. Where there is ambiguity, listed degrees without dates are counted.

"Postgraduate" degrees are not assumed to be masters, as they could be short-cycle certificate programs. Without additional specificity, legislators with postgraduate degrees are coded as holding bachelors degrees. This ambiguity is especially prevalent in South American countries. "Specialist" degrees in post-Soviet countries are coded as masters, as they typically last about five years. Legislators with incomplete Ph.D.s are coded as holding bachelors, unless there was evidence of obtaining a masters degree. Military degrees are typically coded as equivalent to bachelors degrees.

`notes`

Contains any legislator-specific information that explains legislator inclusion or coding decisions for specific variables.

`file`

The standardized file from which the legislator's data originated.

`ISCO08`

The three-digit version of the International Standard Classification of Occupations (ISCO-08), coded using `last_occupation` entries. Developed under the auspices of the International Labour Organization, ISCO is the most widely accepted classification of occupations. The official classification explanations are documented at https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf

A summary of the three-digit labels is provided in Table 4.

Table 4: ISCO-08 Codes - Official

ISCO Code	Occupation
11	Commissioned Armed Forces Officers
21	Non-commissioned Armed Forces Officers
31	Armed Forces Occupations, Other Ranks
111	Legislators and Senior Officials
112	Managing Directors and Chief Executives
121	Business Services and Administration Managers
122	Sales, Marketing and Development Managers
131	Production Managers in Agriculture, Forestry and Fisheries
132	Manufacturing, Mining, Construction and Distribution Managers
133	Information and Communications Technology Services Managers
134	Professional Services Managers
141	Hotel and Restaurant Managers
143	Other Services Managers
211	Physical and Earth Science Professionals
212	Mathematicians, Actuaries and Statisticians
213	Life Science Professionals
214	Engineering Professionals
216	Architects, Planners, Surveyors and Designers
221	Medical Doctors
222	Nursing and Midwifery Professionals
224	Paramedical Practitioners
225	Veterinarians

Table 4: ISCO-08 Codes - Official *(continued)*

ISCO Code	Occupation
226	Other Health Professionals
231	University and Higher Education Teachers
232	Vocational Education Teachers
233	Secondary Education Teachers
235	Other Teaching Professionals
241	Finance Professionals
242	Administration Professionals
243	Sales, Marketing and Public Relations Professionals
251	Software and Applications Developers and Analysts
252	Database and Network Professionals
261	Legal Professionals
262	Librarians, Archivists and Curators
263	Social and Religious Professionals
264	Authors, Journalists and Linguists
265	Creative and Performing Artists
311	Physical and Earth Science Professionals
312	Mining, Manufacturing and Construction Supervisors
313	Process Control Technicians
314	Life Science Technicians and Related Associate Professionals
315	Ship and Aircraft Controllers and Technicians
321	Medical and Pharmaceutical Technicians
325	Other Health Associate Professionals
331	Financial and Mathematical Associate Professionals

Table 4: ISCO-08 Codes - Official *(continued)*

ISCO Code	Occupation
332	Sales and Purchasing Agents and Brokers
333	Business Services Agents
334	Administrative and Specialized Secretaries
335	Government Regulatory Associate Professionals
341	Legal, Social and Religious Associate Professionals
342	Sports and Fitness Workers
343	Artistic, Cultural and Culinary Associate Professionals
351	Information and Communications Technology Operations and User Support Technicians
352	Telecommunications and Broadcasting Technicians
412	Secretaries (general)
421	Tellers, Money Collectors and Related Clerks
422	Client Information Workers
431	Numerical Clerks
432	Material Recording and Transport Clerks
441	Other Clerical Support Workers
511	Travel Attendants, Conductors and Guides
513	Waiters and Bartenders
514	Hairdressers, Beauticians and Related Workers
516	Other Personal Services Workers
522	Shop Salespersons
523	Cashiers and Ticket Clerks
524	Other Sales Workers

Table 4: ISCO-08 Codes - Official *(continued)*

ISCO Code	Occupation
531	Child Care Workers and Teachers' Aides
532	Personal Care Workers in Health Services
541	Protective Services Workers
611	Market Gardeners and Crop Growers
612	Animal Producers
621	Forestry and Related Workers
622	Fishery Workers, Hunters and Trappers
634	Subsistence Fishers, Hunters, Trappers and Gatherers
711	Building Frame and Related Trades Workers
712	Building Finishers and Related Trades Workers
721	Sheet and Structural Metal Workers, Moulders and Welders, and Related Workers
722	Blacksmiths, Toolmakers and Related Trades Workers
723	Machinery Mechanics and Repairers
731	Handicraft Workers
732	Printing Trades Workers
741	Electrical Equipment Installers and Repairers
742	Electronics and Telecommunications Installers and Repairers
751	Food Processing and Related Trades Workers
753	Garment and Related Trades Workers
754	Other Craft and Related Workers
811	Mining and Mineral Processing Plant Operators
812	Metal Processing and Finishing Plant Operators
815	Textile, Fur and Leather Products Machine Operators

Table 4: ISCO-08 Codes - Official *(continued)*

ISCO Code	Occupation
816	Food and Related Products Machine Operators
817	Wood Processing and Papermaking Plant Operators
831	Locomotive Engine Drivers and Related Workers
832	Car, Van and Motorcycle Drivers
833	Heavy Truck and Bus Drivers
921	Agricultural, Forestry and Fishery Labourers
931	Mining and Construction Labourers
932	Manufacturing Labourers
933	Transport and Storage Labourers
952	Street Vendors
962	Other Elementary Workers

We developed specific classifications and new codes for occupations not clearly specified in the ISCO-08 official documentation. This was necessary to standardize the dataset, especially due to the often vague nature of the information collected for individual legislators. Six codes were created to specify certain vague occupations: unspecified military personnel (10), science researchers (210), teachers (230), social science researchers (240), IT specialists (250), and science research assistants (310). Three additional codes were created for our purposes: student (990), unemployed (991), and retired (992). All `ISCO08` coding clarifications and additions are listed in the Table 5.

Table 5: ISCO-08 Codes - PI Clarifications and Additions

ISCO Code	Occupation
10	(unspecified) Military personnel
111	Political (party/municipal/ward) secretary, Head of party committee
111	Minister, Council member, Commissioner
111	Senior civil servant
111	Head of government agency/department/board/commission/office/registry/program
111	Think tank director
111	Government director
111	Public/government/municipal administrator
111	Director of charity organization
111	Union leader/secretary/official
111	Director of legislator's office
111	Founder of special-interest organization
112	Businessman (unspecified), entrepreneur
112	Director (unspecified), company director, regional director/manager, executive
112	Director of business (specified)
112	Director general, deputy director general, general director, managing director, manager dir
112	Director/business director of multiple businesses or conglomerate
112	Owner of multiple businesses
112	Bank director/executive/governor
112	Head/director/chair of cooperative
121	Small business managing director
121	Department/office/branch/local director or manager of enterprises

Table 5: ISCO-08 Codes - PI Clarifications and Additions

(continued)

ISCO Code	Occupation
121	Project manager/director
121	Business founder/owner/director
121	Manager (unspecified), (party/regional) Coordinator
121	Real estate developer
121	Union manager
121	Treasurer
122	Development/communications manager, editor-in-chief
122	Commercial director, sales director
122	Account executive/director/manager
131	Plantation owner, head of a farm, director of a farm, farm owner
132	Factory owner/director
133	Technical director
134	Hospital directors, health administrator/coordinator
134	Hospital/school/university/museum administrator
134	Head of department in universities/hospitals
134	Bank manager
142	Shop/dealership owner
210	Science researcher (ambiguous)
213	Chemical company employee
214	Engineer
214	Materials engineer/counselor
216	Architecture company emplmoyee

Table 5: ISCO-08 Codes - PI Clarifications and Additions

(continued)

ISCO Code	Occupation
226	Chemical hazardous materials specialist
230	Teacher
235	Education officer, Housemaster
240	Social Science Researcher
241	Banking (ambiguous), investor
241	Chamber of commerce employee
242	Administrator, administrative officer, administrative
242	Advisor, congressional staffer, Chief of Staff
242	Project coordinator, project officer
242	Consultant, mediator
242	Government delegate, Foreign service officer
242	Human resources
242	Human resources
242	Union administrative
243	Campaign organizer/director, party spokesperson, representative
242	Business consultant/analyst
242	Company employee, company office worker, corporate employee
243	Activist for specific organization
243	Sales (ambiguous), Comms (ambiguous), HR (ambiguous)
250	IT consultant/professional/specialist
261	Law firm founder/partner/senior associate

Table 5: ISCO-08 Codes - PI Clarifications and Additions

(continued)

ISCO Code	Occupation
261	General/assistant counsel
263	Religious/Church leader
264	Publisher
265	Public/motivational speaker
310	Science research assistant
311	Chemical lab technician/assistant
312	Foreman
313	Gas enterprise/company employee or worker
313	Control agent
321	Employee in pharmaceutical industry
331	Financial/securities trader
332	Petroleum trader
333	Band manager
334	Policy/legislative/political assistants, political aide
334	Secretary for high-ranking official or department, executive secretary, legal secretary
334	Project assistant
334	Baggage handling coordinator
335	Bureaucrat, Civil servant/administrator
335	(Tax/government/hotel/financial/fiscal) Inspector, civil registry officer
341	Social science research assistants
343	Culinary instructor

Table 5: ISCO-08 Codes - PI Clarifications and Additions

(continued)

ISCO Code	Occupation
351	IT service worker
412	Assistant
422	Airport staff
422	Hotel employee
515	Laundry supervisor
522	Salesperson
522	Textile merchant
541	Corrections officer
611	Agricultural producer
731	Artisan, textile worker, Jute
932	Semi-skilled worker; Industry/factory worker
933	Warehouse worker
990	Student
991	Unemployed (political worker, career politician, intern, volunteer, activist, militia/guerilla
992	Retired

We followed a few additional general coding rules:

When more than one occupation is listed in `last_occupation`, `ISCO08` reflects whichever occupation is more likely to be a full-time job, then whichever occupation has a higher ISCO code.

Retired military officers are coded as their military occupation if retired less than two years,

and "retired" otherwise.

Except for civil servants and information/health officers, senior/chief/head/deputy/assistant positions are classified in the same occupation group as their respective unspecified positions. For instance, chief economist is classified in the same category as economist.

Unless otherwise specified, ministers, secretaries, and department heads are assumed to be government positions.

Internships are not coded as occupations.

Managers of specific companies or businesses are coded as managers for that industry. For instance, "manager of fishery business" is coded as a fishery manager.

Unspecified business owners are coded as 121, and owners of specified businesses are coded as specified managers. It should be noted that distinguishing between CEOs (112) and managers (121) is frequently difficult based on the information available.

Shop/store/cafe owners are classified as managers (100s) rather than shopkeepers (500s).

Country-specific coding decisions are located in the Section on Country Specific Data Collection and Coding Procedures.

Coding procedures: We utilized the Computer Assisted Structured Coding Tool (CASCOT), Version 5.6.1 (1) to aid in translating `last_occupation` entries into `ISCO08` codes. The software is copyright (2004–2020) of the University of Warwick and was licensed by the University of Warwick for use with this project. CASCOT is a computer program designed to make the coding of text information to standard classifications simpler, quicker, and more reliable (see Elias P, Halstead K and Prandy K (1993) Computer Assisted Standard Occupational Classification. London: HMSO. And Jones R (2004) CASCOT (Computer Aided Structured Coding Tool). Warwick: University of Warwick). CASCOT can be accessed at: https://warwick.ac.uk/fac/soc/ier/software/cascot/. We utilized CASCOT to reduce

the time necessary to code all occupational data into ISCO codes. We also attempted fuzzy string matching but found CASCOT to be far superior for our purposes.

We conducted a trial run of the CACOT machine by manually coding a sample of 1,987 legislators' `last_occupation` entries into ISCO-08 codes and comparing these to the machine output. CASCOT produces both a suggested `ISCO08` code and a confidence score ranging from 1 to 100. We found that using a cutoff confidence score of 87 resulted in obtaining a 95 percent match rate between manual and CASCOT output. In the sample, 53 percent of entries received CASCOT scores of 87 or above.

All occupational entries were first processed through CASCOT. After processing, those entries with CASCOT scores below 87 were hand-coded by a team of two trained research assistants. Entries with scores of 87 or above were also given a cursory check to ensure accuracy. Research assistants discussed all difficult coding decisions, and brought particularly tough cases to the full research group for discussion. Where occupation descriptions were found to be too vague or incorrect (i.e., not following proper procedure, coding elected political positions, including vague banker or union jobs, etc.), research assistants attempted to find more specific occupational data for that legislator. Three countries — Ecuador, France, and Iceland — had significant percentages of vague codes and were heavily amended at this stage.

Research assistants wrote all coding changes from the CASCOT output to `last_occupation` and `ISCO08` into R for reproducibility. We also ran checks to verify that the final `ISCO08` output matched the manual coding decisions. Some `last_occupation` changes were made directly to the standardized country files. This was the case for Bangladesh, Estonia, France, Iceland, Romania, and Zambia. Note that some ISCO codes had been manually entered during the first phase of data collection but these were not used in the final dataset. Nor were the codes used in the initial CASCOT trial run used.

Because of this two-step process for coding `last_occupation` into `ISCO08` codes, countries were processed in batches to allow manual `ISCO08` coding of countries completed earlier in

the second phase to be completed simultaneously with the cleaning and standardizing of other countries.

year_of_election

The year that the election deciding the parliament captured was held.

parliamentary_period

The years for which the parliament captured was in power.

legislature_number

The formal number of the parliament captured, for countries that number their parliaments. For countries that have had multiple constitutions and/or iterations of their national lower chamber, we used the latest numbering system.

total_mps

The number of legislators in the parliament captured.

total_mps_in_data

The number of legislators in the parliament captured for which there are entries in the dataset. Discrepancies, if they exist, are typically small and due to the inability to locate a canonical list of MPs elected at the beginning of the parliamentary term.

date_verified

Date of final verification of the country's data in the second phase of data collection.

name_source

Source for the legislator's first and last name.

party_source

Source for the legislator's party.

dob_source

Source for the legislator's date of birth.

education_source

Source for the legislator's educational attainment.

occupation_source

Source for the legislator's last occupation held before holding full-time, paid elective office.

source

General source specific to the legislator.

source_2

Additional general source specific to the legislator.

Country-specific data collection and coding procedures

Note that countries without specific information are not listed in this section.

Albania

Primary source used was archived data from the parliamentary website (https://www.parlament.al/deputies/by-regions/?lang=en). MPs who have since been arrested had their info wiped from the internet, so it was difficult to acquire and verify their data. Some occupational data is from http://www.shekulli.com.al/p.php?id=30155. CVs of

official ministers is from http://illyriapress.com/cv-te-zyrtare-te-ministrave-edi-rama-nuk-permend-diplomen/. Other sources used include the Partia Socialiste party website (http://ps.al/new/), Une Votoj (http://www.unevotoj.org/zgjedhjet09/index2.php), and OSCE (http://www.osce.org/albania/88530?download=true).

Some occupations seem to be listed in official sources based solely on education data. This is used where additional information is unavailable.

Argentina

The first phase of data collection was erroneously for Senators rather than members of Argentina's lower chamber. Therefore, all information was collected in the second phase of data collection. Outreach to individual legislators was unsuccessful.

No suitable list of the original parliament could be found to verify the list of legislators.

Australia

Legislator list was verified from https://www.aph.gov.au/Senators_and_Members/Members/Register/Previous_Parliaments/44P_Members_Interest_Statements.

Bahamas

Legislator list was verified using the Bahamas Election Centre (http://www.caribbeanelections.com/bs/elections/bs_results_2012.asp) and Wikipedia (https://en.wikipedia.org/wiki/2012_Bahamian_general_election).

The primary source used was Bahamas' Official Parliament website (https://www.bahamas.gov.bs/).

Bangladesh

Most data was obtained from the official parliamentary website, http://www.parliament.gov.bd//index.php/en/mps/members-of-parliament/former-mp-s/list-of-10th-parliament-members-english.

"Political worker" occupation was coded as 991 (unemployed). Members who both held a job and were involved in their family business are coded as having both occupations, i.e. advocate and industrialist. Many economic and political elites in Bangladesh own textile/garment factories, and thus are coded as industrialists in addition to their other occupation. Legislators that were no longer involved in business immediately before becoming an MP are coded as holding their other occupation (i.e, a doctor and industrialist who was no longer involved in their business right before election is coded as doctor).

Many Bangladeshi legislators are involved in politics starting early from college or due to family party ties. MPs with this background have never engaged in any formal work other than political campaigning for the party they belong to. They are coded as "Political Worker" for `last_occupation` and their `ISCO08` code is 991, for unemployed.

Belgium

Archived data was obtained from the National Assembly website using the Way Back Machine.

Belize

The National Assembly website does not provide information beyond names and party affiliation. Occupation and education were primarily obtained using party websites.

Benin

For party labels, Forces Cauris pour un Bénin émergent (FCBE) was translated as "Cauri Forces for an Emerging Benin".

Bhutan

Individual contact to legislators with missing information was unsuccessful.

Bolivia

An external dataset, the Diccionario biográfico de parlamentarios 1979–2019 by Salvador Romero Ballivián, was used.

Research assistants attempted email and telephone contact for Bolivia's Parliament but were unsuccessful. Most missing information is for indigenous legislators who belong to the Movimiento al Socialismo (MAS). Attempts were made to contact the party directly but these were also unsuccessful.

Bosnia Herzegovina

The two parliamentary lists used to verify legislators are https://www.parlament.ba/Representative/List?mandateId=8&memberType=1 and http://archive.ipu.org/parline-e/reports/2039_E.htm.

For `education`, translations were made according to http://www.euroeducation.net/prof/boherco.htm. "Faculty" was coded as bachelors, "Magister" was coded as masters, and "Doctorate" was coded as Ph.D.

Primary sources used include the official parliament website (https://www.parlament.ba/sadrzaj/poslanici/p/Archive.aspx?m=3&langTag=en-US), the CIN Database (https://translate.google.com/translate?hl=en&sl=hr&u=http://database.cin.ba/baza/biography.php%3Fid%3D60&prev=search and https://translate.google.com/translate?hl=en&sl=hr&u=http://imovinapoliticara.cin.ba/profil.php%3Fprofil%3D60&prev=search), and news articles about legislator occupations (https://translate.google.com/translate?hl=en&sl=hr&u=http://dnevni-list.ba/web1/profesori-orijentalnih-jezika-i-romanistike-

geodeti-veterinari-defektolozi/&prev=search and https://translate.googleusercontent.com/translate_c?depth=1&hl=en&prev=search&rurl=translate.google.com&sl=sr&u=http://www.klix.ba/vijesti/bih/koliko-su-obrazovani-drzavni-parlamentarci-ljekari-pravnici-ekonomisti/141110085&usg=ALkJrhg8ctIPMnvhO7Rq6NPIU-aIC87fHw).

Botswana

Attempts to contact the parliament for information were unsuccessful.

Bulgaria

For many legislators, `last_occupation` listed is based on their university profession rather than their actual occupational record. These cases were highlighted in blue in the clean country file. The Clerk of Parliament was contacted but does not collect occupational data for MPs.

The primary source used was the official parliament website (http://www.parliament.bg/en/MP).

Canada

An external dataset was received from Janet Bennett at the Canadian Library of Parliament, Collections Access and Preservation. This dataset detailed the occupations of all members.

Cape Verde

Some legislator CVs were accessible from the official legislature website.

Cote d'Ivoire

An external dataset that contained names, party affiliation, and occupation data was received from Giulia Piccolino, a country specialist. Relevant data was translated by a country

specialist. We were unable to gather additional data behind what was provided.

Official election results were used to verify the list of legislators (https://www.abidjan.net/elections/legislatives/2011/resultats/resultats.asp).

Croatia

Sources are sometimes contradictory, especially between English and Croatian versions of the parliamentary website. The research assistant attempted to verify entries with additional secondary sources as much as possible.

Ministers are appointed rather than elected, and thus are not treated as full-time elected positions for `last_occupation`.

Cyprus

Legislator list verified using a combination of https://en.wikipedia.org/wiki/2016_Cypriot_legislative_election#cite_ref-6, https://www.electionguide.org/elections/id/2536/, and https://greekreporter.com/2016/05/23/56-mps-for-new-cyprus-parliament-officially-announced/.

Some of the occupational information listed from sources seems to derive from the subject legislators studied at university rather than their career. Uncertainties were highlighted in blue in the clean country file.

Primary sources used were the official parliament website (http://www.parliament.cy/easyconsole.cfm/id/186 and https://translate.google.com/translate?hl=en&sl=el&u=http://www.parliament.cy/&prev=search) and a news article on legislator occupations (http://cyprus-mail.com/2016/05/23/new-mp-contingent-not-entirely-made-lawyers/). Some members' Linkedin and personal web pages were also used.

Phone outreach to the Parliament and email outreach to individual legislators were unsuccessful.

Czech Republic

The following doctorate degrees are equivalent to a masters in the Czech Republic and are coded as such: Doctor of General medicine (MUDr.), Doctor of Dental Medicine (MDDr.), Doctor of Veterinary Medicine (MVDr.), Doctor of Natural Science (RNDr.), Doctor of Pharmacy (PharmDr.), Doctor of Philosophy (PhDr.), Doctor of Law (JUDr.), and Doctor of Pedagogy (PaedDr.). Engineering degrees are also coded as masters.

Mayors, governors, and their deputies are all elected by city or regional assemblies from their own ranks. These positions are considered full-time political (elected) occupations and are treated accordingly.

Denmark

Temporary (substitute) MPs are excluded. Senior union positions are usually elected in Denmark, so these are considered political. `last_occupation` reflects the position one held prior to being elected as chairperson/secretary of a union, for those legislators. "Politisk konsulent" is coded as policy analyst/political advisor for `last_occupation`.

Party numbers were checked using https://www.robert-schuman.eu/en/doc/oee/oee-1600-en.pdf.

Dominican Republic

Attempts were made to contact individual legislators with missing information but these were unsuccessful.

East Timor

Very little information could be obtained through internet searches. The research assistant had to recode many occupations listed as "MP" to "NA".

Ecuador

Attempts were made to contact individual legislators (through email and Twitter) and Parliament directly but these did not yield additional information.

Inconsistencies in party composition were corrected primarily by using an archived page of the Ecuadorian National Assembly website (https://web.archive.org/web/20170518215559/http://www.asambleanacional.gob.ec:80/es/pleno-asambleistas, May 18, 2017). The Spanish-language Wikipedia page on legislator details was also used (https://es.wikipedia.org/wiki/Anexo:Asamble%C3%ADstas_del_Tercer_per%C3%ADodo_legislativo_de_la_Asamblea_Nacional_del_Ecuador).

El Salvador

The list of legislators was verified using two sources, https://es.wikipedia.org/wiki/Anexo:Diputados_de_la_Asamblea_Legislativa_de_El_Salvador_Per%C3%ADodo_2015-2018 and http://diario1.com/politica/2015/03/lista-de-diputados-electos-para-periodo-2015-2018/.

Some information was coded in the first phase of data collection but could not be verified because the sources have since been removed without being archived. This meant that they were irretrievable. This was the case for a significant number of legislators' education and occupations. These cases are highlighted in blue in the clean country file.

Formal information requests were submitted to the Recepción Oficina de Información Pública and the Tribunal Supremo Electoral. Two helpful documents were received from the Tribunal Supremo Electoral. One provided all legislator ages and professions, and the other provided some education information.

Estonia

Late in the process of data verification, it was discovered that a snapshot of a later parliament was used for Estonia rather than the list of legislators elected at the beginning of the parliamentary term. The number of legislators was corrected to match the size of the parliament but the included list does not reflect the parliament initially elected.

Fiji

The primary source for legislator names and parties was the official parliament website (http://www.parliament.gov.fj/Members/Parliamentery-Parties). Information on First Party members is from http://fijivillage.com/news-feature/FijiFirst-Party-announces-all-their-proposed-candidates-for-the-2014-general-election-k295sr, data on Fiji First members is from https://www.facebook.com/pg/thejetnewspaper/photos/?tab=album&album_id=704921506211951, and information on members who served in the military is from http://fijisun.com.fj/2016/01/21/in-the-line-of-duty/. Articles from the Fiji Sun and Fiji Times were also used.

Attempts to acquire information from the parliamentary research office and clerk were unsuccessful.

Finland

The list of legislators was verified using https://en.wikipedia.org/wiki/List_of_members_of_the_Parliament_of_Finland,_2015%E2%80%932019.

France

The list of legislators was verified using https://web.archive.org/web/20120805002116/http://www.assemblee-nationale.fr/qui/xml/liste_alpha.asp?legislature=14.

Georgia

The primary source used, Georgia's parliamentary website, only gives names of institutions of degrees rather than the degree obtained. Dates of educational attendance are not available either. `education` coded is therefore a best guess based on the institution listed (e.g., bachelors if a university is listed).

Germany

Diplomas are coded as bachelors, apprenticeships and job training are coded as short-cycle tertiary, and "degrees" are coded as bachelors if a masters is not explicitly mentioned. If a second state exam is mentioned for those with teacher and lawyer `last_occupation` then `education` is coded as masters. If those with a lawyer `last_occupation` do not have evidence of completing a second state exam, then `education` is coded as LLB.

Ghana

Post-College degrees (PGD) are coded as masters.

Primary sources used include the parliamentary website (http://www.parliament.gh/mps), the Electoral Commission (http://www.ec.gov.gh/resources/downloads/profiles-of-2016-parliamentary-candidates.html), and news articles (http://www.graphic.com.gh/news/politics/22-year-old-lady-wins-npp-primary-at-kwabre-east-constituency.html, http://www.myjoyonline.com/politics/2016/March-1st/science-teacher-gifty-twum-ampofo-wins-abuakwa-north-npp-primaries.php, and https://www.modernghana.com/news/534275/meet-yaw-buaben-asamoa-the-next-general-secretary-of-npp.html).

Greece

Many`last_occupation` entries reflect educational profession rather than verified occupation.

Guatemala

The legislator list used for verification was an archived version of the official legislature website from July 18, 2017 (accessed using the Way Back Machine). Since this is in the middle rather than at the beginning of the parliamentary term, it is plausible that there are differences from the initial elected list (i.e., legislators changed parties, or were arrested and lost their mandates). Both are common occurrences in Guatemala.

Guyana

The legislator list used was from https://en.wikipedia.org/wiki/2015_Guyanese_general_election.

A research assistant contacted the parliamentary clerk but that person was not able to give out date of birth and did not provide information on legislators with missing info.

Hungary

Educational qualifications do not translate cleanly into the standardized categories. Unspecified degrees that are five years in length are coded as masters, and those that are shorter are coded as bachelors unless otherwise specified. LLBs are codied as JDs only if legislators' bios specify a doctorate was earned.

Iceland

The following letter translation decisions were made for Icelandic to English: ð = D, þ = TH, and æ = ae.

India

Occupational info provided on the parliamentary website is vague for most politicians, e.g. agriculturist, industrialist, and business man. The research assistant attempted to narrow

down these descriptions to a more specific career where possible but otherwise used the parliament website info.

Jamaica

Primary data sources used were the official parliament website (http://www.japarliament.gov.jm/) and the Jamaica Information Service (http://jis.gov.jm/). Members' party affiliations are from http://www.caribbeanelections.com/knowledge/parliament/jm_parliament/jm_house.asp. Some information on Labour Party members is from their party website (http://www.jamaicalabourparty.com/membersofparliament). The Jamaica Observer was also used.

The Clerk of Parliament was contacted but does not collect occupation or education data for legislators.

Japan

The legislator list was sourced from https://ja.wikipedia.org/wiki/%E7%AC%AC47%E5%9B%9E%E8%A1%86%E8%AD%B0%E9%99%A2%E8%AD%B0%E5%93%A1%E7%B7%8F%E9%81%B8%E6%8C%99. Party-switching is common in Japan, so care was taken to include the party legislators were originally elected under.

Kenya

An external dataset was received from the Clerk of the Parliament. The data was prepared on the basis of information received from legislators between 2013 and 2014.

The legislator list was verified using https://en.wikipedia.org/wiki/List_of_members_of_the_National_Assembly_of_Kenya,_2013%E2%80%932017#Nominated_Representatives_(12) and https://info.mzalendo.com/position/member-national-assembly/?view=grid.

Kosovo

The primary data source used was the official parliament website (http://www.kuvendikosoves.org/?cid=2,102). Other sources include Members' personal LinkedIn profiles, Kallxo.com (for biographies of members), and the news sites Telegrafi and Express News.

A research assistant attempted to contact both the parliamentary secretary and the Democratic Institute of Kosovo for additional information but was unsuccessful.

Latvia

The primary sources used were the official parliament website (http://titania.saeima.lv/personal/deputati/saeima12_depweb_public.nsf/deputies?OpenView&lang=EN&count=1000), the Central Election Commission (https://www.cvk.lv/pub/public/28361.html) and the Vienotiba party website (http://www.vienotiba.lv/).

Lesotho

An external dataset was received from Shana S. Warren (drawing on research conducted for her New York University Ph.D. dissertation research) containing data on names, party, and gender. A dataset on party information was also used, extracted from http://www.electionpassport.com/.

Attempts by a research assistant to contact the Clerk of the House, the Speaker of the House, and the Deputy Speaker to acquire additional information were unsuccessful.

A small discrepancy in the number of legislators was due to the fact that no definitive list of legislators for the 10th Parliament could be located. For constituency MPs, an existing dataset is used. PR list MPs are corroborated using the number of list seats each party received and the party list names, which were located in PDF form. electionpassport.com was used.

For `education`, "diploma" is coded as short-cycle tertiary.

Liberia

Reports from National Democratic Institute/USAID ("Know Your Representatives") and the European Commission to Monrovia/KAF ("A Profile of Members of the 52nd Legislature of Liberia") were used.

Lithuania

Primary data sources were the official parliament website (http://www.lrs.lt/sip/portal.show?p_r=8801&p_k=2&filtertype=0) and the NGO Reitinguok (http://www.reitinguok.lt/). Member personal websites were also used where available.

The source used to verify legislators was https://en.wikipedia.org/wiki/Twelfth_Seimas_of_Lithuania.

Ministerial appointments can be unelected positions in Lithuania (i.e., not members of parliament), so these are counted as prior occupation accordingly.

"Diploma" is translated to mean bachelors qualification.

Madagascar

Four initially vacant seats were soon filled in Madagascar and are included in the dataset and with `total_mps`. This results in a legislature size of 155 rather than 151.

The Way Back Machine was used to obtain most information. The official Madagascar National Assembly website provides ages rather than dates of birth. `dob` was coded using estimated year of birth in order to best standardize the entries.

Malta

The list of legislators was verified using https://www.parlament.mt/media/87852/general-elections-results-government-gazette.pdf.

There is often not a clear chronology to occupations, so some occupations listed may be held simultaneously by legislators holding political positions.

Mauritius

The primary data source used was the official parliament website (http://mauritiusassembly.govmu.org/English/hmembers/Pages/default.aspx). Data on parties/alliances is from http://everypolitician.org/mauritius/national-assembly/term-table/2014.html.

For `education`, HSC (higher school certificate) and SC (school certificate) were both coded as secondary.

Much of the first phase data was collected from parliamentary pages that were removed without being archived. This data could not be verified for accuracy.

Mexico

The legislative list used to check and verify MPs is from http://gaceta.diputados.gob.mx/SIL/Legislaturas/Listados.html. A common source of information was the Sisteem De informacion eficiente (SIE).

Alternate deputies are counted as elected politicians for the purposes of coding occupation. Note that most national deputies in Mexico hold internal party elected positions for extensive periods of time. These were generally not coded as occupations, nor considered full-time elected positions.

Moldova

Most legislators report "graduated from [school]" but do not have their specific qualifications listed. A bachelors degree was assumed unless a postgraduate institution is listed.

A research assistant attempted to contact parliament and avere.md (http://www.avere.md) for missing info but was unsuccessful.

Namibia

The list of legislators was validated with the parliamentary website list and Wikipedia (https://en.wikipedia.org/wiki/List_of_members_of_the_6th_National_Assembly_of_Namibia).

From `education`, certificates are coded as secondary qualifications and diplomas are coded as short-cycle tertiary degrees. Almost every single legislator worked for their party before being elected.

Netherlands

"HBO" is coded as a bachelors degree. "Dr.s" is coded as masters.

Norway

The list of legislators was verified using the Wikipedia page.

The Candidatus realium degree is coded as PhD. Cand.socion and Cand.mag are coded as masters degrees.

Many legislators held intermediate jobs between being a municipal councilor and becoming a representative. Municipal councilor is counted as a full-time elected position, so `last_occupation` reflects the job held prior to being municipal councilor.

Pakistan

Most data was externally received. Data on `name` and `party` was received from Miriam Golden (EUI), Saad Gulzar (Stanford University), and Luke Sonnet (UCLA), who originally sourced it from Jake Shapiro (Princeton University). All `dob`, `education`, and `last_occupation` data for Pakistan was obtained from the KP Assembly of Pakistan, which requested the data from from the National Assembly of Pakistan. Data is entered as obtained from this external source because it was detailed enough to describe the characteristics of the members.

The parliamentary website and Wikipedia were used to verify the list of legislators.

Occupations listed as "landlords" in `last_occupation` are coded as unemployed in `ISCO08`.

Panama

Attempts to contact individual legislators were unsuccessful.

Papua New Guinea

The main source of info was the parliamentary website. Occupation listed was sometimes simply "business". In these cases, an additional search for information was conducted. Where the position of CEO could be verified, `last_occupation` was coded as "CEO"; otherwise, it was left blank.

External information was acquired from a database provided by Jasper Cooper (UCSD), who had compiled the data along with Terence Wood (ANU) and Gareth Nellis (UCSD) and that contained information on national parliaments through 2012. Additional external data was obtained directly from the parliament librarian, Elesallah Matatier.

Paraguay

We were unable to obtain much information about legislators who faced corruption charges or were involved in illegal drug dealing activities. Most legislators have deleted their parliamentary emails and thus could not be contacted. Attempts were made to contact Parliament but this did not yield additional information.

Peru

The legislator list was verified through an archived webpage, https://web.archive.org/web/20160708141036/http://elcomercio.pe/especiales/congresistas-electos/. July 7, 2016 was used as the date.

Legislators who use the title of "Dr." but have no evidence of completing a Ph.D. are not coded as obtaining a doctoral degree. It is assumed that lawyers have obtained at least a bachelors degree.

Philippines

Legislator list verified using https://en.wikipedia.org/wiki/17th_Congress_of_the_Philippines#ref_kabayan.

Some information acquired externally from Ceci Cruz (UCLA). This data was originally compiled by the Philippines Center for Investigative Journalism (PCIJ) on their website I-Site, which has since been taken down. They had, in turn, derived this data from the Statement of Assets and Liabilities (SALN) that all politicians are required to fill out.

Attempts to get info from the parliament directly were unsuccessful.

Poland

City councilors have extensive rules regulating their non-political activities (https://pl.wikipedia.org/wiki/Radny) and thus are treated as full-time political positions.

Portugal

For education codes, "Frequência" means not finished, so degrees with this label were not counted.

Romania

There is frequently a vague distinction in `education` between short-cycle tertiary, bachelors, and masters.

The line between CEO and manager is blurred in Romanian `last_occupation` codes.

Senegal

The official website for National Assembly was used for most of the data, which is located at http://www.assemblee-nationale.sn/. We had difficulty obtaining detailed and reliable data for Senegal. Legislators listed in the first 70 entries on the official website tended to have more detailed biographies than those with later entries. Wikipedia and news articles with relevant information were rarities.

Some external data was received from Catherine Lena Kelly (National Defense University), including her book, *Party Proliferation and Political Contestation in Africa: Senegal in Comparative Perspective.*

The list of legislators was verified using http://www.assemblee-nationale.sn/anciennes-legislatures-12-t1-assnat-p6.xml.

Serbia

Several `last_occupation` entries taken from the parliamentary website are likely just areas of study rather than occupations. These are colored in blue on the country's clean file. Additionally, there is oftentimes not a clear distinction between secondary school teachers and university professors.

Legislators listed as graduates but without specified university degrees are coded as having bachelors. Medical doctors without listed education are coded as having bachelors.

Most information was from the Open Parliament Initiative (https://otvoreniparlament.rs/o-nama). Attempts to contact the Serbian parliament were unsuccessful.

Sierra Leone

Three initial vacancies were created due to legal reasons but filled days after the initial election. These three seats are included in the dataset and with `total_mps`. This results in a parliament size of 112 rather than 109.

Some external data was received from Shana Warren (NYU Ph.D.; now IPA).

The primary data sources used were PDFs from the National Electoral Commission (Notice of Certified Final Results of Parliament Candidates, 2012; list of nominated candidates) and the National Democratic Institute (A Directory of the Parliament of Sierra Leone 2007–2012).

Slovak Republic

"štátny tajomník" is coded as deputy minister.

South Africa

Most legislators have a history as trade union leaders or holding previous political office but it was not always clear if these positions are listed in their bios. Many legislators with

occupations coded as teachers were more akin to 'pseudo' teachers, spending much of their time as activists.

Attempts were made to acquire information from the Parliamentary secretary and the chief party whips but these were unsuccessful.

A definitive legislator list of the beginning of the Parliament could not be located. While alternative lists were available, it was deemed too difficult to reconcile these conflicting lists. Therefore, there is a discrepancy between the number of legislators in the parliament and the number of MPs in the dataset. Some legislators in the dataset entered parliament mid-session, well after the election. The parliament captured is a snapshot from 2017, whereas the general election was held in 2014.

South Korea

The legislator list used was the country's National Election Committee's rosters of elected legislators (http://info.nec.go.kr/main/showDocument.xhtml?electionId=0000000000&topMenuId=EP&secondMenuId=EPEI01). The Parliamentary website, Wikidata, and Naver were the main sources for individual data.

For birth dates, Wikidata was used to ensure that `dob` captures the standard Gregorian calendar rather than the lunar calendar (DOBs listed on the parliamentary website were in both formats).

`first_name` and `last_name` are reversed, and `name` lists surname first.

Spain

Individual party names are used, rather than parliamentary name groups.

Suriname

The official National Assembly website (https://web.archive.org/web/20140730234618/http:/dna.sr/het-politiek-college/leden/) was used for most Suriname entries. Educational degrees were taken from names on this website where provided.

Sweden

For `education`, "Subject teacher degree" is coded as bachelors. "Folk High School" is coded as short-cycle tertiary, as they provide study opportunities similar to universities but cannot hand out degrees.

Switzerland

A verified list of elected legislators for the 2015–2019 period could not be located. The parliamentary website lists only 199 MPs, even though 200 were elected (https://www.parlament.ch/en/%C3%BCber-das-parlament/archives/groups-archive).

Taiwan

The sources used to verify the 9th Legislative Yuan are https://en.wikipedia.org/wiki/9th_Legislative_Yuan and https://www.ly.gov.tw/EngPages/List.aspx?nodeid=221.

Tanzania

The primary source used was the official parliament website (http://www.parliament.go.tz/mps-list), particularly archived information (https://web.archive.org/web/20120626090722/http://parliament.go.tz/index.php/members/memberslist/all/all/2010-2015). News and social media sources were used for some individual members.

Included legislators are 264 directly elected from constituencies, 113 special seats elected from women-only lists, and 5 members elected indirectly by the Zanzibar House of Representatives.

The Attorney General (Ex officio member) and 10 MPs appointed by the President are not included, yielding a total legislature size of 382 rather than 393.

`dob` was mostly sourced from https://www.parliament.go.tz/.

Trinidad and Tobago

Outreach to individual legislators with missing data was attempted unsuccessfully.

Tunisia

The French website for the Tunisian website was found to be inaccessible. A supplementary website was located instead: https://majles.marsad.tn/2014/fr/elus/, accessed 12/08/2020.

Turkey

Legislators who entered politics after doing mandatory military service (after they graduated from university) are coded as students.

Ukraine

While the official number of MPs is 450, many of these seats are vacant because the Ukrainian government no longer controls the territories from which these legislators should have been elected (Crimea, and some districts in Lugans and and Donetsk). The 418 legislators captured are those that were seated in the first session of the 8th convocation, on November 27, 2014. There were 32 vacancies total at this time.

For `party`, legislators elected through the party list are coded as affiliated with that party, regardless of whether they are formal members of the party. Similarly, every legislator nominated by a party in a single-member district is coded with that party affiliation.

Most Ukrainian politicians got their degrees in Soviet or early post-Soviet times, which means

they are neither bachelors nor masters but something in-between. These are generally labeled as "Specialists" degrees and typically take 5 years to acquire. They are coded as masters.

Legislators who were well-known opposition journalists before the revolution and were promoted to political office (as government officials) shortly thereafter are coded according to their pre-revolutionary occupation (journalist), as this better captures their social background. Those who got elected after being military volunteers in ATO zones are coded as "military" despite the fact that they were only in the army for a short period of time, since it is by being in the army that they gained popular support to run for office. Those who were in charge of joint-stock companies are coded as CEOs, whereas those who ran limited liability companies are coded as presidents of LLC. In most cases, LLCs are smaller than JSCs. However, both receive the same code (112) in the ISCO-08 classification system.

Many MPs were previously deputies of city councils. In most cases, these were not full-time elected positions and thus are treated accordingly (in a few rare cases, especially for bigger cities, these are full-time positions).

External data downloaded from https://data.rada.gov.ua/open/data/mps-all and entitled "General information about People's Deputies of Ukraine (for all convocations)". The original data was compiled by the Open Data Portal from Ukraine's official parliamentary website. This contained names, DOB, and gender.

United Kingdom

Data comes externally sourced from Jennifer vanHeerse-Hudson and Rose Campbell (vanHeerde-Hudson, J. and R. Campbell (2015). Parliamentary Candidates UK Dataset (v. 1). www.parliamentarycandidates.org.). The data was originally collected by Jennifer vanHeerde-Hudson and Rosie Campbell with the support of the Leverhulme Trust (RPG-2013-175). Because most of the data comes from this source, no `date_verified` is entered.

Data for Northern Ireland MPs was added. Some additional missing info was filled in and about 15 percent of occupations originally listed as elected positions (councilor, mayor, MEP, trade union official) were recoded. "SpAd" coded as "Political Adviser (SpAd)".

For `education`, tertiary incomplete qualifications were coded as secondary. PhD and masters were distinguished for "postgrad" codes using a list of MPs with PhDs (http://virtualstoa.net/2016/08/28/doctors-in-the-house/). Those with "postgrad" listed but were not on the PhD list are coded as having masters.

`dob` was only the year of birth in the original dataset, so all values are defaulted to January 1.

United States

Non-voting delegates and resident commissioners are included, raising the total number of legislators from 435 to 441. The non-voting members included are Jennifer Gonzalez-Colon (Puerto Rico), Eleanor Holmes Norton (District of Columbia), Stacey Plaskett (U.S. Virgin Islands), Amata Coleman Radewagen (American Samoa), Gregorio Sablan (Northern Mariana Islands), and Michael San Nicolas (Guam).

Some data was acquired from *CQ*, which requires a subscription to access.

Zambia

The primary data source was the official parliament website (http://www.parliament.gov.zm/members-of-parliament). `dob` was mostly sourced from http://www.parliament.gov.zm/. Some occupational data was obtained from news sources.

Zambian education was standardized using information from https://www.sheffield.ac.uk/international/entry-requirements/zambia.

Description of workflow and R procedures

Directory Layout

This section gives an overview of each main directory folder.

`01_carnes_lupu_data`

This subdirectory contains all materials collected by Nick Carnes and Noam Lupu during the first phase of data collection. They collected data on legislators in 103 democracies over 2016–2017 with a team of research assistants under the supervision of Emily Noh. Noh standardized these files, and verified 25 of them. This file also contains documentation on the ISCO-08 coding system and source documents from the first phase of data collection.

`02_other_sourced_data`

This subdirectory contains all records of correspondence to request and/or verify data, as well as all external datasets received during the second phase of the project. Miriam Golden led this phase, with a team of research assistants supervised by Esme Lillywhite.

`03_initial_cleaned_data`

This subdirectory contains a folder of cleaned country files (`/01_clean_files`), standardized files prepared for R processing (`/02_stnd_files`), and files used to collect missing `dob` data from a web scraper (`/03_msngdobs_files`). This subdirectory also contains country-level contextual data and the template research assistants used to create clean country files.

`04_code`

This subdirectory contains all R code files, as well as archived and trial code. The coding pipeline is explained in more detail in the Section on the Coding Pipeline.

05_cascot_input_data

This subdirectory contains batched files with assigned `unique_id` (`/01_stnd_with_id_batch_files`) and batched files prepared for CASCOT input (`/02_cascot_input_batch_files`).

06_cascot_output_data

This subdirectory contains output from the CASCOT machine (`/01_cascot_isco_output`), batched files with the merged CASCOT output (`/02_cascot_and_manual_isco`), finalized `ISCO08` codes for each legislator and verification files for the ISCO-08 coding (`/03_cascot_and_manual_isco`), and the final database (`/04_processed_country_files`).

07_lab_notebooks

This subdirectory contains all research assistant lab notes, the research assistant handbook, `ISCO08` coding decisions, and verification spreadsheets (checking for duplicate entries, cleaning `education`, listing missing `dob` entries, and checking matching of sources). It also contains amendments made to the dataset during processing (`/batch_4`, `/batch_5`, and `05_Modified_MPs_xlsx`), and archived files (`/_archive`).

08_codebook

This subdirectory contains the R code and a copy of the GLD codebook.

crossnational.Rproj

This is the R Project file for all code files. Those seeking to replicate the production of the dataset from the standardized country files should start by opening this file.

global_legislator_dataset_status.xlsx

An overview of countries coded, the research assistant assigned to each country, missingness, hours spent on completion, and date verified.

Batches

Countries were processed in batches as they were finished. Each batch was cleaned using R code, run through the CASCOT machine, and then had `ISCO08` entries finalized with manual coding. These batches were then merged together for final cleaning steps. Batches 1, 2, 3, and 6 include individual country files. Batches 4 and 5 add legislators that were left out during the initial processing of each country in batches 1–3. For these two batches, separate lists were also created for legislators that needed to be deleted or for whose `party` needed to be amended.

Batch 1

Albania, Argentina, Australia, Austria, Bahamas, Bangladesh, Belize, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Canada, Chile, Colombia, Costa Rica, Cote d'Ivoire, Croatia, Cyprus, Denmark, Dominican Republic, Ecuador, El Salvador, Fiji, Georgia, Germany, Ghana, Guatemala, Guyana, Hungary, Ireland, Israel, Jamaica, Kenya, Kosovo, Lithuania, Malta, Mauritius, Mexico, Moldova, Namibia, Pakistan, Papua New Guinea, Paraguay, Peru, Philippines, Portugal, Romania, Senegal, Serbia, Solomon Islands, South Africa, South Korea, Suriname, Sweden, Taiwan, Tanzania, Trinidad and Tobago, Ukraine, United States, Uruguay, Zambia

Batch 2

Belgium, Benin, Cape Verde, Czech Republic, East Timor, Estonia, Finland, France, Greece, Iceland, Italy, Japan, Latvia, Lesotho, Liberia, Luxembourg, Mongolia, Montenegro, Netherlands, New Zealand, Nigeria, Norway, Panama, Poland, Sierra Leone, Slovakia, Slovenia, Spain, Switzerland, Turkey, United Kingdom

Batch 3

Madagascar, Tunisia

Batch 4

Correcting discrepancies in OECD countries

Batch 5

Correcting discrepancies in non-OECD countries

Batch 6

India

Coding Pipeline

This section describes the coding pipeline used to process, merge, and clean the dataset.

Manual Cleaning

Individual country files were cleaned and standardized, as explained in the Section on Second Phase Procedures.

`00_merging_dobs`

The DOB process for Emily Noh countries (most of batch 2) was partially automated by the use of a web scrapper developed by Ivan Formichev (found in `/web_scripts`). The web scrapper was run on these countries. Then, research assistants manually located the DOBs that the scrapper failed to find. Finally, the results were merged into the individual standardized country files.

The DOB scrapper was used for Australia, Austria, Canada, Czech Republic, Estonia, Finland, France, Greece, Italy, Japan, Latvia, Luxembourg, Netherlands, New Zealand, Nigeria, Norway, Slovenia, Spain, Switzerland, and Turkey. The DOB scrapper was not used

for Benin, Cape Verde, Liberia, Mongolia, or Panama, as it returned virtually no useful data. These countries' DOBs were only coded manually.

`01_merge_country_data`

This file merges each batch of standardized country files into a single file and checks that the merged columns are identical.

`02_clean_country_data`

This file turns all missing entries to "NA", identifies and fixes `dob` errors, evaluates for MP gender balance, and adds a unique identifier to each entry. Major issues with `dob` were discovered that had to be manually corrected in the standardized files. All country .xlsx files completed in USA defaults had to have their `dob` columns transformed from M/D/Y to D/M/Y (this was the case for Bosnia, Colombia, Costa Rica, El Salvador, Lithuania, Mexico, Pakistan, Peru, Taiwan, United States, and Zambia). Bulgaria was completely recoded manually. Canada's `dob` entries were found to be inaccurate, so the country was processed through Formichev's web scrapper. Sporadic `dob` issues identified through the R code were manually fixed in the standardized country files.

`03_prepare_cascot_input`

This files creates a new shortened occupation column, cleans and recodes certain occupations, and creates the batched files that were then inputted into the CASCOT machine. The shortened occupation was created and used for the CASCOT machine because longer word strings tended to produce less accurate results.

`04_clean_cascot_output`

This file cleans up the returned CASCOT machine output files and joins them with the cleaned country data produced from `02_clean_country_data`. The CASCOT machine produces

four-digit ISCO-08 codes, so the last digit was removed to yield three-digit codes. A column with three-digit ISCO descriptions was also created to aid the manual coding process.

`05_manual_isco_code`

This file merges all batched files together, corrects `last_occupation` and `ISCO08`, and verifies the results with the manually coded ISCO-08 entries. All changes produced through the manual coding procedures outlined in the Section on Coding Procedures were reconciled in this file.

`06_final_cleaning`

This file completes several cleaning and standardization tasks to produce the finalized database. It corrects and standardizes `education`, removes deleted legislators and corrects `party` entries (both of which are identified in batches 4 and 5), adds missing `dob` entries from the web scrapper, adds country-level variables, removes all accents, evaluates and removes duplicate entries, merges in source data, standardizes capitalization, trims remaining white space, and saves the final output.

Unique `education` entries are identified, and a handful of changes were made to the individual standardized country files for vague or improperly coded data.

During the process of verifying legislator lists and reconciling discrepancies between `total_mps` and `total_mps_in_data`, some legislators needed to be added and some needed to be removed. Additions were made in batches 4 and 5. Subtractions were made in this file. This was necessary to preserve `unique_id` identifiers, which were assigned after the individual standardized country files were merged into batches. Changes to party affiliation were also made in this file.

The web scrapper used earlier on the Emily Noh countries was utilized on all remaining entries with missing `dob` information (where manual attempts to find the info had already

failed). The scrapper was able to reduce missingness by about 10 percent. These located `dob` entries were merged into the dataset.

Unnecessary columns were removed, and a few columns were renamed. Any rows with blank names were removed. As a result, hundreds of blank entries in Jamaica and Senegal that were in the standardized files were removed.

Country-level contextual data was merged into the dataset as well. This information consisted of a UN country code; the year of election, parliamentary period, and legislature coded; the total number of legislators for that parliament and the number of legislators captured in the dataset; and the final date of data verification for each county.

Legislator-level sources were merged in from the clean country files. Because these files were not cleaned or standardized in any of the previous R code and did not include unique identifiers, accurately merging the data required significant manual cleaning. Legislators were matched on `country_name` and `name`. Checks were conducted to ensure that the sourcing information merged correctly.

Potential duplicate entries were identified and, where found to be actual duplicates, removed. Finally, capitalization for `name`, `first_name`, `last_name`, and occupation columns were each standardized, and white space in party entries was removed.