Research Article

Research and Politics
July-September 2019: 1–8

DOI: 10.1177/2053168019879561
journals.sagepub.com/home/rap

What makes experts reliable? Expert reliability and the estimation of latent traits

Kyle L. Marquardt[1], Daniel Pemstein[2], Brigitte Seim[3] and Yi-ting Wang[4]

Abstract

Experts code latent quantities for many influential political science datasets. Although scholars are aware of the importance of accounting for variation in expert reliability when aggregating such data, they have not systematically explored either the factors affecting expert reliability or the degree to which these factors influence estimates of latent concepts. Here we provide a template for examining potential correlates of expert reliability, using coder-level data for six randomly selected variables from a cross-national panel dataset. We aggregate these data with an ordinal item-response theory model that parameterizes expert reliability, and regress the resulting reliability estimates on both expert demographic characteristics and measures of their coding behavior. We find little evidence of a consistent substantial relationship between most expert characteristics and reliability, and these null results extend to potentially problematic sources of bias in estimates, such as gender. The exceptions to these results are intuitive, and provide baseline guidance for expert recruitment and retention in future expert coding projects: attentive and confident experts who have contextual knowledge tend to be more reliable. Taken as a whole, these findings reinforce arguments that item response theory models are a relatively safe method for aggregating expert-coded data.

Keywords

Cross-national panel data, expert surveys, measurement, IRT models, Bayesian methods

Introduction

Many political science datasets use experts to code concepts that are difficult to directly assess (Bakker et al., 2012; Buttice and Stone, 2012; Kitschelt and Kselman, 2012; Castles and Mair, 1984; Clinton and Lewis, 2008). Although modeling rater-level bias and reliability when aggregating codings is of clear importance (Johnson and Albert, 1999; Maestas et al., 2014; Wagner et al., 2010), there has been little exploration of either the factors that influence reliability in political science contexts or their implications for model design. Such exploration is essential for both assessing the validity of data-aggregation methods and determining criteria for expert retention and recruitment.

Here we analyze potential correlates of expert reliability in the context of a cross-national survey of political traits: the Varieties of Democracy (V–Dem) Dataset (Coppedge et al., 2018a), which employs a diverse body of over 3000 experts to code over 121 ordinal variables covering a variety of regime traits from 1900–2017.[1] This diversity of experts and contexts provides an ideal laboratory for analyzing coder reliability.

We measure reliability using expert-specific discrimination (reliability) parameters from six randomly selected

[1]School of Politics and Governance and International Center for the Study of Institutions and Development, National Research University Higher School of Economics, Russian Federation
[2]Department of Criminal Justice and Political Science, North Dakota State University, USA
[3]Department of Public Policy, University of North Carolina, Chapel Hill, USA
[4]Department of Political Science, National Cheng Kung University, Taiwan

Corresponding author:
Kyle L. Marquardt, National Research University Higher School of Economics, Krivokolennyi Pereulok 3, Moscow, 401000, Russian Federation.
Email: kmarquardt@hse.ru

V–Dem variables. In the item response theory (IRT) context, reliability parameters represent the degree to which an expert randomly diverges from other experts who code the same cases; experts who code in patterns similar to those of their peers receive higher reliability scores and thus contribute more to the estimation of the latent concept. This operationalization aligns with classic definitions of reliability (Carmines and Zeller, 1979), as well as work examining convergence among crowd-platform coders (Benoit et al., 2016).

In this analysis, we regress these reliability parameters on both expert coding behavior and demographic characteristics. Doing so provides insight into the degree to which the prominent method for aggregating expert-coded data—an IRT model that accounts for both variation in expert reliability and scale perception (Clinton and Lewis, 2008; Pemstein et al., 2019)—provides substantively unbiased estimates of latent concepts.

In general, we find a weak and inconsistent relationship between reliability and expert characteristics. Most of these null findings regard variables that could constitute problematic sources of bias in the the estimation procedure, such as gender. The exceptions are intuitive. Reliable experts tend to be those who: (a) are more confident in their codings; (b) vary their codings; and (c) evince contextual knowledge of an important concept. Cumulatively, these findings indicate that IRT models incorporating expert reliability and scale perception parameters are a safe method for aggregation.

Reliability in the V–Dem model

We use a modified version of the V–Dem measurement model (Pemstein et al., 2019) to estimate expert reliability for each of the six variables.[2] This model derives from the basic assumption that each expert r provides a coding of the latent trait z in country-year ct with error, such that

$$\tilde{y}_{ctr} = z_{ct} + e_{ctr}, \quad e_{ctr} \sim N(0, \sigma_r) \tag{1}$$

where $\tilde{y}_{ctr}$ is the expert's perception of the latent trait, e_{ctr} the perceptual error in each observation, and σ_r the rater-specific error variance. The model resembles a standard Bayesian ordinal IRT model (Johnson and Albert, 1999), and recovers latent trait estimates as well as or better than other standard methods for aggregating expert scores (Marquardt and Pemstein, 2018). Equation 2 shows the partial likelihood.

$$\Pr(y_{ctr} = k) = \phi\left(\gamma_{r,k} - \beta_r z_{ct}\right) - \phi\left(\gamma_{r,k-1} - \beta_r z_{ct}\right) \tag{2}$$

ϕ is the normal cumulative distribution function and y_{ctr} is the ordinal survey response of expert r for country-year ct. We assume $z_{ct} \sim N(0,1)$.

Two sets of parameters in this model are of particular importance. First, γ is a k-length vector of threshold parameters for each expert r. These parameters model variation in expert strictness, accounting for non-linearity across the response scale.[3] In doing so, they allow the model to account for systematic biases in how experts translate perceptions into ratings (differential item functioning, or DIF), a common concern in surveys involving multi-rater judgment (Aldrich and McKelvey, 1977; Bakker et al., 2014; Hare et al., 2015).

Second, β is a vector of expert-specific reliability parameters. We model $\beta_r \sim N(1,1)$, restricted to positive values.

In IRT terminology, β is a "discrimination" parameter: $\beta_r = \frac{1}{\sigma_r}$. Because σ_r is rater r's error variance, β measures reliability. After accounting for DIF through γ, experts with higher β scores are those who stochastically diverge less from other experts who code the same cases.[4]

Benefits of analyzing reliability correlates

Analyses of potential reliability correlates provide a diagnostic of a measurement model. In the model we use, experts with lower reliability scores contribute less to the estimation of country-year latent traits, the parameters of interest in most applications. Systematic biases that are inconsistent with model assumptions—notably case-varying systematic differences across experts—will appear to the model like random error, resulting in lower reliability scores among experts who exhibit such biases. Although certain coder characteristics—such as conceptual knowledge—should correlate with reliability, other traits should not. Analyses of respondent-level reliability can therefore provide insight into potential threats to validity by highlighting classes of experts for which selection procedures and modeling assumptions do not effectively adjust for systematic bias.

A key example is gender. A majority of V–Dem experts are men. If women systematically perceive a latent trait differently than men, and this systematic bias is not adequately modeled through threshold estimates, women could receive lower reliability scores even though their viewpoint is equally valid. Such a result would indicate problematic bias in the measurement process.

Analyses of reliability correlates also provide tentative evidence regarding the characteristics of more reliable experts, which may facilitate decisions on expert recruitment and retention. Although research stresses that expertise is important for data validity (Maestas et al., 2014), potential correlates of intra-expert variation in this context remain largely unexplored.

Variables and descriptive statistics

Reliability

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

We randomly selected all six variables.[5] We selected one variable (*Female freedom* of discussion [**v2cldiscw**]) from the set of gender-specific variables, because these are the most likely cases for gender-based systematic differences in reliability. We selected the remaining five from the set of all Likert-scale expert-coded variables: (a) *Executive oversight* by non-legislature bodies (**v2lgotovst**), (b) opposition *Party autonomy* (**v2psoppaut**), (c) the degree to which officials offer *Reasoned justification* (**v2dlreason**) for their decisions, (d) government *Domestic autonomy* (**v2svdomaut**) from other states, and (e) the level of *Journalist harassment* (**v2meharjrn**).[6]

We use Markov chain Monte Carlo (MCMC) methods to estimate the IRT model for each of the variables included in the analysis.[7] 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.

Correlates of reliability

We discuss potential sets of reliability correlates in turn. All variables related to coding characteristics regard the variable being analyzed; self-reported confidence and coding variation variables use reduced data.[8] Online Appendix C presents descriptive statistics.

Demographics. Previous research illustrates that a rater's background can influence their perception of latent traits (Cumming, 1990; Michael et al., 1980; Royal-Dawson and Baird, 2009), and raters with greater expertise are more reliable when rating complex or broad tasks (Schoonen et al., 1997). We therefore include measures of education and university employment, which indicate relevant expertise and thus potentially greater reliability.

We trichotomize education: experts with a (a) *PhD* (reference level), (b) *Professional degree* such as a Master of Business Administration or Doctor of Jurisprudence, or (c) *MA or lower*. We analyze employment with four indicators: employees of a *Public university* (the reference level), *Private university*, the *Government*, and *Other* (non-governmental, non-academic employment). We separate public and private employment because experts in the private sector may be more reliable, because they are potentially less susceptible to government pressure or other incentives to provide biased estimates.

Because gender may influence reliability for reasons previously discussed, we include the dichotomous indicator *Female*. We also include the natural logarithm of a respondent's *Age* . Finally, we include an indicator for *Historical coders*, or those coders who coded data for a select set of cases back to 1789. These coders are generally the sole coder for pre-1900 data, which could mechanically affect their reliability.

Knowledge. We *a priori* expect all experts to have a high level of knowledge about the cases and concepts they code. Equally knowledgeable experts should provide similar coding patterns, although their codings may vary due to DIF or case-level stochastic error. However, if some experts know less about a concept or case, their codings may vary in a fashion that is not attributable to DIF or case-level stochastic error. For example, a less knowledgeable expert may miss changes in latent concept values. As a result, less knowledgeable experts should receive lower reliability scores.

Measuring knowledge is difficult in the absence of concrete data (e.g., responses to factual questions about a case). We therefore use three proxies to measure different types of knowledge. Because these proxies are not comprehensive, results should be interpreted with caution.

We proxy lower case knowledge with an indicator for experts who are *Not resident* in the country they are coding, assuming that residing in a country can provide an expert experience with a case. 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—a principle that underpins most definitions of democracy—important to the broader concept of democracy. The indicator *Low knowledge* represents experts who either consider (a) very democratic Sweden to be non-democratic or (b) very non-democratic North Korea to be democratic.[9]

Democracy in residence country. Experts living in democratic countries may have better access to information and may thus be more knowledgeable than experts residing in autocracies. They may also be less concerned by potential government sanction, allowing them to more accurately code sensitive concepts and cases. For both of these reasons, such experts may be more reliable. *Democracy* represents the average level of V–Dem's electoral democracy index from 2008 to 2017 for an expert's residence country.

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. This measure provides a rough

estimate of an expert's knowledge about the variable they are coding; experts who are generally not confident are potentially signaling low knowledge.[10] For the reasons detailed in the previous section, lower knowledge could result in lower reliability.

Attentiveness. Less attentive experts may be less reliable, because they will be less sensitive to changes in latent traits for the variables they code than more attentive experts. We measure attentiveness with two sets of indicators. First, because most countries vary in political traits, the degree to which an expert varies their scores may proxy their attentiveness. Second, because expertise likely varies over time and across countries, attentive experts should vary in self-reported confidence. We measure both variation in coding and confidence with two indicators each. *Coding variation* and *Confidence variation* indicate if an expert changed their scores on either metric at least once. Because the extent to which an expert varied their coding or confidence may also be important for reliability, we also include *Coding sd* and *Confidence sd* to measure an expert's standard deviation on these metrics, with those who did not vary coded as zero.

Volume. High coding volume may lead experts to overextend themselves, causing them to either be less attentive or code cases and concepts with which they are less familiar. Such overextended experts may therefore be less reliable. 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, although 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 they coded. We include both measures in the model because the difference between coding one and two countries may be different from the difference between coding any number of countries after two.

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.[11] 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.[12]

Figure 1 presents coefficient estimates by variable, with points representing the bootstrapped median coefficient estimate and horizontal lines the 90% highest bootstrapped density about this estimate. The vertical line aligns with an effect magnitude of zero; we center the intercept at zero for illustration purposes.

Demographics. The difference between female and male coders is generally low in magnitude and inconsistent across variables, indicating the model does not erroneously penalize female experts. Age and employment also show 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, although these estimates are based on a relatively small number of experts; results regarding experts with a Master's degree or lower level of education are ambiguous. 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), although this result may be a relic of differences in the cases these experts code.

Democracy in residence country. Democracy shows an ambiguous relationship with reliability, evincing little relationship in four of the six variables and contradicting signs in the remaining two.

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

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.

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.

Volume. Neither the number of country-years an expert coded nor the number of variables they coded shows a relationship with reliability in any variable. Results regarding the number of unique countries an expert coded are inconsistent; volume and reliability are uncorrelated for two variables, negatively correlated for three variables, and positively correlated for one variable.

Predicted reliability

The coefficient plots also show a high level of uncertainty in the intercept, which indicates they may be misleading in

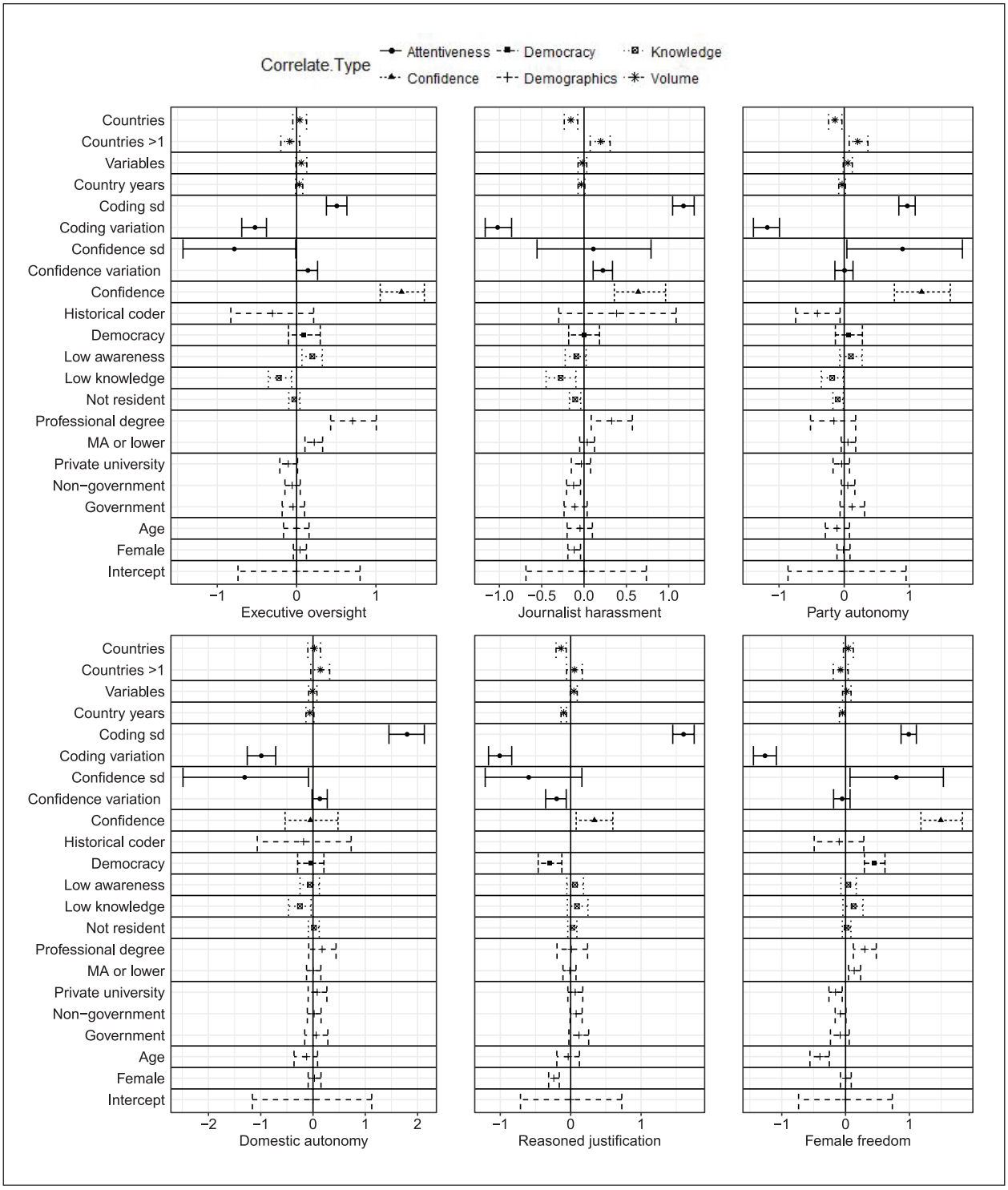

Figure 1. Bootstrapped posterior coefficient estimates of correlates of reliability.
Intercept estimate centered at zero for illustration purposes. Models include country and year fixed effects.

terms of the substantive importance of reliability correlates. Figure 2 presents the predicted reliability of experts with different characteristics across variables. Points represent the bootstrapped predicted median reliability for experts with given certain demographic or coding characteristics, holding all other correlates constant at their mean or mode.[13] The range represents each variable's posterior median range of reliability scores.

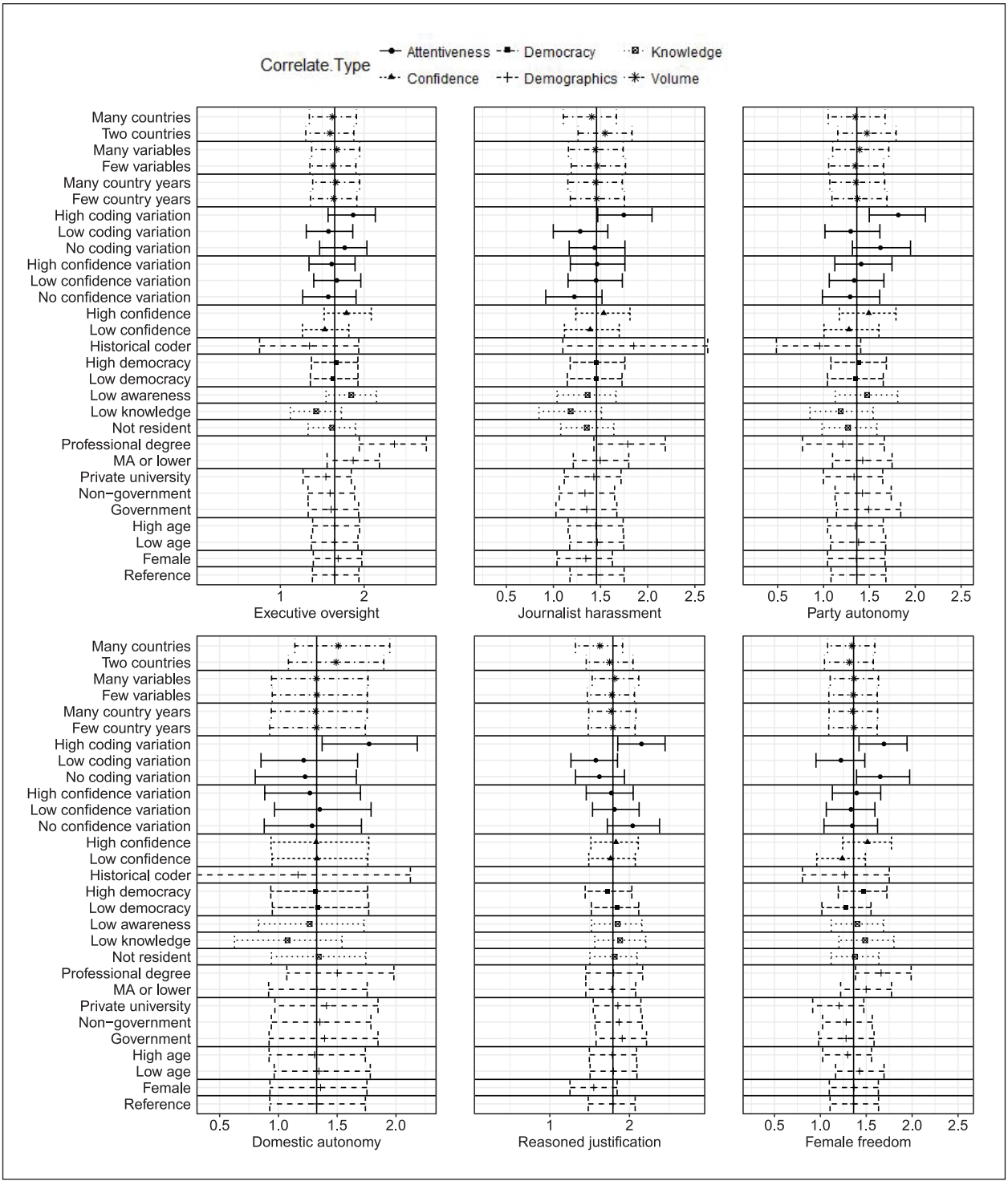

Figure 2. Posterior bootstrapped predicted reliability of experts with different characteristics.
Models include country and year fixed effects.

As Figure 2 makes clear, once we incorporate overall posterior uncertainty into the assessment, the substantive relationship between the correlates of reliability and this outcome is generally minimal. The main exceptions to this rule are Confidence, Low knowledge, and Coding variation, which retain their relatively strong correlation with reliability. In four of the six indicators, experts with high average confidence are more reliable than those with lower confidence, and experts with low knowledge tend to be less reliable. Across all variables, experts who vary their coding are more reliable than those who do not or do so minimally.

Conclusion

The analyses in this paper assess the correlates of expert reliability in the context of cross-national panel data. Most potential correlates show little substantive relationship with reliability; these null results provide evidence that the IRT model is well specified in this context and, more generally, that IRT models are a safe method for aggregating expert-coded data.

The most notable exception to this rule regards coding variation, which is positively correlated with reliability. This result provides a simple heuristic for evaluating respondents to expert surveys: of those experts who vary their codings, those that vary the most will tend to be most reliable. Other results are more tentative, albeit intuitive: lower conceptual knowledge and lower confidence predict lower reliability. This suggests that expert-coding enterprises should endeavor to recruit experts who have knowledge of the concepts they are coding and are confident in their knowledge.

This paper also suggests directions for further research. Although the analyses here focus on expert-level correlates of reliability, they provide tentative evidence that task difficulty also matters: the country and year being coded explains a great deal of variation in reliability (Online Appendix Table E.1), and the distribution of reliability scores varies substantially across questions (Online Appendix Figure C.1). Although it is important to not overinterpret these results, future scholarship would do well to probe them.

Acknowledgements

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 the editor and two anonymous reviewers for their valuable insights. The authors also thank Staffan Lindberg and other members of the V--Dem team for their suggestions and assistance. Regionala etikprövningsnämnden i Göteborg 1080-16 provided ethics approval, including informed consent guidelines.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors acknowledge research support from 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. Marquardt acknowledges the support of the HSE University Basic Research Program and funding by the Russian Academic Excellence Project '5-100.' The authors performed simulations and other computational tasks using resources provided by the Swedish National Infrastructure for Computing at the National Supercomputer Centre in Sweden (SNIC 2017/1-406 and 2018/3-133, PI: Staffan I. Lindberg).

ORCID iD

Kyle L. Marquardt https://orcid.org/0000-0003-2043-5880

Supplemental materials

The supplemental files are available at http://journals.sagepub.com/doi/suppl/10.1177/2053168019879561
The replication files are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LSLVLT

Notes

1. For details regarding the recruitment of V–Dem experts, see Coppedge et al. (2018b).
2. We do not pool variables in the estimation process because scales and reliability may vary substantially.
3. Each threshold γ_k for expert r is hierarchically clustered: first by thresholds for experts who were recruited to code the same main country, then by a universal threshold. The first cluster rests on the assumption that experts with experience in similar countries will have similar patterns of DIF, whereas the second facilitates identification.
4. See Online Appendix A and Pemstein et al. (2019) for additional details. Note that reliability is not necessarily the same as accuracy (Maestas et al., 2014); assessing accuracy directly is an impossible task in this dataset, given there is no concrete reference point for coding accuracy of latent variables.
5. To increase the applicability of our results to other expert-coded datasets (Bakker et al., 2012), we also analyze two additional variables that measure concepts related to government ideology. Online Appendix F presents the results from these analyses, which largely align with those in the article.
6. For additional details on the variables, see the V–Dem Codebook. All variables are based on five-point Likert scale questions, with the exceptions of Domestic autonomy and Reasoned justification, which have three and four points, respectively.
7. We conduct all analyses using the statistical software Stan (Stan Development Team, 2018). See Online Appendix B for additional details.
8. We reduce data to regimes—country-year observations where at least one expert changes their coding or self-reported confidence—in the estimation process to prevent inaccurate estimates of uncertainty (Pemstein et al., 2019).

9. Experts rank 12 countries on a 0–100 scale, with high scores representing more democracy; a score on either side of 50 indicates democracy versus non-democracy.
10. Many experts provide a static value of one for their confidence, and average confidence may vary due to factors unrelated to the underlying construct. We account for the first concern with the indicator *Confidence variation*. As regards the second concern, gender is perhaps the most likely confounding variable; Online Appendix D explores this possibility.
11. Online Appendix H presents results from analyses that only analyze the relationship between the correlates and the posterior median, whereas Online Appendix G provides analyses that (a) only include coding characteristics and (b) only include expert characteristics from the post-survey questionnaire.
12. 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 0.22 (Online Appendix E).
13. 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; variation in coding and confidence), we report the relationship between the dichotomous indicator of variation (*Two countries coded, No coding*, and *No confidence*) and $High$ and Low estimates based on the quantiles of experts who show variation (the exception is unique countries, because most experts who coded more than one country coded only two).

Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

References

Aldrich JH and McKelvey RD (1977) A method of scaling with applications to the 1968 and 1972 presidential elections. *American Political Science Review* 71(1): 111–130.

Bakker R, de Vries C, Edwards E, et al. (2012) Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics* 21(1): 143–152.

Bakker R, Jolly S, Polk J, et al. (2014) The European common space: Extending the use of anchoring vignettes. *The Journal of Politics* 76(4): 1089–1101.

Benoit K, Conway D, Lauderdale BE, et al. (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review* 110(2): 278–295.

Buttice MK and Stone WJ (2012) Candidates matter: Policy and quality differences in congressional elections. *Journal of Politics* 74(3): 870–887.

Carmines EG and Zeller RA (1979) *Reliability and validity assessment*. Thousand Oaks, CA: Sage Publications.

Castles FG and Mair P (1984) Left-right political scales: Some 'expert' judgments. *European Journal of Political Research* 12(1): 73–88.

Clinton JD and Lewis DE (2008) Expert opinion, agency characteristics, and agency preferences. *Political Analysis* 16(1): 3–20.

Coppedge M, Gerring J, Knutsen CH, et al. (2018a) V–Dem Dataset v8. *Technical report, Varieties of Democracy Project*. https://ssrn.com/abstract=3172819.

Coppedge M, Gerring J, Lindberg SI, et al. (2018b) Varieties of Democracy Methodology v8. *Technical report, Varieties of Democracy Project: Project Documentation Paper Series*. https://ssrn.com/abstract=3172796.

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

Hare C, Armstrong DA, Bakker R, et al. (2015) Using Bayesian Aldrich-Mckelvey scaling to study citizens' ideological preferences and perceptions. *American Journal of Political Science* 59(3): 759–774.

Johnson VE and Albert JH (1999) *Ordinal Data Modeling*. New York: Springer.

Kitschelt H and Kselman DM (2012) Economic development, democratic experience, and political parties' linkage strategies. *Comparative Political Studies* 46: 1453–1484.

Maestas CD, Buttice MK and Stone WJ (2014) Extracting wisdom from experts and small crowds: Strategies for improving informant-based measures of political concepts. *Political Analysis* 22(3): 354–373.

Marquardt KL and Pemstein D (2018) IRT models for expert-coded panel data. *Political Analysis* 26(4): 431–456.

Michael WB, Cooper T, Shaffer P, et al. (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 D, Marquardt KL, Tzelgov E, et al. (2019) The V-Dem measurement model: Latent variable analysis for cross-national and cross-temporal expert-coded data. *Varieties of Democracy Institute Working Paper* 21: 1–30.

Royal-Dawson L and Baird JA (2009) Is teaching experience necessary for reliable scoring of extended english questions? *Educational Measurement: Issues and Practice* 28(2): 2–8.

Schoonen R, Vergeer M and Eiting M (1997) The assessment of writing ability: expert readers versus lay readers. *Language Testing* 14(2): 157–184.

Stan Development Team (2018) RStan: the R interface to Stan, R package version 2.17.3. http://mc-stan.org.

Wagner SM, Rau C and Lindemann E (2010) Multiple informant methodology: A critical review and recommendations. *Sociological Methods and Research* 38(4): 582–618.