V-Dem: Aggregating Expert Opinion on Democracy with IRT Models

Dan Pemstein, *North Dakota State University*
Kyle L. Marquardt, *University of Gothenburg*
Eitan Tzelgov, *University of East Anglia*

Yi-ting Wang, *National Cheng Kung University*
Farhad Miri, *ex-University of Gothenburg*
Joshua Krusell, *ex-University of Gothenburg*

Anna Lührmann, *University of Gothenburg*
Laura Maxwell, *University of Gothenburg*
Johannes von Römer, *University of Gothenburg*

Staffan I. Lindberg, *University of Gothenburg*
Juraj Medzihorsky, *University of Gothenburg*

Varieties of Democracy

Varieties of Democracy (V-Dem) is an approach to conceptualizing and measuring democracy that goes beyond the simple presence of elections and distinguishes between high-level principles of democracy: electoral, liberal, participatory, deliberative, and egalitarian. The V-Dem Dataset covers 201 countries and dependent territories from 1789 to 2017 on about 400 indices of which over 350 are specific and 52 aggregated. About a half of the specific indices are based on documentary information and the rest on expert judgments. With six Principal Investigators, 14 Project Managers responsible for issue areas, more than 30 Regional Managers, 170 Country Coordinators, Research Assistants, and 3,000 Country Experts from 177 countries, the V-Dem project is one of the largest social science data collection projects. The dataset has a wide and diverse user base, with over 24,000 downloads from more than 150 countries.

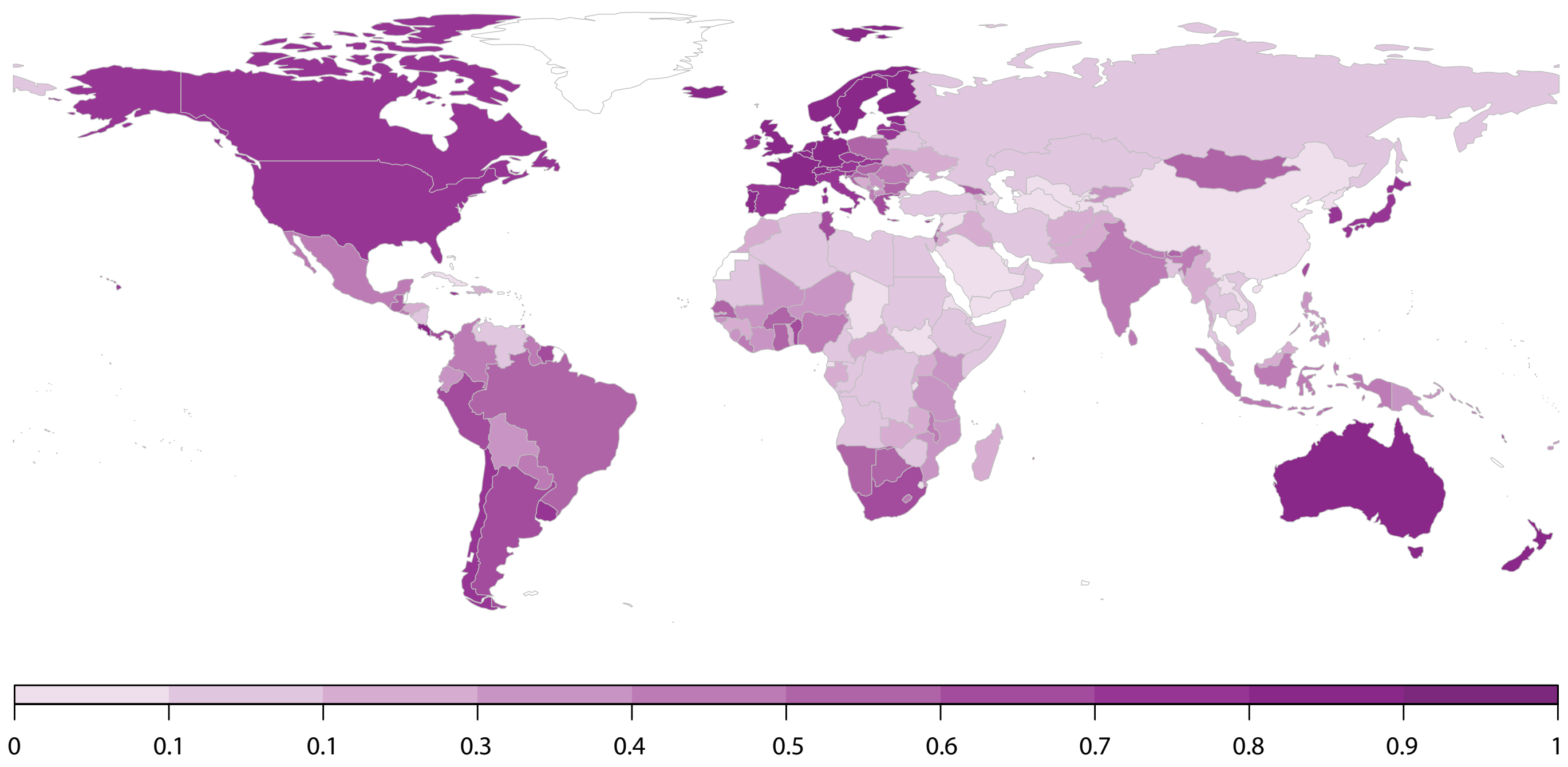
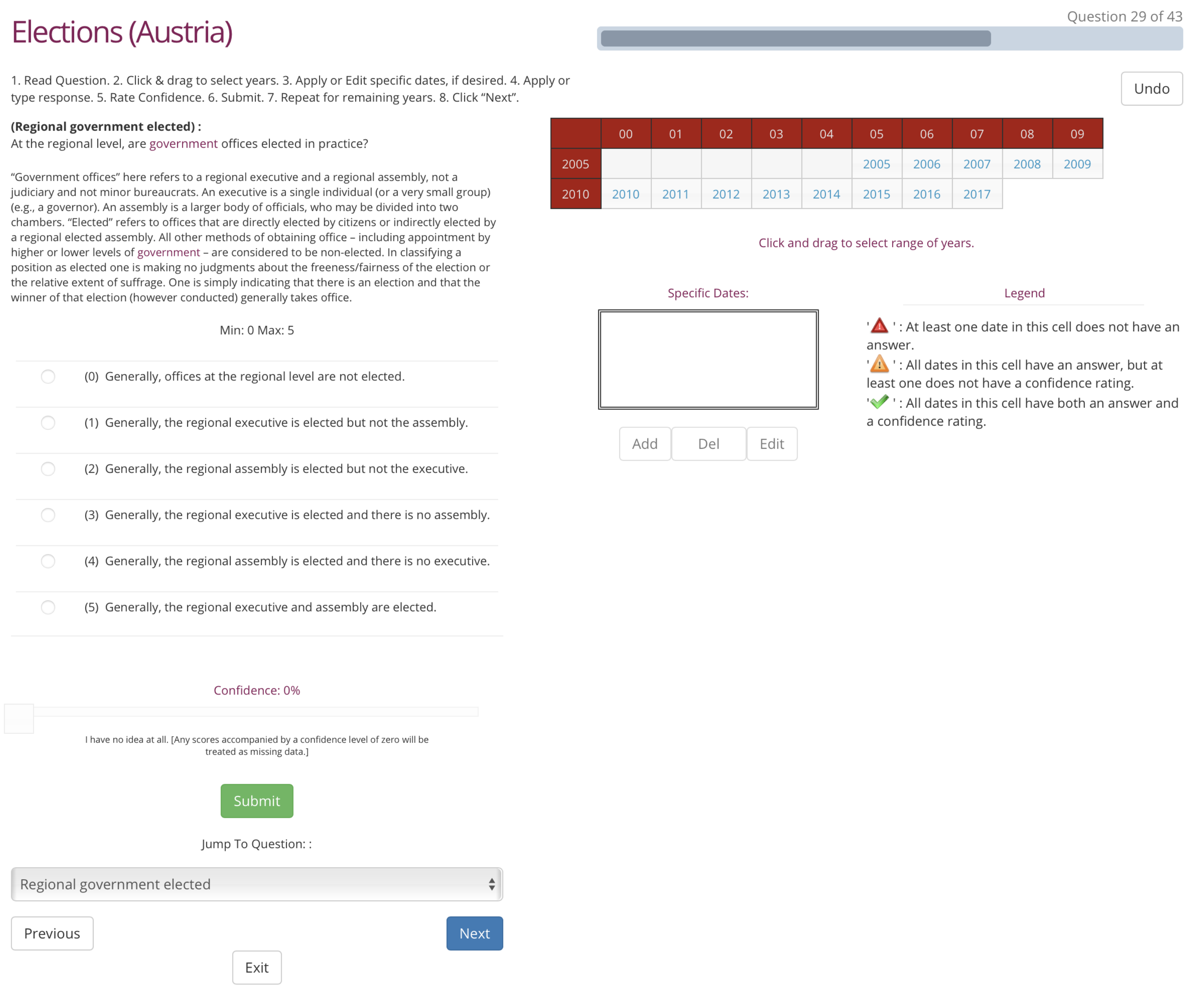


Figure 1: The V-Dem Liberal Democracy Index across the world in 2017.

Collecting expert judgments on democracy

The 3,000 experts provide their judgments through an online survey. The overwhelming majority of the survey items use ordinal scales, and also let the experts report how confident they are in their ratings. The interface, shown below, allows the experts to answer each question for multiple years in the same country.



Elections (Austria)

Question 29 of 43

1. Read Question. 2. Click & drag to select years. 3. Apply or Edit specific dates, if desired. 4. Apply or type response. 5. Rate Confidence. 6. Submit. 7. Repeat for remaining years. 8. Click "Next".

(Regional government elected) :
At the regional level, are government offices elected in practice?

"Government offices" here refers to a regional executive and a regional assembly, not a judiciary and not minor bureaucrats. An executive is a single individual (or a very small group) (e.g., a governor). An assembly is a larger body of officials, who may be divided into two chambers. "Elected" refers to offices that are directly elected by citizens or indirectly elected by a regional elected assembly. All other methods of obtaining office – including appointment by higher or lower levels of government – are considered to be non-elected. In classifying a position as elected one is making no judgments about the freeness/fairness of the election or the relative extent of suffrage. One is simply indicating that there is an election and that the winner of that election (however conducted) generally takes office.

Min: 0 Max: 5

○ (0) Generally, offices at the regional level are not elected.
○ (1) Generally, the regional executive is elected but not the assembly.
○ (2) Generally, the regional assembly is elected but not the executive.
○ (3) Generally, the regional executive is elected and there is no assembly.
○ (4) Generally, the regional assembly is elected and there is no executive.
○ (5) Generally, the regional executive and assembly are elected.

Confidence: 0%

I have no idea at all. [Any scores accompanied by a confidence level of zero will be treated as missing data.]

	00	01	02	03	04	05	06	07	08	09
2005						2005	2006	2007	2008	2009
2010	2010	2011	2012	2013	2014	2015	2016	2017		

Click and drag to select range of years.

Specific Dates:

Legend

'⚠' : At least one date in this cell does not have an answer.
'⚠' : All dates in this cell have an answer, but at least one does not have a confidence rating.
'✓' : All dates in this cell have both an answer and a confidence rating.

Figure 2: The expert survey interface.

The experts are recruited based on their knowledge of a specific country, which they are asked to rate on multiple questions over a longer year span. The goal is to have at least five experts of different backgrounds judge each country-year. To improve the comparability of the judgments across countries, the survey implements

- batteries of anchoring vignettes;
- bridge rating, where an expert judges an additional country over a longer span;
- lateral rating, where an expert judges several additional countries over a shorter span.

The V-Dem measurement model

Each rating collected in the survey is nested in expert, question, and, hierarchically, country-year in country. In turn, each expert is nested in their primary country of expertise. For pragmatic reasons, the judgments are aggregated by question. In aggregating the judgments, V-Dem takes into account two ways in which the experts might differ. First, they might perceive the thresholds on the ordinal scales differently. Second, they might not be equally knowledgeable and as a result their judgments might contain different amounts of stochastic noise. The judgments are aggregated with an ordinal IRT model that includes expert-specific threshold and reliability parameters.

More formally, y_{uj} is the rating of country-year u by judge j that imperfectly reflects the latent value z_u. While y_{uj} is on an ordinal scale with K categories, $z_u \in \mathbb{R}$. Under the model,

$$y_{u[i]j[i]} \sim \text{Categorical}(p_{u[i]j[i]}),$$

where p_{uj} is the simplex of category probabilities of expert j rating country-year u. The probabilities are a function of the latent score z_u and the expert-specific thresholds $\{\tau^J_j\}$ and slope β_j. Specifically,

$$p_{u[i]j[i]k} = \phi(\tau^J_{j[i]k+1} - \beta_{j[i]} z_{u[i]}) - \phi(\tau^J_{j[i]k} - \beta_{j[i]} z_{u[i]})$$

where ϕ is the unit normal CDF. The judge-specific thresholds are in turn modeled as

$$\tau^J_{jk} \sim \text{Normal}(\tau^C_{c[j]k}, 0.25),$$

where τ^J_{jk} is the threshold of judge j and τ^C_{ck} of country c that separate categories k and k + 1. The country thresholds are modeled as

$$\tau^C_{ck} \sim \text{Normal}(\tau^W_k, 0.25),$$

with independent uniform priors for the world-level thresholds,

$$\tau^W_k \sim \text{Uniform}(-6, 6).$$

The judge slopes have independent priors,

$$\beta_j \sim \text{Normal}^+(1, 1).$$

The country-year latent values receive informative normal priors,

$$z_u \sim \text{Normal}(\bar{z}_u, 1),$$

where $\bar{z}_u$ is the weighted mean of expert ratings,

$$\bar{z}_u = \sum_j w_{uj} y_{uj},$$

where w_{uj} is the weight computed from the self-reported confidences. The current development implementation of the model is shown below.

```
data {
    int<lower=2> K;     // no. response categories
    int<lower=1> J;     // no. judges (experts)
    int<lower=1> C;     // no. countries
    int<lower=1> N;     // no. country-years
    int<lower=1> R;     // no. ratings
    int<lower=1, upper=C> judge_country[J]; // the country of each judge
    int<lower=1, upper=K> y[R];         // ratings
    int<lower=1, upper=J> j_id[R];      // judge ids
    int<lower=1, upper=C> c_id[R];      // country ids
    int<lower=1, upper=N> cy_id[R];     // country-year ids
    real<lower=0> sigma_judge;          // judge sd around country threshold
    real<lower=0> sigma_country;        // country sd around world threshold
    vector[N] z_bar;    // country-year prior means
}
parameters {
    vector[N] z_star;                               // country-year raw values
    real<lower=0> beta[J];                          // judge reliability
    vector<lower=-6, upper=6>[K-1] tau_raw_world;   // world raw thresholds
    vector[K-1] tau_raw_country[C];                 // country raw thresholds
    ordered[K-1] tau_raw_judge[J];                  // judge raw thresholds
}
transformed parameters {
    vector[N] z;                                // country-year values
    ordered[K+1] tau[J];                        // judge thresholds
    z = z_bar + z_star;
    for (j in 1:J) {
        tau[j,1] = -1e6;
        tau[j,K+1] = 1e6;
        for (k in 2:K)
            tau[j,k] = tau_raw_judge[j,k-1];
    }
}
model {
    real p;
    z_star ~ normal(0, 1);
    for (j in 1:J)
        beta[j] ~ normal(1, 1)T[0,];
    for (c in 1:C)
        tau_raw_country[c] ~ normal(tau_raw_world, sigma_country);
    for (j in 1:J)
       tau_raw_judge[j] ~ normal(tau_raw_country[judge_country[j]], sigma_judge);
    for (r in 1:R) {
        p = Phi_approx(tau[j_id[r], y[r]+1] - z[cy_id[r]]*beta[j_id[r]])-
            Phi_approx(tau[j_id[r], y[r]] - z[cy_id[r]]*beta[j_id[r]]);
        target += log(p);
    }
}
```

Updating the dataset

The institute annually updates the dataset with information on the previous year and hitherto excluded country-years, as well as new information on already included ones. The models are fitted on the Kebnekaise computer of the Swedish National Infrastructure for Computing, in about 250 jobs, one per index. Most jobs involve four chains with 10,000 iterations each, with some increased to 20, 40, or 80 thousand. The input data for all the jobs takes about 75MB in .rds files. The total output is around 22GB of samples from the posteriors stored in .rds files. In the dataset, the posteriors are summarized with point estimates and standard deviations. The complete samples from the posteriors are available separately.

Acknowledgements

This project was supported by

- National Science Foundation, Grant SES-1423944, PI: Daniel Pemstein
- Riksbankens Jubileumsfond, Grant M13-0559:1, PI: Staffan I. Lindberg
- Swedish Research Council, 2013.0166, PI: Staffan I. Lindberg and Jan Teorell
- Knut and Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg
- University of Gothenburg, Grant E 2013/43 as well as internal grants from the Vice-Chancellor's office, the Dean of the College of Social Sciences, and the Department of Political Science
- European Research Council, Grant 724191, PI: Staffan I. Lindberg

The project uses computational resources provided by the Notre Dame Center for Research Computing (CRC) through the High Performance Computing section and the Swedish National Infrastructure for Computing (SNIC) at the National Supercomputer Centre in Sweden. We specifically acknowledge the assistance of In-Saeng Suh at CRC and Johan Raber at SNIC in facilitating our use of their respective systems.