What Makes Experts Reliable?

Kyle L. Marquardt, Daniel Pemstein, Brigitte Seim, Yi-ting Wang

June 2018

Working Paper
SERIES 2018:68
THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (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. The project includes a worldwide team with six 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.

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

V–Dem Working Papers are available in electronic format at www.v-dem.net.

What makes experts reliable?*

Kyle L. Marquardt
V–Dem Institute, Department of Political Science, University of Gothenburg

Daniel Pemstein
Department of Criminal Justice and Political Science, North Dakota State University

Brigitte Seim
Department of Public Policy, University of North Carolina, Chapel Hill

Yi-ting Wang
Department of Political Science, National Cheng Kung University

*Kyle L. Marquardt is first author; second authors listed alphabetically. Earlier drafts presented at the 2016 MPSA Annual Conference, 2016 EIP/V–Dem APSA Workshop, 2018 SPSA Annual Conference and 2018 Annual V–Dem Conference. The authors thank David Armstrong, Ryan Bakker, Ruth Carlitz, Chris Fariss, John Gerring, Adam Glynn, Kristen Kao, Laura Maxwell, Juraj Medzihorsky, Jon Polk, Sarah Repucci, Jeff Staton, Laron Williams and Matthew Wilson for their comments on earlier drafts of this paper, as well as Staffan Lindberg and other members of the V–Dem team for their suggestions and assistance. This material is based upon work supported by the National Science Foundation (SES-1423944, PI: Daniel Pemstein), Riksbankens Jubileumsfond (M13-0559:1, PI: Staffan I. Lindberg), the Swedish Research Council (2013.0166, PI: Staffan I. Lindberg and Jan Teorell); the Knut and Alice Wallenberg Foundation (PI: Staffan I. Lindberg) and the University of Gothenburg (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 at University of Gothenburg. We performed simulations and other computational tasks using resources provided by the Notre Dame Center for Research Computing (CRC) through the High Performance Computing section and the Swedish National Infrastructure for Computing at the National Supercomputer Centre in Sweden (SNIC SNIC 2018/3-133, PI: Staffan I. Lindberg).

Abstract

Many datasets use experts to code latent quantities of interest. However, scholars have not explored either the factors affecting expert reliability or the degree to which these factors influence estimates of latent concepts. Here we systematically analyze potential correlates of expert reliability using six randomly selected variables from a cross-national panel dataset, V–Dem v8. The V–Dem project includes a diverse group of over 3,000 experts and uses an IRT model to incorporate variation in both expert reliability and scale perception into its data aggregation process. In the process, the IRT model produces an estimate of expert reliability, which affects the relative contribution of an expert to the model. We examine a variety of factors that could correlate with reliability, and find little evidence of theoretically-untenable bias due to expert characteristics. On the other hand, there is evidence that attentive and confident experts who have a basic contextual knowledge of the concept of democracy are more reliable.

Many important datasets use experts to code values that are difficult to directly estimate. Weighting estimates of these latent concepts with a measure of relative coder expertise—often proxied by reliability—is of clear importance (Pemstein, Meserve and Melton, 2010). However, the factors that influence expert reliability remain unexplored, as do their implications for model design.

Here we analyze potential correlates of expert reliability in the context of a cross-national expert survey of political concepts. This exploratory analysis provides insight into the degree to which a method—a modified Item Response Theory (IRT) model—provides substantively unbiased estimates of latent concepts using expert-coded data. The analysis also offers evidence regarding characteristics that make some experts more reliable than others, awareness of which will assist future expert-coding endeavors.

We investigate reliability using data from the Varieties of Democracy (V–Dem) Dataset (Coppedge, Gerring, Knutsen, Lindberg, Skaaning, Teorell, Altman, Bernhard, Fish, Cornell, Dahlum, Gjerlow, Glynn, Hicken, Krusell, Lührmann, Marquardt, McMann, Mechkova, Medzihorsky, Olin, Paxton, Pemstein, Pernes, von Römer, Seim, Sigman, Staton, Stepanova, Sundström, Tzelgov, Wang, Wig, Wilson and Ziblatt, 2018). The V–Dem dataset is an ideal laboratory for these analyses because the project utilizes a diverse body of over 3,000 experts to code over 121 ordinal variables covering a variety of regime traits; these variables cover almost all states and many colonies from 1900-2017, as well as a more limited set of variables for 91 cases from 1789-1900 (Coppedge, Gerring, Lindberg, Skaaning, Teorell, Krusell, Marquardt, Medzihorsky, Pemstein, Pernes, Stepanova, Tzelgov, Wang and Wilson, 2018). This variation over cases, concepts and time—as well as differences in both expert demographics and coding characteristics—provides ample data for analyzing the characteristics that could affect expert reliability in a variety of contexts.

We measure reliability using data from six randomly-selected expert-coded variables. As a measure of reliability we use the expert-specific discrimination parameters from an IRT model which aggregates these data. These parameters represent the degree to which an expert diverges from other experts who code the same cases. This operaionalization aligns with classic definitions of reliability (Carmines and Zeller, 1979), as well as recent empirical work examining convergence among workers on crowd-sourcing platforms when coding the same cases (Benoit et al., 2016; Marquardt et al., 2017). As potential correlates of reliability, we use both demographic data from a post-survey questionnaire and the coding characteristics of experts.

In general, the correlates of reliability vary across variables. Most of these null findings regard variables that could constitute potentially problematic sources of bias in the the estimation procedure, such as gender. The exceptions to the null results are intuitive: 1) more confident experts tend to be more reliable than less confident experts; 2) experts who vary their codings tend to be more reliable than experts who vary less; and 3)

experts who evince general knowledge of a concept integral to the coding enterprise tend to be more reliable. Cumulatively, these findings reinforce the argument that IRT models that account for variation in expert reliability and scale perception are a safe method for aggregating expert-coded data (Marquardt and Pemstein, In press).

1 Reliability in the V–Dem model

We use the V–Dem measurement model (Pemstein et al., 2018) as a baseline to estimate expert reliability. Previous research illustrates that this model generally outperforms other standard methods for aggregating expert-coded data in terms of recovering latent trait estimates (Marquardt and Pemstein, In press). However, the degree to which it is subject to other forms of bias remains unexplored.

The model closely resembles a standard Bayesian ordinal IRT model (Johnson and Albert, 1999), with a partial likelihood taking the form of Equation 1.

$$\Pr(y_{ctr} = k) = \phi\left(\gamma_{r,k} - \beta_r z_{ct}\right) - \phi\left(\gamma_{r,k-1} - \beta_r z_{ct}\right). \tag{1}$$

Here, ϕ is the cumulative distribution function of a normal distribution, y_{ctr} is the ordinal response of expert r for country-year ct and z_{ct} is the latent concept being estimated for country-year ct. We follow standard convention and *a priori* assume $z_{ct} \sim \mathcal{N}(0, 1)$.

In this paper we focus on β, the reliability parameter unique to each expert r. In IRT terminology, β is a "discrimination" parameter: $\beta_r = \frac{1}{\sigma_r}$, where σ_r represents each expert's stochastic error variance.[1] We model $\beta_r \sim \mathcal{N}(1, 1)$ *a priori*, restricted to positive values. This truncation is necessary for identification purposes, and rests on the relatively safe assumption that experts code in the right direction.

Note that the model also accounts for systematic biases in how experts translate perceptions into ratings—a common concern in surveys that rely on multi-rater judgment (Aldrich and McKelvey, 1977; Bakker et al., 2014; Hare et al., 2015)—through γ, a $k = 1, ..., n$ vector of threshold parameters specific to each expert. As a result, β estimates the degree to which experts stochastically diverge from other experts who coded the same cases, conditional on her scale perception. Higher scores indicate that the expert diverges less from other experts than experts with lower scores, which we interpret as proxying her reliability in the more colloquial sense.[2]

[1] See Pemstein et al. (2018) for a derivation of this relationship. Appendix A provides more details on the modeling strategy.

[2] Reliability is not necessarily the same as accuracy (Maestas, Buttice and Stone, 2014). However, assessing accuracy directly is an impossible task in this dataset, given that there is no concrete reference point for coding accuracy of latent variables.

2 Benefits of analyzing reliability correlates

A primary benefit of analyzing reliability correlates is that doing so provides a useful diagnostic of the measurement method. In this modified IRT model, experts with lower reliability scores contribute less to the estimation of country-year latent traits, the parameters of interest in most applications. Incorrect measurement of reliability could thus lead to inaccurate latent trait estimates.

In principle, the stringent recruitment criteria of the V–Dem project means that all experts should be exchangeable in their expertise. In practice, there are many ways in which the IRT approach to modeling reliability could go awry, systematically penalizing colloquially "reliable" experts. A key example is gender. A majority of V–Dem experts are men. If women systematically perceive a latent trait differently than men, minority status could lead women to receive lower reliability scores even though their viewpoint is equally valid. Such a result would indicate that there is theoretically untenable bias in the measurement process.

A second benefit of analyzing the correlates of reliability is that it provides tentative evidence regarding the correlates of more reliable experts. This evidence may facilitate decisions regarding expert recruitment and retention in future projects. While previous research illustrates that experts provide better coding in the context of cross-national panel data regarding latent concepts than do laypersons (Marquardt et al., 2017), potential correlates of intra-expert variation in this context remain unexplored.

3 Variables and descriptive statistics

3.1 Reliability

We analyze reliability (β) scores from six of the 121 expert-coded ordinal V–Dem variables over all expert-country-year observations. While the limited number of variables means that our analyses are not exhaustive, the diversity of variables coded militates against finding trends across them: consistent trends are thus likely a function of a consistent relationship between reliability and certain correlates.

We randomly selected all six variables, five from the universe of variables and one from the set of gender-specific variables.[3] The five fully randomly-selected variables (codebook identifier in bold) are: 1) *Executive oversight* by bodies other than the legislature (**v2lgotovst**), 2) opposition *Party autonomy* (**v2psoppaut**) from the ruling regime,

[3]For additional details on the variables, see the V–Dem Codebook (Coppedge, Gerring, Knutsen, Lindberg, Skaaning, Teorell, Altman, Bernhard, Cornell, Fish, Gjerlow, Glynn, Hicken, Krusell, Lührmann, Marquardt, McMann, Mechkova, Olin, Paxton, Pemstein, Seim, Sigman, Staton, Tzelgov, Uberti, Wang, Wig and Ziblatt, 2018). All variables are based on five-point Likert scale questions, with the exception of domestic autonomy and reasoned justification, which have three and four points, respectively.

3) the degree to which officials offer *Reasoned justification* (**v2dlreason**) for their decisions, 4) the degree to which a government has *Domestic autonomy* (**v2svdomaut**) from other states, and 5) the degree to which a state engages in *Journalist harassment* (**v2meharjrn**).

We also randomly selected one variable from the universe of gender-specific variables: *Female freedom* of discussion (**v2cldiscw**). This variable represents a most likely case where we would expect theoretically-untenable systematic differences in reliability with regard to gender.

We use Markov chain Monte Carlo (MCMC) methods to estimate the IRT model for each of the variables included in the analysis.[4] MCMC methods generate samples from the posterior distributions of model parameters; we use the full posterior of reliability estimates across iterations of the MCMC algorithm to account for measurement error.

3.2 Correlates of reliability

We discuss potential sets of reliability correlates in turn. All coding characteristic variables regard expert behavior in coding the variable for which reliability is being analyzed; variables regarding self-reported confidence and coding variation use reduced data.[5] Appendix B presents descriptive statistics.

3.2.1 Demographics

Gender may influence reliability for the theoretically-untenable reasons previously discussed; we therefore include the dichotomous indicator *Female*. We also control for educational background and employment, which have a more theoretical tenable connection to reliability. Previous research demonstrates that experts with different academic and professional backgrounds can have different knowledge and thus vary in their perception of aspects of latent traits (Cumming, 1990; Michael et al., 1980; Royal-Dawson and Baird, 2009). Similarly, raters with higher levels of expertise are more reliable when rating complex or broad tasks (Schoonen, Vergeer and Eiting, 2016). A majority of V-Dem experts hold a PhD and/or work at a university, both of which potentially indicate expertise and thus greater reliability.

[4]We conduct all analyses using the statistical software Stan (Stan Development Team, 2015). We ran eight chains with 10,000 iterations (burn-in of 1,000 iterations and thinning interval of 20), quintupling the iterations, burn-in and thinning interval if the initial run did not converge. We assess convergence using the Gelman diagnostic, considering reliability scores to have reached convergence if less than 10 percent of reliability scores have values below 1.1.

[5]When running models we reduce the data to regimes—country-year observations where at least one expert changes her coding or self-reported confidence—to prevent inaccurately low estimates of uncertainty (Pemstein et al., 2018). We analyze coding variation at the reduced level because these cases are those in which at least one expert has changed her codings, indicating a potential change in the latent trait value.

We trichotomize education: the reference level is individuals with a PhD, while *Professional degree* indicates that the expert holds a degree such as an MBA or JD, while *MA or lower* is self-explanatory. We analyze employment with a set of four indicators: employees of a *Public university* (the reference level), *Private university*, the *Government* (non-university government employment, including employees of regional governments and state-owned enterprises), and *Other* (non-governmental non-academic employment). We disaggregate public and private employment because experts in the private sector may be more reliable, since they are potentially less susceptible to government pressure or other incentives to provide biased estimates.

Finally, we include the natural logarithm of a respondent's *Age* as a standard control. Similarly, we include an indicator for *Historical coders*, or those coders who coded pre-1900 data. These coders diverge from others in that they are generally the sole coder for pre-1900 data, which could affect their reliability for mechanical—as opposed to substantive—reasons.

3.2.2 Democracy in residence country

Experts living in democratic countries may not be concerned by potential government sanction, and may have better access to information. Both of these factors may increase their reliability. *Democracy* represents the average level of democracy from 2008 to 2017 for experts' residence country, measuring democracy with the V–Dem electoral democracy index ($v2x\_polyarchy$).

3.2.3 Knowledge

We proxy case knowledge with an indicator for experts who are *Not resident* of the country they are coding. We also measure both conceptual awareness and general knowledge. The indicator *Low awareness* represents experts who reported in a post-survey questionnaire that they do not consider electoral democracy important to the concept of democracy. Since electoral democracy underpins most definitions of democracy, experts who are not aware of this connection may be less reliable. The indicator *Low knowledge* represents experts who either consider 1) Sweden to be non-democratic or 2) North Korea to be democratic.[6] Since Sweden consistently ranks among the most democratic countries in the world and North Korea among the least, an expert who miscodes either case betrays a lack of understanding of the concept of democracy or extremely limited understanding of the worldwide context. Such an expert may be less reliable.

[6]Experts rank 12 countries on a 0-100 scale, with high scores representing more democracy. We consider a score on either side of 50 to represent democracy vs. non-democracy.

3.2.4 Confidence

Experts self-report their case-level *Confidence* on a 0-1 scale, which we aggregate to an expert's average over a given variable. We expect more confident experts to be more reliable.

3.2.5 Attentiveness

More attentive experts may be more reliable. We measure attentiveness with two sets of indicators. First, most countries vary in political traits over time. After controlling for country-coded effects, the degree to which an expert varies her scores may therefore proxy her attentiveness. Second, since expertise likely varies over time and countries, attentive experts should vary in self-reported confidence. We measure both variation in coding and confidence with two indicators each. *Coding* and *Confidence variation* indicate if an expert changed her scores on either metric at least once, while *Coding* and *Confidence sd* measure an expert's standard deviation on these metrics.

3.2.6 Volume

High coding volume may lead experts to overextend themselves and thus lower their reliability. We measure coding volume along three dimensions. First, the natural logarithm of the country-years an expert coded, *Country-years*. Second, the natural logarithm of variables an expert coded, *Variables*. Third, though most experts coded only one country, many coded several. We include both *Countries* > 1, which indicates an expert who coded more than one country and *Unique countries*, the natural logarithm of the unique countries she coded.

4 Results

We conduct analyses of each variable's reliability scores individually, regressing each posterior draw of reliability parameters on the complete set of potential correlates.[7] Given that some countries and years may be more difficult to code than others, we include fixed effects for the coded country and year in all analyses. [8]

[7] Appendix D presents results from analyses that only analyze the relationship between the correlates and the posterior median, which are in line with these analyses, albeit with much tighter estimates of uncertainty since they do not incorporate posterior measurement error. Since approximately 50 percent of experts do not complete the post-survey questionnaire, we present replications of analyses with only coding characteristics and an indicator for experts who did not complete the questionnaire in Appendix C. The results are in line with those we present in the body of the text; experts who did not complete the post-survey question tend to have similar reliability to those who did, conditional on their coding characteristics. Finally, Appendix C also includes analyses of models that do not include coding characteristics. Again, results are essentially in line with those in the main text.

[8] Models with only country and year fixed effects explain a fair amount of variance, with bootstrapped posterior median r^2 values ranging from 0.16 to 0.19.

We first show coefficient plots, then discuss their substantive implications with the predicted reliability of experts with different characteristics. Figure 1 presents coefficient estimates by variable, with points representing the bootstrapped median coefficient estimate and horizontal lines the 90 percent highest bootstrapped density about this estimate. The vertical line aligns with effect magnitude of zero; we center the intercept at zero for illustration purposes. We discuss results by expert characteristic type.

4.1 Demographics

The difference between female and male coders is generally low in magnitude and inconsistent across variables, indicating that the model does not erroneously penalize female experts. Similarly, age and employment shows little correlation with reliability. Respondents with a professional degree tend to have higher reliability than experts with a PhD (the reference level) in four of the six variables with a relatively high magnitude, though these estimates are based on a relatively small number of experts. Results regarding the other education indicator—MA or lower education—are ambiguous and relatively small in magnitude. Experts who code historical data tend to be less reliable than other experts in four of the five variables (there are no historical data for Reasoned justification) though this result may be a relic of these experts generally being the sole coders of cases.

4.2 Democracy in residence country

Democracy shows little relationship with expert reliability in four of the six variables, and a negative relationship in one variable (Reasoned justification) and a positive relationship in the remaining variable (Female freedom).

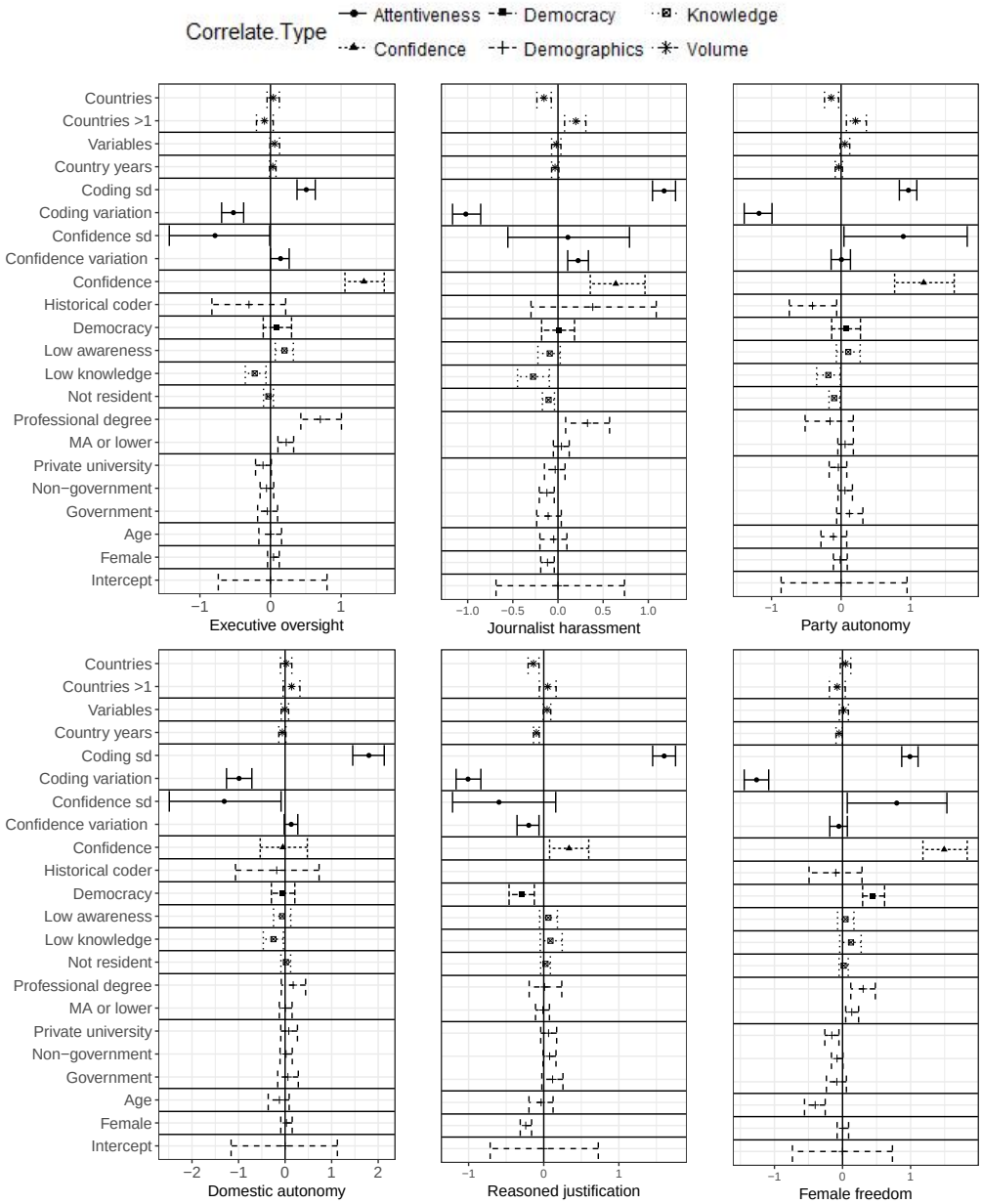
4.3 Knowledge

Experts who betray a lack of general knowledge of democracy are less reliable than other experts in four of the six variables, and slightly more reliable in the remaining two. However, the magnitude of this relationship is generally small. The remaining knowledge measures (Not resident and Low awareness) show little consistent relationship with reliability.

4.4 Confidence

In five of the six variables self-reported confidence shows a positive correlation with reliability; in the remaining variable there is little evidence of a relationship.

Figure 1: Bootstrapped posterior coefficient estimates of correlates of reliability



Intercept estimate centered at zero for illustration purposes. Models include country and year fixed effects; reference level is an expert coding Germany and the year 2012.

4.5 Attentiveness

Variation in coding shows the most consistent results in these analyses: in all variables, experts who varied more in their coding tend to have higher reliability than their peers who varied less. However, results regarding the difference between those experts who did not vary their codings and their peers are inconsistent, which may be due to the relative lack of variation in latent concept levels in some cases across variables. Variation in self-reported confidence shows little correlation with expert reliability.

4.6 Volume

Neither the number of country-years an expert coded nor the number of variables she coded shows a relationship with reliability in any variable. Results regarding the number of unique countries an expert coded are inconsistent: in two variables they show little correlation with reliability; in three, experts who coded more than two variables tend to be slightly less reliable than those who only coded two; and in one variable, experts who coded two or more countries tend to be more reliable than those who only coded one.

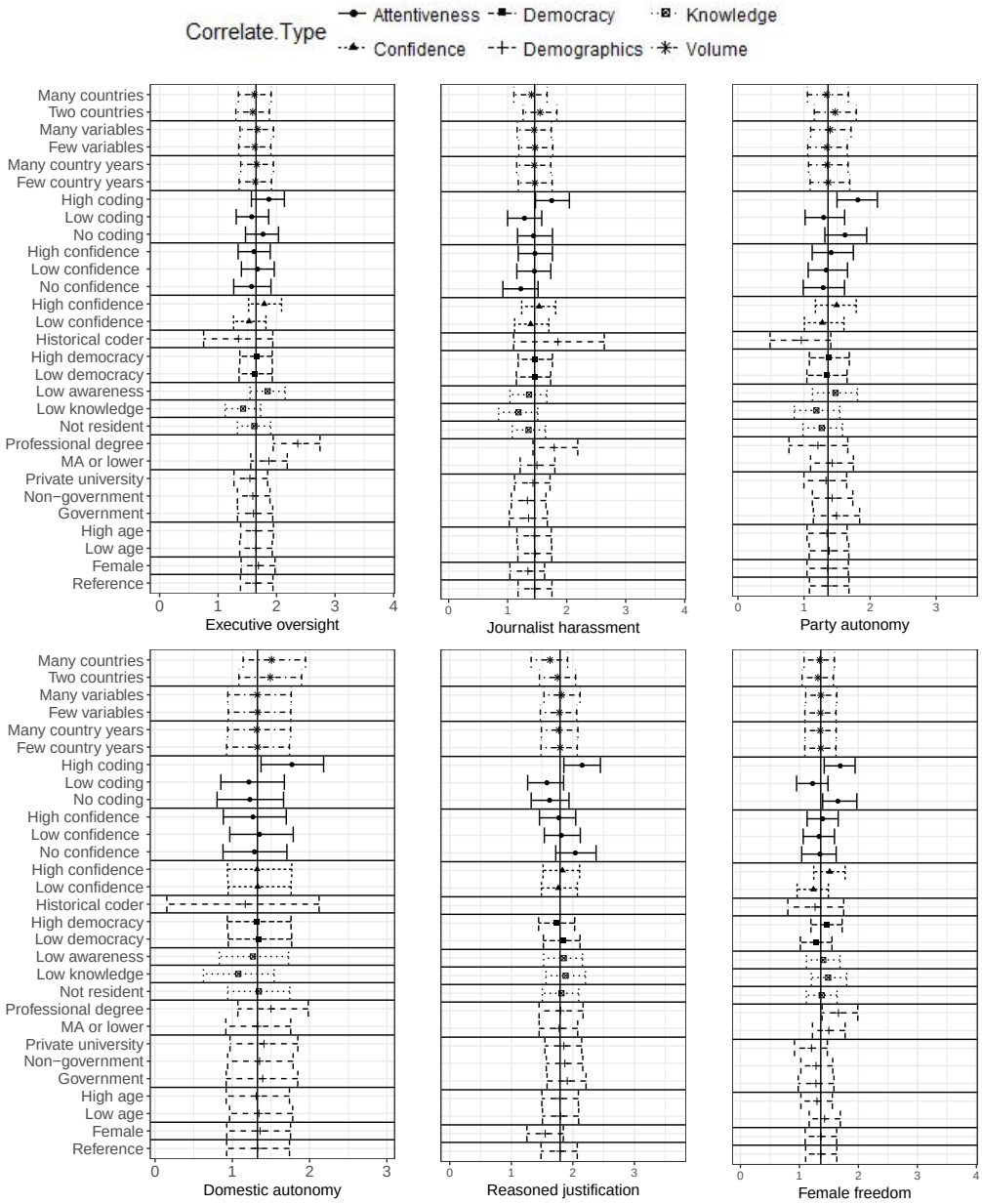
4.7 Predicted reliability

The coefficient plots illustrate that there is a high level of uncertainty in the intercept, which indicates that coefficient plots may be misleading in their illustration of the substantive importance of the covariates. Figure 2 presents the predicted reliability of experts with different characteristics across variables. Points represent the the bootrapped predicted median reliability for experts with given certain demographic or coding characteristics, holding all other correlates constant at their mean or mode.[9] The range represents the posterior median range of reliability scores for a given variable.

As Figure 2 makes clear, once we incorporate overall posterior uncertainty into the assessment of the relationship between the correlates of reliability and this outcome, the substantive relationship is generally minimal. The main exceptions to this rule are Confidence, Low knowledge, and Coding variation, which retain their relatively large correlation with reliability. However, this estimation procedure likely underestimates uncertainty; if we were to illustrate the results by using the method of composition the relationship between these variables and reliability would likely further diminish.

[9] In the case of continuous correlates, we plot predicted values at their second and fourth quantile ("low" and "high"); for those variables that include a dichotomous indicator (unique countries coded, and variation in coding and confidence), we report both the relationship between the dichotomous indicator of some variation (*Two countries coded*, *No coding*, and *No confidence*) and *High* and *Low* estimates based on the quantiles of experts who show variation in either self-reported confidence or coding (most experts who coded more than one country coded only two).

Figure 2: Posterior bootstrapped predicted reliability of experts with different characteristics



Models include country and year fixed effects; reference level is an expert coding Germany and the year 2012.

5 Conclusion

The analyses in this paper represent a valuable first step in analyzing correlates of expert reliability, using the diverse body of experts who code a variety of political traits cross-nationally and cross-temporally for the V–Dem Project. Most potential correlates of reliability show little substantively important correlation with the measure; in general, these null results provide evidence that the IRT model we use in the estimation procedure does not provide results that are biased for untenable reasons. The main exception to this rule is coding variation: experts who vary their codings less than others tend to be less reliable across variables, conditional on them changing their codings at least once. This result indicates that more attentive experts tend to be more reliable.

Other results are more tentative. There is evidence that experts who show low general conceptual knowledge are less reliable than others, and that those experts who are more confident in their codings are more reliable. These results lead to the intuitive conclusion that expert-coding enterprises should endeavor to recruit experts who have knowledge of the concepts they are coding and believe they have knowledge on the concepts and cases.

References

Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1):111–130.

Bakker, Ryan, Seth Jolly, Jonathan Polk and Keith Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *The Journal of Politics* 76(4):1089–1101.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Carmines, Edward G and Richard A Zeller. 1979. *Reliability and validity assessment.* Vol. 17 Sage Publications.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Haakon Gjerlow, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Moa Olin, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Aksel Staton, Jeffreyand Sundstöm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig and Daniel Ziblatt. 2018. Varieties of Democracy Codebook v8. Technical report Varieties of Democracy Project: Project

Documentation Paper Series.
Accessed at: https://ssrn.com/abstract=3172791

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Agnes Cornell, Sirianne Dahlum, Haakon Gjerlow, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2018. V–Dem Dataset v8. Technical report Varieties of Democracy Project.
Accessed at: https://ssrn.com/abstract=3172819

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Joshua Krusell, Kyle L. Marquardt, Juraj Medzihorsky, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov, Yi-Ting Wang and Steven Wilson. 2018. Varieties of Democracy Methodology v8. Technical report Varieties of Democracy Project: Project Documentation Paper Series.
Accessed at: https://ssrn.com/abstract=3172796

Cumming, Alister. 1990. "Expertise in evaluating second language compositions." *Language Testing* 7(1):31–51.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Johnson, Valen E and James H Albert. 1999. *Ordinal Data Modeling*. New York: Springer.

Kozlowski, Steve W. and Keith Hattrup. 1992. "A disagreement about within-group agreement: Disentangling issues of consistency versus consensus." *Journal of Applied Psychology* 77(2):161–167.

Maestas, Cherie D., Matthew K. Buttice and Walter J. Stone. 2014. "Extracting wisdom from experts and small crowds: Strategies for improving informant-based measures of political concepts." *Political Analysis* 22(3):354–373.

Marquardt, Kyle L. and Daniel Pemstein. In press. "IRT models for expert-coded panel data." *Political Analysis* .

Marquardt, Kyle L., Daniel Pemstein, Constanza Sanhueza, Brigitte Seim, Steven L. Wilson, Michael Bernhard, Michael Coppedge and Staffan I. Lindberg. 2017. "Experts, Coders, and Crowds: An analysis of substitutability." *Varieties of Democracy Institute Working Paper* 53.

Michael, William B., Terri Cooper, Phyllis Shaffer and Earl Wallis. 1980. "A Comparison of the Reliability and Validity of Ratings of Student Performance on Essay Examinations by Professors of English and by Professors in Other Disciplines." *Educational and Psychological Measurement* 40(1):183–195.

Pemstein, Daniel, Eitan Tzelgov and Yi-ting Wang. 2015. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." *Varieties of Democracy Institute Working Paper* 1(March):1–53.

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." *Varieties of Democracy Institute Working Paper* 21.

Pemstein, Daniel, Stephen A. Meserve and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.

Royal-Dawson, Lucy and Jo-Anne Baird. 2009. "Is Teaching Experience Necessary for Reliable Scoring of Extended English Questions?" *Educational Measurement: Issues and Practice* 28(2):2–8.

Schoonen, Rob, Margaretha Vergeer and Mindert Eiting. 2016. "The assessment of writing ability: expert readers versus lay readers:." *Language Testing* .

Stan Development Team. 2015. "Stan: A C++ Library for Probability and Sampling, Version 2.9.0.".
Accessed at: http://mc-stan.org/

A Additional model details

The model assumes that experts imperfectly observe an interval-valued latent variable, z_{ct}, when producing their ordinal ratings for each country year, y_{ctr}. These parameters represent the concept that experts are asked to evaluate when answering a given survey question. More reliable raters observe each z_{ct} with less error than less reliable raters. In particular, each rater r perceives the latent trait for case ct such that

$$p_{ctr} = z_{ct} + e_{ctr}, \text{where } e_{ctr} \sim N(0, \sigma_r). \tag{A.1}$$

While the z parameters are the main quantity of interest of the V-Dem measurement model, they are of secondary concern here.[10] Second, to account for idiosyncratic interpretation of the ordinal categories (differential item functioning, or DIF), the model estimates unique thresholds (each γ) for each expert. These thresholds are cutoff points on the underlying latent scale that determine how experts translate interval-level perceptions of the latent traits into ordinal ratings. If rater r perceives p_{ct} such that $p_{ct} < \gamma_{r,1}$, she reports the lowest possible ordinal score for case ct (i.e $y_{ctr} = 0$); if she observes p_{ct} such that $\gamma_{r,1} < p_{ct} < \gamma_{r,2}$, she reports the second lowest ordinal score; and so on. There is a large body of literature on the importance of accounting for variation in how experts code latent concepts (Kozlowski and Hattrup, 1992): namely, experts may disagree about question scales, though they are in agreement about the latent value. As a result, a model that only estimates reliability without adjusting for systematic bias risks conflating bias with a unreliability. For example, an expert who consistently ranks a concept one item higher than other experts, but otherwise follows their trends in coding, is as reliable as the other experts. Expert-specific thresholds ameliorate this concern by allowing for inter-expert variation in how they map their understanding of latent concepts into ordinal codes. They allow the model to adjust for a large class of systematic rater biases that lead to inter-expert disagreement in coding. For example, some experts may have higher standards than others, and threshold parameters account for this sort of systematic error.

We follow the standard V-Dem framework for estimating thresholds hierarchically (Pemstein et al., 2018): each expert's unique thresholds use the same prior as the thresholds of similar experts (i.e., experts who were recruited to code the same country as their main country). Equation A.2 provides a more precise description of this estimation strategy.

[10] The standard V–Dem model uses a confidence-weighted empirical prior for z_{ct} to correct for sparse data in some country years (Pemstein et al., 2018), we use a vague $\mathcal{N}(0, 1)$ prior. The use of this prior has no bearing on estimates of rater reliability, which are purely a function of inter-rater agreement. Thus, we avoid this complication here.

$$\begin{aligned} \gamma_{r,k} &\sim \mathcal{N}(\gamma_k^{c_r}, 0.2) \\ \gamma_k^c &\sim \mathcal{N}(\gamma_k^{\mu}, 0.2) \\ \gamma_k^{\mu} &\sim U(-4, 4) \end{aligned} \tag{A.2}$$

In Equation A.2, γ_k^{μ} represents the overall population threshold μ for category k; γ_k^c the overall threshold for experts with a common main country-of-coding c, and $\gamma_{r,k}$ the expert-r specific threshold. For the purposes of this paper, this framework means that the reliability scores we analyze are mainly a function of similar patterns in coding, as opposed to DIF.

The hierarchical structure facilitates more precise estimation of latent concepts in two ways. First, V–Dem encourages experts to code either another country for the entire time series in addition to their main country (bridge coding), or multiple countries in a given year (lateral coding) (Pemstein, Tzelgov and Wang, 2015). By hierarchically-clustering thresholds about a main country coded, we are able to incorporate information regarding the coding patterns of experts who coded more than one country into the thresholds of experts who only coded one country.

Second, many experts do not code the entire scale (e.g., an expert who only codes Freedom of discussion for women from 2012-2017 in Switzerland will likely never report a scale item representing systematic prevention of political discussion). To accurately estimate such an expert's thresholds, we borrow strength from experts who code both the same country as her, as well as countries with greater variation in this latent concept. In other words, when we lack the information necessary to estimate a particular expert's threshold(s), we assume that she behaves like other, similar, raters.

B Descriptive statistics and regression table

Figure B.1: Posterior median reliability distribution by expert

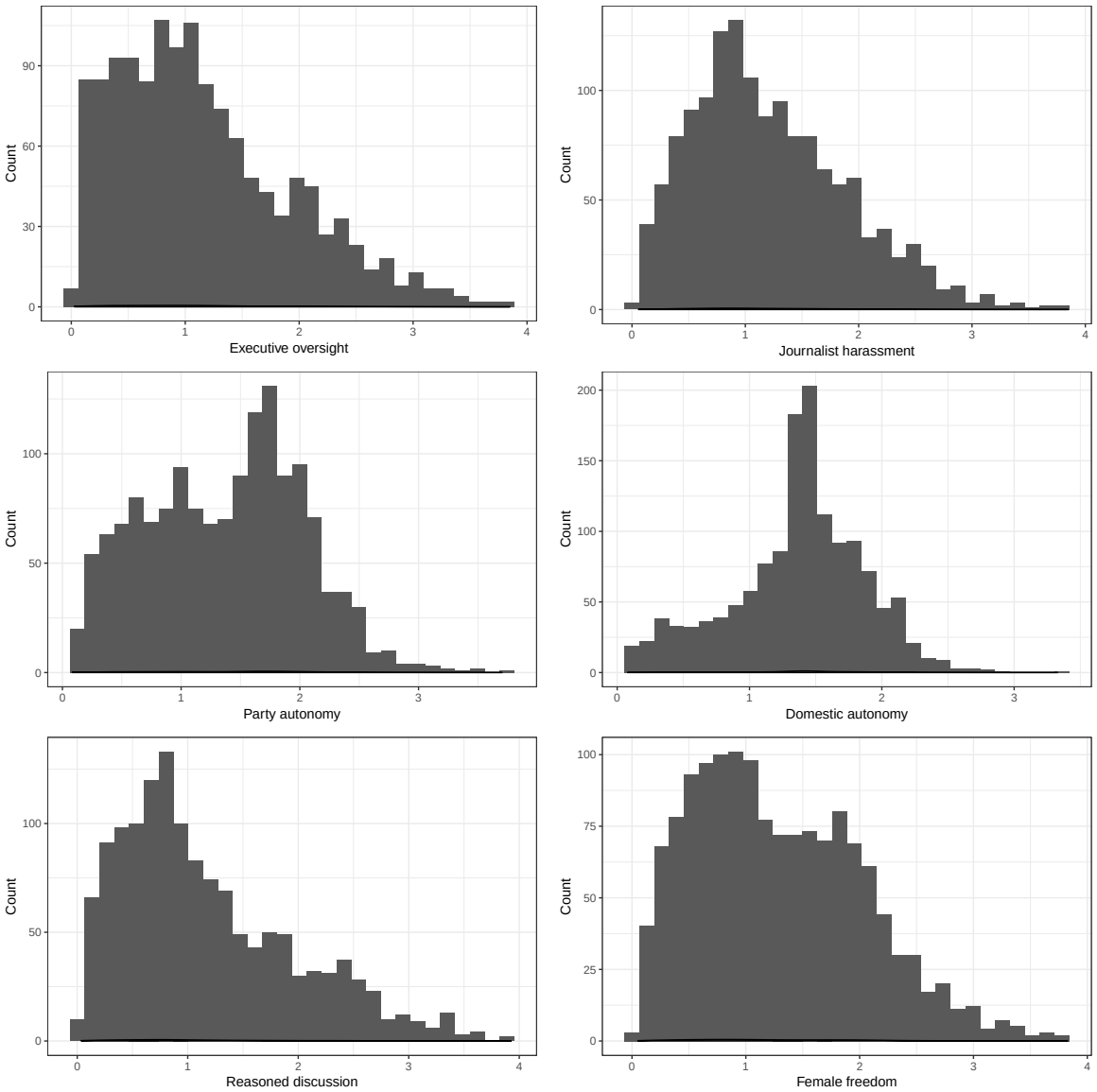


Figure B.2: Posterior distribution of reliability for every fiftieth expert across variables

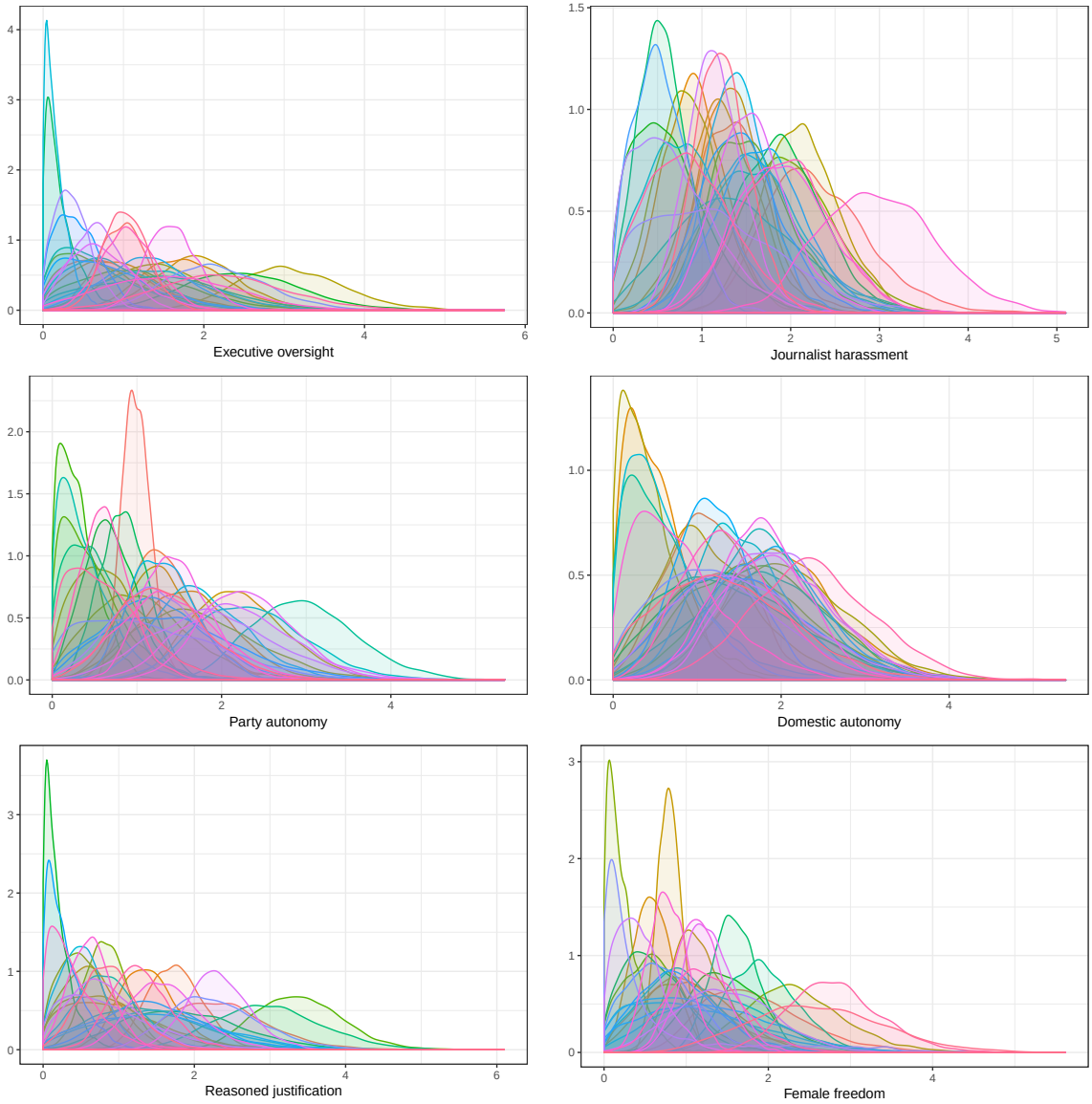


Table B.1: Descriptive statistics

	Executive oversight	Journalist harassment	Party autonomy	Domestic autonomy	Reasoned justification	Female freedom
Posterior median	1.16 (0.77)	1.20 (0.70)	1.37 (0.70)	1.40 (0.54)	1.20 (0.80)	1.26 (0.74)
N unique coders	1,350	1,435	1,467	1,392	1,375	1,439
N	77,462	102,237	94,305	109,553	102,850	113,284
N w/full demographics	42,834	54,327	50,373	54,935	58,923	56,385
N countries	195	197	192	198	182	198
N historical dates	332	323	368	422	375	377
Female	0.22	0.26	0.20	0.22	0.21	0.23
Age	3.82 (0.23)	3.83 (0.23)	3.83 (0.23)	3.82 (0.23)	3.83 (0.22)	3.82 (0.23)
Public university	0.53	0.46	0.50	0.49	0.50	0.47
Private university	0.12	0.12	0.13	0.11	0.11	0.13
Government	0.06	0.07	0.07	0.07	0.07	0.06
Non-government	0.29	0.35	0.30	0.32	0.32	0.33
PhD	0.73	0.69	0.70	0.69	0.71	0.70
Professional degree	0.02	0.02	0.01	0.04	0.02	0.04
MA or lower	0.25	0.29	0.28	0.27	0.26	0.27
Not resident	0.50	0.50	0.46	0.47	0.51	0.50
Low knowledge	0.04	0.04	0.04	0.05	0.05	0.05
Low awareness	0.07	0.08	0.06	0.07	0.06	0.08
Democracy	0.69 (0.23)	0.67 (0.24)	0.68 (0.23)	0.67 (0.24)	0.67 (0.24)	0.67 (0.24)
Historical coder	0.06	0.09	0.07	0.09	NA	0.09
Confidence	0.80 (0.15)	0.74 (0.16)	0.84 (0.14)	0.89 (0.12)	0.77 (0.15)	0.79 (0.15)
Confidence variation	0.78	0.88	0.77	0.70	0.85	0.84
Confidence sd*	0.11 (0.07)	0.13 (0.08)	0.11 (0.08)	0.09 (0.06)	0.10 (0.06)	0.10 (0.06)
Coding variation	0.79	0.89	0.80	0.83	0.90	0.89
Coding sd*	0.99 (0.38)	0.93 (0.33)	1.16 (0.41)	0.73 (0.21)	0.79 (0.26)	1.04 (0.36)
Country-years	4.40 (0.72)	4.65 (0.73)	4.55 (0.74)	4.71 (0.67)	4.68 (0.72)	4.72 (0.69)
Variables	4.26 (0.56)	4.08 (0.72)	4.11 (0.70)	4.13 (0.75)	4.02 (0.85)	4.22 (0.55)
Country > 1	0.40	0.42	0.38	0.42	0.40	0.42
Countries*	1.16 (0.53)	1.18 (0.53)	1.16 (0.52)	1.19 (0.57)	1.17 (0.56)	1.19 (0.57)

All statistics represent the proportion for dichotomous indicators and mean and standard deviation for continuous variables across all observations (i.e. expert $\times$ country-year cases). *Mean and standard deviation for experts who evinced variation in relevant variable.

Table B.2: Bootstrapped posterior regression results

	Executive oversight	Journalist harassment	Female freedom	Party autonomy	Domestic autonomy	Reasoned justification
Intercept	0.17 (-0.56, 0.97)	1.12 (0.44, 1.86)	1.85 (1.12, 2.59)	0.83 (-0.04, 1.78)	2.05 (0.89, 3.18)	2.23 (1.52, 2.95)
Female	0.04 (-0.04, 0.12)	-0.12 (-0.19, -0.04)	0 (-0.08, 0.09)	-0.01 (-0.11, 0.09)	0.03 (-0.09, 0.15)	-0.24 (-0.31, -0.16)
Age	0 (-0.16, 0.16)	-0.05 (-0.2, 0.1)	-0.4 (-0.56, -0.25)	-0.11 (-0.29, 0.08)	-0.12 (-0.36, 0.09)	-0.04 (-0.2, 0.12)
Government	-0.05 (-0.18, 0.1)	-0.11 (-0.24, 0.04)	-0.08 (-0.24, 0.06)	0.12 (-0.06, 0.31)	0.06 (-0.16, 0.28)	0.12 (-0.03, 0.26)
Non-government	-0.06 (-0.15, 0.05)	-0.12 (-0.21, -0.04)	-0.08 (-0.16, 0.01)	0.06 (-0.04, 0.16)	0.02 (-0.11, 0.15)	0.08 (-0.01, 0.16)
Private university	-0.11 (-0.21, 0.01)	-0.03 (-0.15, 0.08)	-0.16 (-0.26, -0.05)	-0.04 (-0.17, 0.08)	0.08 (-0.09, 0.27)	0.06 (-0.04, 0.17)
MA or lower	0.22 (0.11, 0.33)	0.04 (-0.05, 0.12)	0.14 (0.05, 0.24)	0.06 (-0.05, 0.18)	0 (-0.12, 0.15)	-0.01 (-0.11, 0.08)
Professional degree	0.71 (0.43, 1.01)	0.33 (0.08, 0.57)	0.3 (0.12, 0.48)	-0.16 (-0.52, 0.18)	0.17 (-0.08, 0.44)	0.01 (-0.19, 0.24)
Not resident	-0.03 (-0.1, 0.04)	-0.1 (-0.17, -0.04)	0.02 (-0.05, 0.08)	-0.1 (-0.17, -0.01)	0.02 (-0.09, 0.12)	0.02 (-0.04, 0.09)
Low knowledge	-0.22 (-0.36, -0.06)	-0.28 (-0.45, -0.1)	0.12 (-0.04, 0.27)	-0.18 (-0.35, -0.01)	-0.25 (-0.47, -0.04)	0.09 (-0.05, 0.25)
Low awareness	0.2 (0.07, 0.32)	-0.09 (-0.22, 0.02)	0.04 (-0.07, 0.17)	0.1 (-0.07, 0.27)	-0.06 (-0.25, 0.12)	0.06 (-0.05, 0.18)
Democracy	0.09 (-0.1, 0.3)	0 (-0.18, 0.18)	0.45 (0.29, 0.62)	0.07 (-0.14, 0.28)	-0.05 (-0.29, 0.21)	-0.3 (-0.46, -0.13)
Historical coder	-0.3 (-0.83, 0.21)	0.38 (-0.3, 1.09)	-0.1 (-0.49, 0.29)	-0.41 (-0.74, -0.06)	-0.18 (-1.07, 0.73)	NA
Confidence	1.32 (1.05, 1.61)	0.64 (0.36, 0.96)	1.49 (1.18, 1.83)	1.19 (0.77, 1.63)	-0.05 (-0.53, 0.48)	0.34 (0.08, 0.6)
Confidence variation	0.14 (0, 0.26)	0.22 (0.11, 0.34)	-0.05 (-0.19, 0.07)	0.01 (-0.14, 0.13)	0.13 (-0.02, 0.27)	-0.2 (-0.36, -0.06)
Confidence sd	-0.79 (-1.43, -0.01)	0.11 (-0.56, 0.79)	0.8 (0.07, 1.53)	0.89 (0.04, 1.81)	-1.31 (-2.49, -0.09)	-0.6 (-1.21, 0.16)
Coding variation	-0.53 (-0.69, -0.38)	-1.02 (-1.17, -0.86)	-1.26 (-1.44, -1.08)	-1.18 (-1.39, -0.99)	-0.99 (-1.26, -0.71)	-1.01 (-1.17, -0.84)
Coding sd	0.51 (0.38, 0.63)	1.17 (1.05, 1.3)	0.99 (0.87, 1.11)	0.97 (0.84, 1.09)	1.8 (1.45, 2.13)	1.6 (1.45, 1.75)
Country years	0.03 (-0.01, 0.08)	-0.03 (-0.07, 0.01)	-0.05 (-0.1, 0)	-0.03 (-0.08, 0.02)	-0.06 (-0.14, 0.02)	-0.1 (-0.14, -0.06)
Variables	0.06 (-0.01, 0.13)	-0.02 (-0.07, 0.03)	0.02 (-0.05, 0.09)	0.06 (-0.02, 0.12)	-0.01 (-0.08, 0.08)	0.05 (-0.01, 0.09)
Countries >1	-0.09 (-0.2, 0.04)	0.2 (0.07, 0.31)	-0.08 (-0.19, 0.04)	0.21 (0.08, 0.36)	0.14 (-0.04, 0.32)	0.05 (-0.06, 0.16)
Countries	0.04 (-0.05, 0.13)	-0.15 (-0.23, -0.07)	0.04 (-0.03, 0.12)	-0.14 (-0.24, -0.03)	0.03 (-0.1, 0.14)	-0.14 (-0.21, -0.06)
N	42,834	54,327	56,385	50,373	54,935	58,923
r^2	0.35 (0.32, 0.38)	0.51 (0.46, 0.55)	0.54 (0.49, 0.58)	0.44 (0.41, 0.48)	0.43 (0.37, 0.48)	0.42 (0.37, 0.47)

All models include country and historical date fixed effects. Quantities represent the bootstrapped coefficent median over the posterior distribution and 90 percent highest density about this estimate.

C Results from additional analyses

Figure C.1: Demographic correlates of reliability

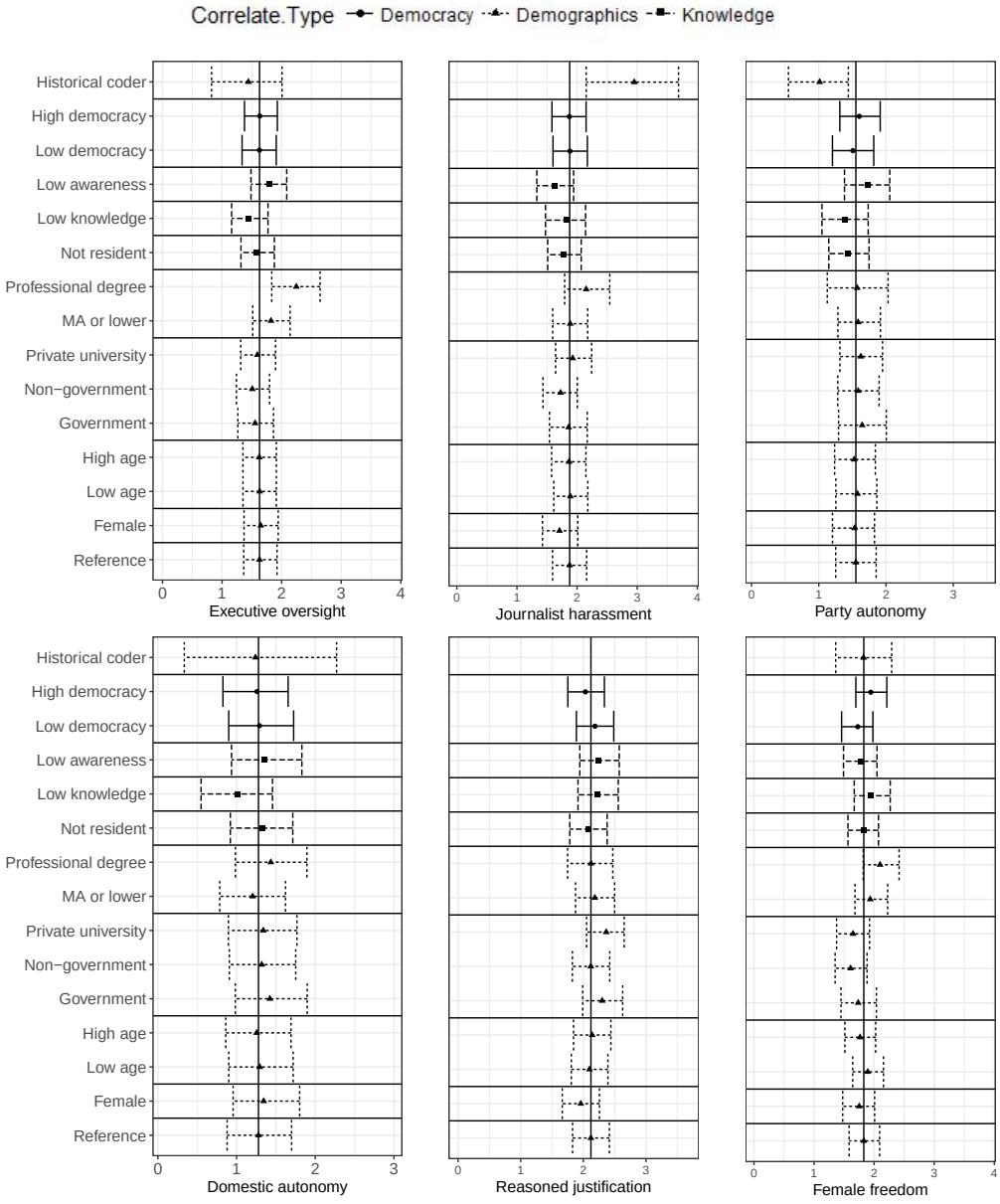


Figure C.2: Coding correlates of reliability

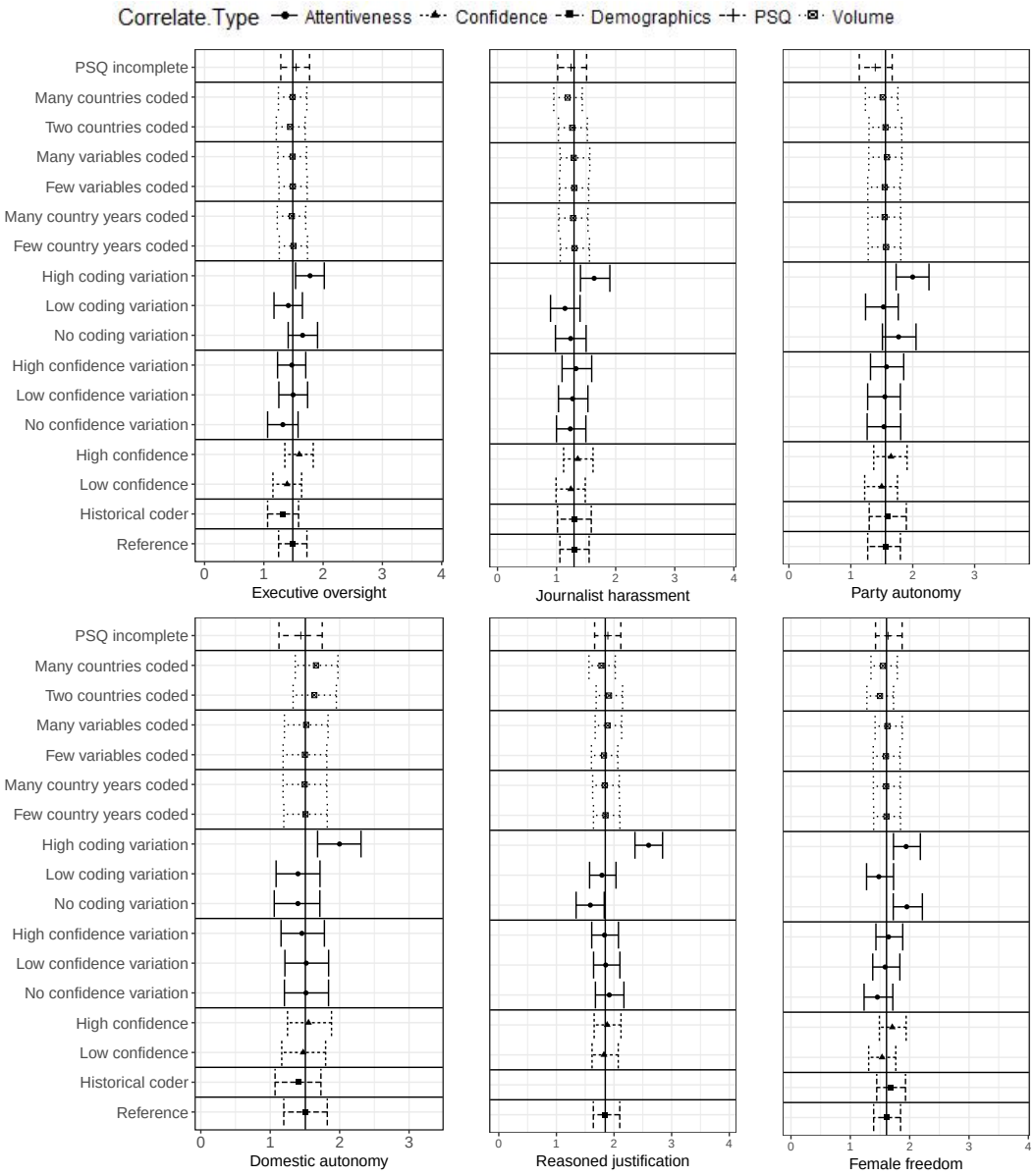


Table C.1: Bootstrapped regression results, demographic correlates

	Executive oversight	Journalist harassment	Female freedom	Party autonomy	Domestic autonomy	Reasoned justification
Intercept	1.72 (1.05, 2.42)	2.18 (1.53, 2.81)	3.03 (2.4, 3.66)	1.94 (1.13, 2.73)	1.78 (0.87, 2.79)	1.85 (1.23, 2.56)
Female	0.02 (-0.06, 0.1)	-0.17 (-0.25, -0.1)	-0.07 (-0.16, 0.01)	-0.02 (-0.12, 0.07)	0.07 (-0.05, 0.2)	-0.16 (-0.24, -0.09)
Age	-0.03 (-0.19, 0.13)	-0.07 (-0.22, 0.08)	-0.4 (-0.55, -0.24)	-0.14 (-0.33, 0.05)	-0.12 (-0.36, 0.09)	0.13 (-0.03, 0.29)
Government	-0.08 (-0.21, 0.07)	-0.02 (-0.15, 0.12)	-0.09 (-0.24, 0.05)	0.09 (-0.09, 0.28)	0.14 (-0.08, 0.35)	0.18 (0.04, 0.32)
Non-government	-0.12 (-0.22, -0.03)	-0.15 (-0.23, -0.07)	-0.22 (-0.31, -0.14)	0.03 (-0.07, 0.13)	0.04 (-0.08, 0.18)	0 (-0.09, 0.08)
Private university	-0.03 (-0.14, 0.08)	0.05 (-0.06, 0.17)	-0.18 (-0.29, -0.08)	0.07 (-0.05, 0.19)	0.06 (-0.12, 0.23)	0.24 (0.14, 0.35)
MA or lower	0.19 (0.08, 0.31)	0.01 (-0.08, 0.1)	0.11 (0.01, 0.2)	0.03 (-0.07, 0.15)	-0.07 (-0.2, 0.07)	0.06 (-0.03, 0.16)
Professional degree	0.61 (0.33, 0.9)	0.27 (0.03, 0.51)	0.27 (0.09, 0.44)	0.01 (-0.33, 0.35)	0.15 (-0.11, 0.41)	0 (-0.2, 0.24)
Not resident	-0.05 (-0.12, 0.02)	-0.09 (-0.16, -0.03)	-0.01 (-0.07, 0.06)	-0.12 (-0.19, -0.04)	0.05 (-0.05, 0.15)	-0.04 (-0.1, 0.03)
Low knowledge	-0.19 (-0.33, -0.04)	-0.06 (-0.23, 0.12)	0.12 (-0.05, 0.27)	-0.18 (-0.35, -0.01)	-0.26 (-0.48, -0.05)	0.11 (-0.03, 0.26)
Low awareness	0.16 (0.03, 0.28)	-0.25 (-0.38, -0.13)	-0.05 (-0.17, 0.07)	0.18 (0, 0.35)	0.07 (-0.12, 0.25)	0.12 (0.01, 0.25)
Democracy	0.02 (-0.18, 0.22)	-0.02 (-0.2, 0.17)	0.5 (0.34, 0.66)	0.24 (0.04, 0.44)	-0.07 (-0.33, 0.18)	-0.37 (-0.55, -0.21)
Historical coder	-0.19 (-0.72, 0.31)	1.07 (0.35, 1.74)	-0.01 (-0.4, 0.37)	-0.54 (-0.88, -0.21)	-0.05 (-0.9, 0.89)	NA
N	42,834	54,327	56,385	50,373	54,935	58,923
r^2	0.27 (0.21, 0.34)	0.34 (0.28, 0.4)	0.41 (0.35, 0.46)	0.31 (0.25, 0.37)	0.34 (0.28, 0.41)	0.26 (0.2, 0.32)

All models include country and historical date fixed effects. Quantities represent the bootstrapped coefficient median over the posterior distribution and 90 percent highest density about this estimate.

Table C.2: Bootstrapped regression results, coding correlates

	Executive oversight	Journalist harassment	Female freedom	Party autonomy	Domestic autonomy	Reasoned justification
Intercept	0.92 (0.53, 1.29)	1.14 (0.8, 1.47)	1.16 (0.79, 1.49)	1.17 (0.73, 1.58)	1.23 (0.73, 1.75)	1.25 (0.92, 1.56)
Historical coder	-0.16 (-0.33, -0.01)	0.01 (-0.17, 0.19)	0.06 (-0.08, 0.18)	0.03 (-0.14, 0.2)	-0.1 (-0.27, 0.06)	NA
Confidence	0.98 (0.77, 1.18)	0.54 (0.33, 0.75)	0.9 (0.71, 1.12)	0.82 (0.55, 1.09)	0.41 (0.11, 0.77)	0.3 (0.1, 0.5)
Confidence variation	0.19 (0.09, 0.28)	0 (-0.08, 0.09)	0.09 (0.01, 0.18)	0 (-0.09, 0.1)	0.05 (-0.06, 0.14)	-0.05 (-0.14, 0.04)
Confidence sd	-0.28 (-0.76, 0.2)	0.55 (0.06, 0.98)	0.78 (0.24, 1.27)	0.29 (-0.23, 0.82)	-0.82 (-1.58, -0.02)	-0.23 (-0.75, 0.34)
Coding variation	-0.63 (-0.75, -0.51)	-0.95 (-1.05, -0.83)	-1.14 (-1.26, -1.02)	-1.04 (-1.17, -0.91)	-0.98 (-1.15, -0.8)	-1.14 (-1.26, -1.02)
Coding sd	0.59 (0.5, 0.68)	1.2 (1.12, 1.3)	0.86 (0.77, 0.95)	0.89 (0.81, 0.98)	1.78 (1.56, 1.97)	1.52 (1.4, 1.62)
Country years	-0.05 (-0.08, -0.01)	-0.07 (-0.1, -0.04)	-0.04 (-0.08, -0.01)	-0.06 (-0.09, -0.02)	-0.06 (-0.11, -0.01)	-0.03 (-0.06, 0)
Variables	0 (-0.05, 0.05)	-0.01 (-0.05, 0.03)	0.03 (-0.02, 0.07)	0.04 (-0.01, 0.09)	0.02 (-0.02, 0.08)	0.07 (0.04, 0.11)
Countries >1	-0.09 (-0.18, 0.02)	0.04 (-0.05, 0.15)	-0.17 (-0.25, -0.08)	0.05 (-0.07, 0.17)	0.11 (-0.04, 0.24)	0.19 (0.1, 0.29)
Countries	0.06 (-0.01, 0.13)	-0.11 (-0.18, -0.04)	0.08 (0.01, 0.14)	-0.07 (-0.16, 0.01)	0.03 (-0.06, 0.13)	-0.19 (-0.25, -0.12)
PSQ complete	0.05 (0, 0.11)	-0.05 (-0.1, 0)	0.02 (-0.03, 0.08)	-0.17 (-0.23, -0.1)	-0.06 (-0.14, 0.01)	0.04 (-0.01, 0.1)
N	77,462	102,237	113,284	94,305	109,553	102,850
r^2	0.22 (0.2, 0.25)	0.34 (0.29, 0.38)	0.31 (0.27, 0.35)	0.31 (0.29, 0.34)	0.32 (0.27, 0.37)	0.28 (0.24, 0.33)

All models include country and historical date fixed effects. Quantities represent the bootstrapped coefficent median over the posterior distribution and 90 percent highest density about this estimate.

D Analysis of posterior medians

Figure D.1: Correlates of reliability