V-Dem INSTITUTE

POLICY BRIEF

No. #17, 2018. Laura Maxwell, Kyle L. Marquardt and Anna Lührmann

The V-Dem Method for Aggregating Expert-Coded Data

Photo by Markus Spiske.

V-Dem has developed innovative methods for aggregating expert judgments in a way that produces valid and reliable estimates of difficult-to-observe concepts. This aspect of the project is critical because many key features of democracy are not directly observable. For example, it is easy to observe and code whether or not a legislature has the legal right to investigate the executive when it engages in corruption. However, assessing the extent to which the legislature actually does so requires the evaluation of experts with extensive conceptual and case knowledge.

In general, expert-coded data raise concerns regarding comparability across time and space. Rating complex concepts requires judgment, which may vary across experts and cases. Moreover, because even equally knowledgeable experts may disagree, it is imperative to report measurement error to the user. We address these issues using both cutting-edge theory and methods, resulting in valid estimates of concepts relating to democracy.

We have recruited over 3,000 country experts to provide their judgment on different concepts and cases. These experts come from almost every country in the world, which allows us to leverage the opinions of experts from a diverse set of backgrounds. We typically gather data from five experts for each observation, which enables us to statistically account for both uncertainty about estimates and potential biases that experts may evince, using a custom-built Bayesian measurement model.

We ask our experts very detailed questions about specific concepts. In addition to being of interest in their own right, experts are better suited to the task of coding specific concepts rather than broader concepts such as "democracy." Box M.1 provides the V-Dem question on academic freedom as an example.

As Box 1 makes clear, we endeavor to both make our questions clear to experts and craft response categories that are not overly open to interpretation. However, we cannot ensure that two experts understand descriptions such as 'somewhat respected' in a uniform way (a response of "2" in Box M.1)—even when 'somewhat' is accompanied by a carefully formulated description. Put simply, one expert's 'somewhat' may be another expert's 'weakly' (a response of "1" in Box M.1), even if they perceive the same level of freedom of expression in a particular country. Of equal importance, all experts code more than one indicator over time, and their level of expertise may vary, making them more or less reliable in different cases.

Pemstein et al. (2018) have developed a Bayesian Item-Response Theory (IRT) estimation strategy that accounts for many of these concerns, while also providing estimates of remaining random measurement error. We use this strategy to convert the ordinal responses experts provide into continuous estimates of the concepts being measured. The basic logic

BOX M1. Question: Is there academic freedom and freedom of cultural expression related to political issues?

Responses:

0: Not respected by public authorities. Censorship and intimidation are frequent. Academic activities and cultural expressions are severely restricted or controlled by the government.

1: Weakly respected by public authorities. Academic freedom and freedom of cultural expression are practiced occasionally, but direct criticism of the government is mostly met with repression.

2: Somewhat respected by public authorities. Academic freedom and freedom of cultural expression are practiced routinely, but strong criticism of the government is sometimes met with repression.

3: Mostly respected by public authorities. There are few limitations on academic freedom and freedom of cultural expression, and resulting sanctions tend to be infrequent and soft.

4: Fully respected by public authorities. There are no restrictions on academic freedom or cultural expression.

TABLE M.1: VERSIONS OF THE V-DEM INDICATORS.

SUFFIX	SCALE	DESCRIPTION	RECOMMENDED USE
None	Interval	Original output of the V-Dem measurement model	Regression analysis
_osp	Interval	Linearized transformation of the measurement model output on the original scale	Substantive interpretation of graphs and data
_ord	Ordinal	Most likely ordinal value taking uncertainty estimates into account	Substantive interpretation of graphs and data
_codelow / _codehigh	Interval	Values approximately one standard deviation above (_codehigh) and below (_codelow) the point estimate	Evaluating differences over time within units
_sd	Interval	Standard deviation of the interval estimate	Creating confidence intervals based on user needs

behind these models is that an unobserved latent trait exists, but we are only able to see imperfect manifestations of this trait. By taking all of these manifest items (in our case, expert ratings) together, we are able to provide an estimate of the trait. In the dataset, we present the user with a best estimate of the value for an observation *(the point estimate)*, as well as an estimate of uncertainty *(the credible regions*, a Bayesian corollary of confidence intervals).

The IRT models we use allow for the possibility that experts have different thresholds for their ratings. These thresholds are estimated based on patterns in the data, and then incorporated into the final latent estimate. In this way, we are able to correct for the previously-discussed concern that one expert's "somewhat" may be another expert's "weakly" (a concept known as Differential Item Functioning). Apart from experts holding different thresholds for each category, we also allow for their reliability (in IRT terminology, their "discrimination parameter") to idiosyncratically vary in the IRT models, based on the degree to which they agree with other experts. Experts with higher reliability have a greater influence on concept estimation, accounting for the concern that not all experts are equally expert on all concepts and cases.

BOX M.2. KEY TERMS.

Point Estimate: A best estimate of a concept's value.
Confidence Intervals: Credible regions for which the upper and lower bounds represent a range of probable values for a point estimate. These bounds are based on the interval in which the measurement model places 68 percent of the probability mass for each score, which is generally approximately equivalent to the upper and lower bounds of one standard deviation from the median.
Significant Differences or Changes: When the upper and lower bounds of the confidence intervals for two point estimates do not overlap, we are confident that the difference between them is real and not a result of measurement error.

To facilitate cross-country comparability, we have encouraged country experts to code multiple countries using two techniques. We refer to the first as **bridge coding**, in which an expert codes the same set of questions for the same time period as the original country they coded. This form of coding is particularly useful when the two countries have divergent regime histories because experts are then more likely to code the full range of the ordinal question scale, providing us with more information as to where an expert's thresholds are. By extension, this information also provides us with a better sense of the thresholds of her colleagues who only coded one of the countries she coded. The second technique is **lateral coding**. This has the purpose of gaining a great deal of information regarding an individual expert's thresholds by asking her to code many different cases that utilize a wide variety of other experts. By comparing her codings to those of many other experts, we are able to gain a greater sense of how she systematically diverges from experts who code other cases; conversely, we also gain information on how those other experts diverge from her. Both of these techniques provide us with more precise and cross-nationally comparable concept estimates.

Finally, we employ **anchoring vignettes** to further improve the estimates of expert-level parameters and thus the concepts we measure. Anchoring vignettes are descriptions of hypothetical cases that provide all the necessary information to answer a given question. Since there is no contextual information in the vignettes, they provide a great deal of information about how individual experts understand the scale itself. Furthermore, since all experts can code the same set of vignettes, they provide insight into how experts systematically diverge from each other in their coding. Incorporating information from vignettes into the model thus provides us with further cross-national comparability in the concept estimates, as well as more precision in the estimates themselves.

The output of the IRT models is an interval-level point estimate of the latent trait that typically varies from -5 to 5, along with the credible regions. These estimates are the best to use for statistical analysis. However, they are difficult for some users to interpret in substantive terms (what does -1.23 mean with regard to the original scale?). We therefore also provide interval-level point estimates that have been linearly transformed back to the original coding scale that experts use to code each case. These estimates typically run from 0 to 4, and users can refer to the V-Dem codebook to substantively interpret them. Finally, we also provide ordinal versions of each variable. Each of the latter two is also accompanied by credible regions.

The end result of this process is a set of versions of indicators of democratic institutions and concepts, along with estimates of uncertainty, allowing both academics and policy-makers alike to understand the features of a polity of interest to them. Table 1 summarizes the output with which we provide users.

REFERENCES

- Marquardt, Kyle L. and Daniel Pemstein. Forthcoming. "IRT Models for Expert-Coded Panel Data." *Political Analysis*.
- Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Joshua Krusell, and Farhad Miri. 2018. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." *University of Gothenburg, Varieties of Democracy Institute:* Working Paper No. 21, 3d edition.
- Pemstein, Daniel, Eitan Tzelgov and Yi-ting Wang. 2015. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." *University of Gothenburg, Varieties of Democracy Institute:* Working Paper No. 1.

ABOUT V-DEM INSTITUTE

V-Dem is a new approach to conceptualization and measurement of democracy. The headquarters – the V-Dem Institute – is based at the University of Gothenburg with 17 staff, and a project team across the world with 6 Principal Investigators, 14 Project Managers, 30 Regional Managers, 170 Country Coordinators, Research Assistants, and 3,000 Country Experts, the V-Dem project is one of the largest ever social science research-oriented data collection programs.

V-Dem INSTITUTE

Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, PO 711
SE 405 30 Gothenburg Sweden
contact@v-dem.net
+46 (0) 31 786 30 43
www.v-dem.net
www.facebook.com/vdeminstitute
www.twitter.com/vdeminstitute