SUPPORTING INFORMATION

Reconciling the Theoretical and Empirical Study of International Norms: A New Approach to Measurement

American Political Science Review

Tyler Girard

Department of Political Science

The University of Western Ontario

tgirard2@uwo.ca

Contents

1 Item Response Theory (IRT) Primer

1.1 Applications in Political Science

Item Response Theory (IRT) models were initially developed in psychological and educational research (Armstrong II et al., 2014; Patz and Junker, 1999; Baker, 2001), but are now used across a range of substantive domains in political science. Notwithstanding the different issue areas, the form of the data is consistent: we have units (countries, individuals, etc.) and indicators associated with those units (votes, survey responses, treaty ratifications, etc.) that are understood to be the observed manifestation of an underlying concept (ideology, state preferences, etc.). To use a running example from survey research (Treier and Hillygus, 2009), an individual's responses to a set of survey questions (such as support or opposition to policies on abortion, same-sex marriage, and gun control) can be used to estimate that individual's political ideology (an unobserved continuum from liberal to conservative). Importantly, while the level of measurement of the indicators may vary across applications (binary, ordinal, nominal, continuous, or a mixture), our estimation strategies and intepretation of results often remain similar.

The following examples help to further clarify not only how these models are applied and interpreted, but also some of the salient issues for the proposed application to measuring norm adoption. Martin and Quinn (2002) estimate the ideology of the U.S. Supreme Court Justices from 1953-1999. Here, political ideology is the "latent dimension" that we measure by considering the voting patterns of Justices. In turn, we can then situtate each Justice on the underlying political ideology spectrum. Similarly, Clinton, Jackman and Rivers (2004) use U.S. Congressional roll call data (the indicators) to estimate legislator ideal points (the latent dimension) and Bailey, Strezhnev and Voeten (2017) use votes in the United Nations General Assembly (the indicators) to estimate country ideal points (the latent dimenion). Several scholars have developed IRT models to estimate the ideological position of political parties or individual legislators from text, where words or sentences operate as indicators in the same manner as survey questions or votes (Slapin

and Proksch, 2008; Benoit et al., 2016). As argued in the main manuscript, the measurement of norm adoption fits neatly within this framework, such that country policies and laws serve as indicators of the degree of norm adoption (the latent dimension) in each country-year.

In applications involving time-series cross-sectional data, an important element to consider is *time*. When using IRT models to measure democracy (Treier and Jackman, 2008; Pemstein, Meserve and Melton, 2010) and human rights abuses (Fariss, 2014; Schnakenberg and Fariss, 2014), for instance, we generally expect the position of a country on the latent dimension in a given year to be related to the position of the same country on the latent dimension in the previous year. As Reuning, Kenwick and Fariss (2019) argue, "static" approaches that treat each year as independent ignore the time-series properties of the data and the unobserved concept. Alternatively, "dynamic" IRT models (that smooth the latent estimates over time) provide an improved theoretical connection while also potentially "over-smoothing" rapid changes in a country's position on the latent dimension.

When compared to the initial development of IRT models in other disciplines, applications in political science tend to emphasize different features of the model. As explained by Armstrong II et al. (2014, 222): "[W]hile the focus in testing applications of the IRT model is on the estimated values of the item parameters (to determine how well test items are constructed), political scientists' quantity of interest is usually the individual parameters [i.e. the position of each subject on the latent dimension]" (see also Clinton, Jackman and Rivers 2004). To be clear, the model specification is the same, what changes is the focus of our attention and how we interpret the different components of the model. In the application presented here, however, we gain insight on different elements of norm adoption by considering all components of the model. Rather than focus solely on the latent estimates (degree of norm adoption), the item/indicator parameters can shed light on the relationship between observed policies/laws and the degree of norm adoption over time.

1.2 Understanding the Model

To provide additional technical details on the model specification, I return to the example of estimating an individual's political ideology from a set of survey questions. We assume that all individuals have a latent political ideology along a single dimension, which we infer from the modeled relationship between the latent political ideology and the responses to a battery of questions (or "items," in the psychology/education terminology). For ease of understanding, given my focus on the adoption (or non-adoption) of policies/laws associated with international norms, we'll assume that the questions are structured to solicit binary responses, where "support" is the conservative response.

For each question, the individual is predicted to respond with "support" if his or her ideological position is above a particular threshold. This threshold is referred to as the question difficulty parameter. An item characteristic curve (ICC), also known as an item response function, is estimated for each question, which maps the probability of a "support" response given specific levels of the latent ability and the value of the difficulty parameter. The ICC monotonically increases over the latent dimension. In other words, as you move from left to right on the latent dimension (towards higher levels of conservatism), the probability of the individual providing a "support" (i.e. conservative) response to the question increases. Different versions of this model are achieved by allowing the shape of the ICC (captured by the discrimination parameter) to vary for each question.

It is important to note that we cannot estimate a latent concept like ideology using a single question (Fariss, 2018). In this scenario, we are unable to distinguish between variation in the observed response due to measurement error versus variation in the latent political ideology. However, by adding additional questions, we can produce a more precise estimate of the individual's political ideology. Further, we assume that the responses to any two questions are independent conditional on the individual's latent political ideology. In other words, we assume that two question responses are only related because each is an observed outcome of the latent political ideology. Likewise, the adoption of

policies or laws by a country is related through the country's underlying degree of norm adoption (the latent dimension). Consequently, we do not expect policies and laws to be unrelated; rather, they are related to each other through the country's degree of norm adoption.

Using the running political ideology example, our data would consist of $i = 1, ..., N$ individuals and $j = 1, ... J$ questions. The probability distribution for the "support" (conservative) response by individual i on question j, where $F(\cdot)$ identifies the logistic cumulative distribution function, is:

$$P[y_{ij} = 1] = F(\alpha_j + \beta_j \theta_i) \tag{1}$$

The parameter α represents the question *difficulty*, the parameter β represents the question *discrimination*, and the parameter θ represents an individual's unobserved *political ideology*. The likelihood function is thus:

$$L(\alpha, \beta, \theta | y) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[F(\alpha_j + \beta_j \theta_i)^{y_{ij}} * (1 - F(\alpha_j + \beta_j \theta_i))^{(1-y_{ij})} \right] \tag{2}$$

However, the running example has only considered a group of individuals and set of questions at a single time point. As discussed above, time-series cross-sectional data requires us to consider the temporal features of the concept and data. To do so, we can index the units of time (e.g. years) as $t = 1, ..., T$, thus adjusting the probability distribution:

$$P[y_{itj} = 1] = F(\alpha_j + \beta_j \theta_{it}) \tag{3}$$

The likelihood function is:

$$L(\alpha, \beta, \theta | y) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=1}^{J} \left[F(\alpha_j + \beta_j \theta_{it})^{y_{itj}} * (1 - F(\alpha_j + \beta_j \theta_{it}))^{(1-y_{itj})} \right] \tag{4}$$

The use of Bayesian, rather than maximum likelihood, methods allows us to simultane-

ously estimate the difficulty, discrimination, and latent dimension parameters. The Gibbs sampler allows for efficient sampling of the conditional posterior densities of each parameter from a high-dimensional posterior density (Armstrong II et al., 2014). Further, while all IRT models face problems of identification[1], Bayesian methods are easily capable of handling this problem through parameter constraints and theoretically informed priors. The use of priors also aids with "dynamic" models (with latent concepts estimated over time). By using a random walk prior on the latent concept, we can set our prior about the value of the latent variable at time t by the value at time t-1.

[1]Specifically, invariance to reflection (multiplying all parameters by -1 would not affect the likelihood function) and invariance to rotation (different parameter sets suggest the same probability distribution, given the data).

2 Model Data and Model Fit

As described in the main manuscript (p. 7), the data consist of 13 policies/laws compiled using reports from the International Lesbian, Gay, Bisexual, Trans and Intersex Association (ILGA). The information from these reports was combined with the Correlates of War (COW) state system membership data (manually extended to include 2017; Correlates of War 2017). Further, I use data from the Archigos project (Goemans, Gleditsch and Chiozza, 2009) to identify regime changes for each country. In order to both extend the Archigos data to cover the full time period[2] and include all COW states, I relied on the original sources used for the Archigos project.[3] The resulting data set covers 196 countries between 1990-2017.

To evaluate model fit, we often use tools like information criteria (e.g. Akaike Information Criterion (AIC)) to compare the fit of competing models. This approach is not helpful in this application since the goal is not to adjudicate between rival models. Instead, posterior predictive checks provide an opportunity to identify potential problems with model fit. The essential idea is that data simulated from the parameters of a well-fitting model should closely resemble the original data (Gelman and Hill, 2007; Gelman et al., 2013). Using the posterior samples, 50,000 data sets were simulated in each year (1990-2017). Subsequently, I compare the proportion of positive responses (indicating the policy/law was adopted) for each policy/law in each year between the original data and the simulated data.[4] Comparing the results graphically allows us to easily identify systematic discrepancies between the simulated data and the original data, which can be due to model misfit or chance. The results of the posterior predictive checks for four years (1990, 2000, 2010, 2017) are presented below (Figure 1 to Figure 4) and strongly suggest that the model fits the data well. The posterior predictive checks for every year (1990-2017) are available in the Methodological Appendix.

[2]The original Archigos data ends in 2015.
[3]Specifically, www.rulers.org.
[4]For years in which a policy/law had not yet been adopted by any country (e.g. same-sex marriage, 1990-2000), no parameter estimates are generated. Consequently, no posterior predictive check is performed for such policies/laws.

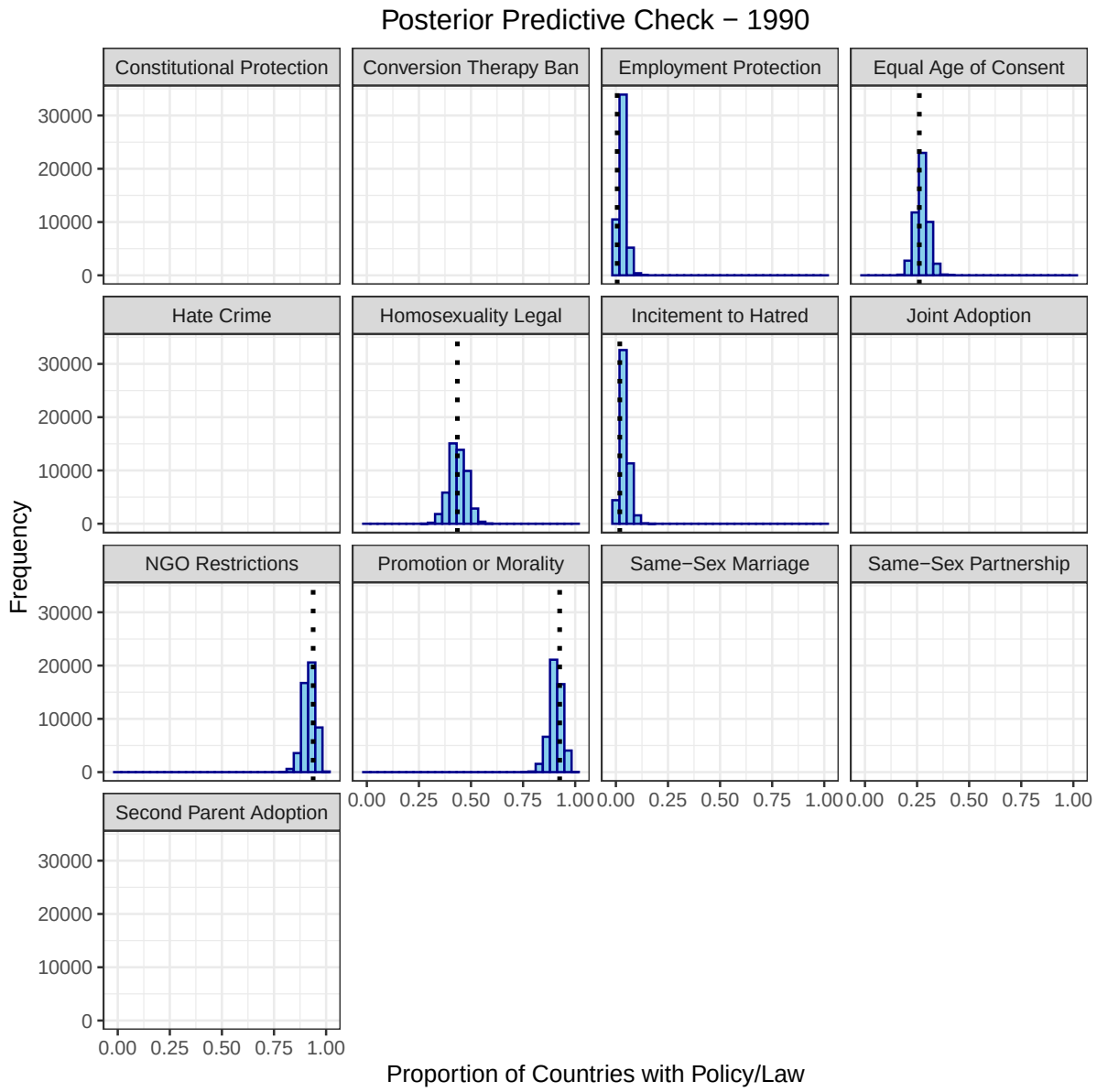

Figure 1: The dotted line indicates the proportion of countries with that policy/law in a given year in the original data. The bars indicate the frequency of proportions from 50,000 simulated data sets using the posterior estimates.

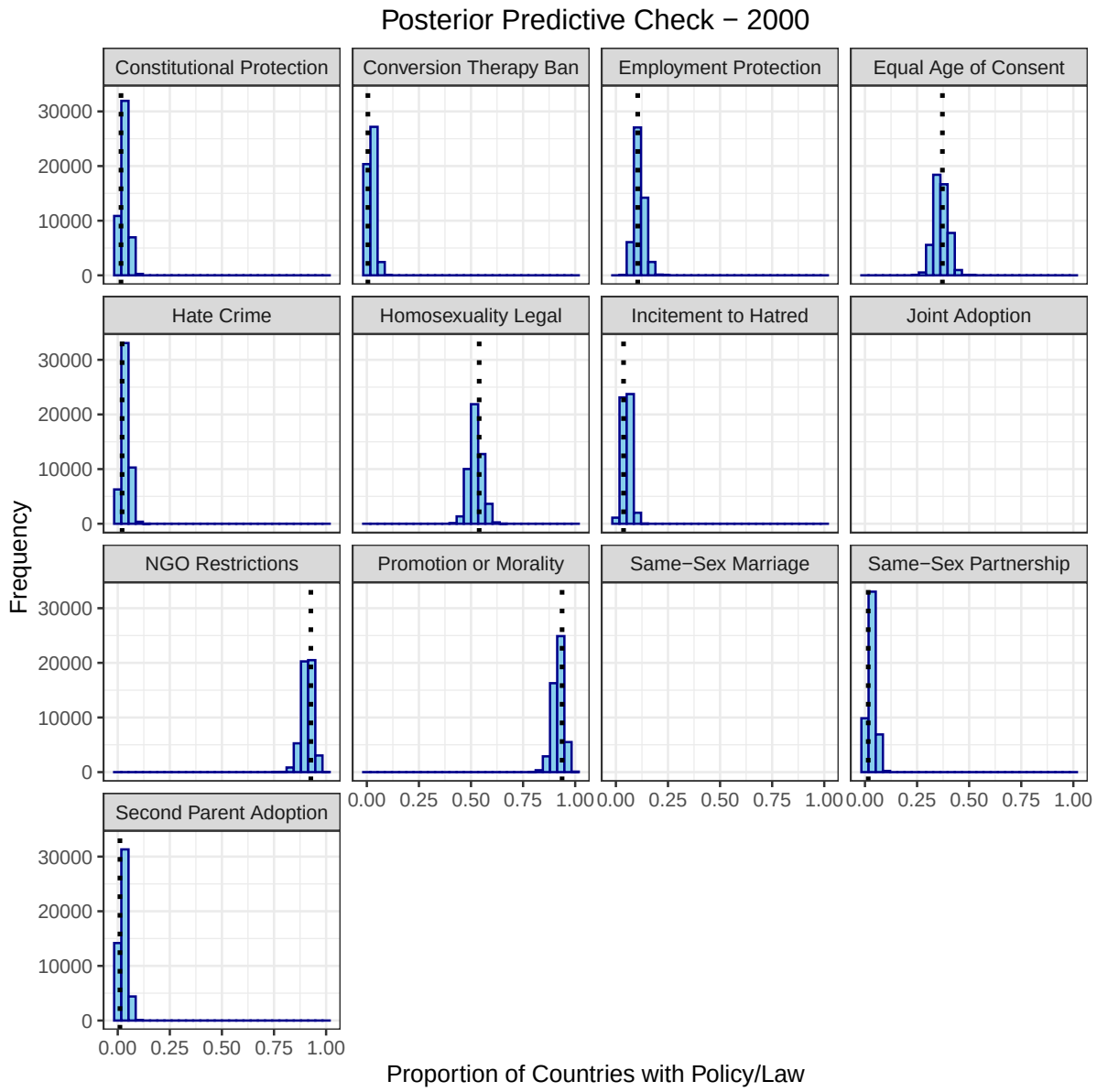

Figure 2: The dotted line indicates the proportion of countries with that policy/law in a given year in the original data. The bars indicate the frequency of proportions from 50,000 simulated data sets using the posterior estimates.

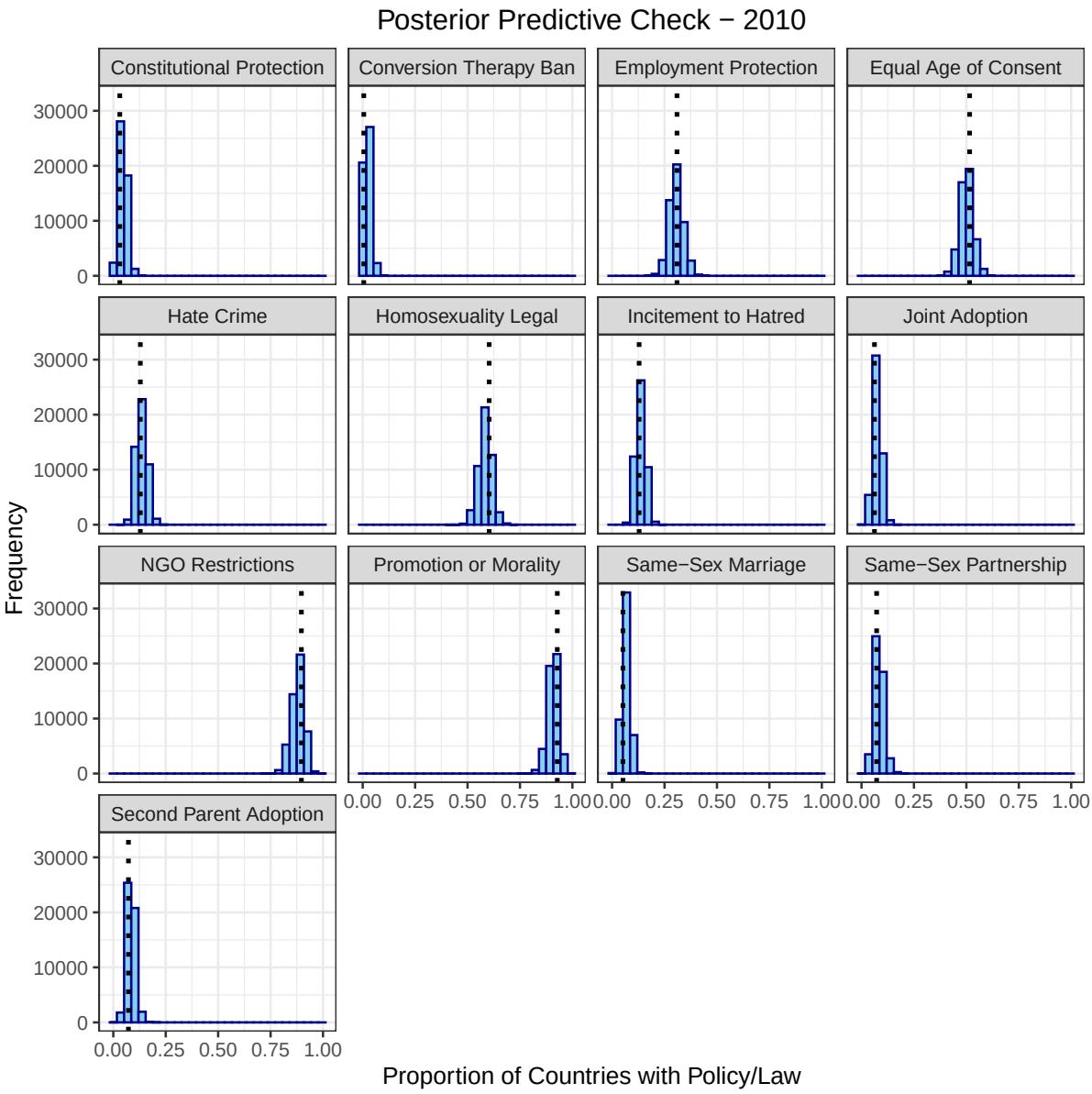

Figure 3: The dotted line indicates the proportion of countries with that policy/law in a given year in the original data. The bars indicate the frequency of proportions from 50,000 simulated data sets using the posterior estimates.

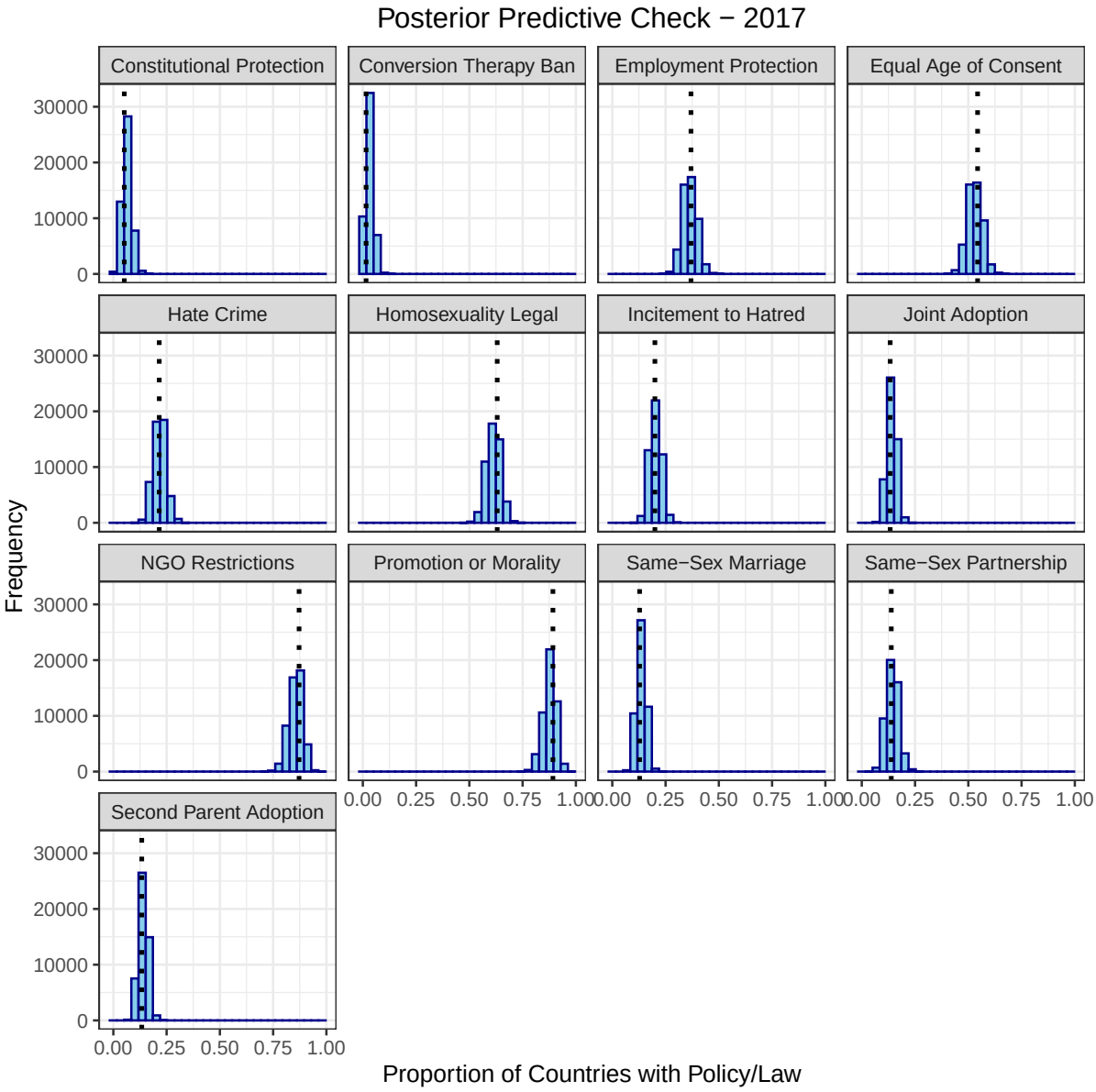

Figure 4: The dotted line indicates the proportion of countries with that policy/law in a given year in the original data. The bars indicate the frequency of proportions from 50,000 simulated data sets using the posterior estimates.

3 Results - Latent Estimates (Norm Adoption)

3.1 Global Adoption, All Years

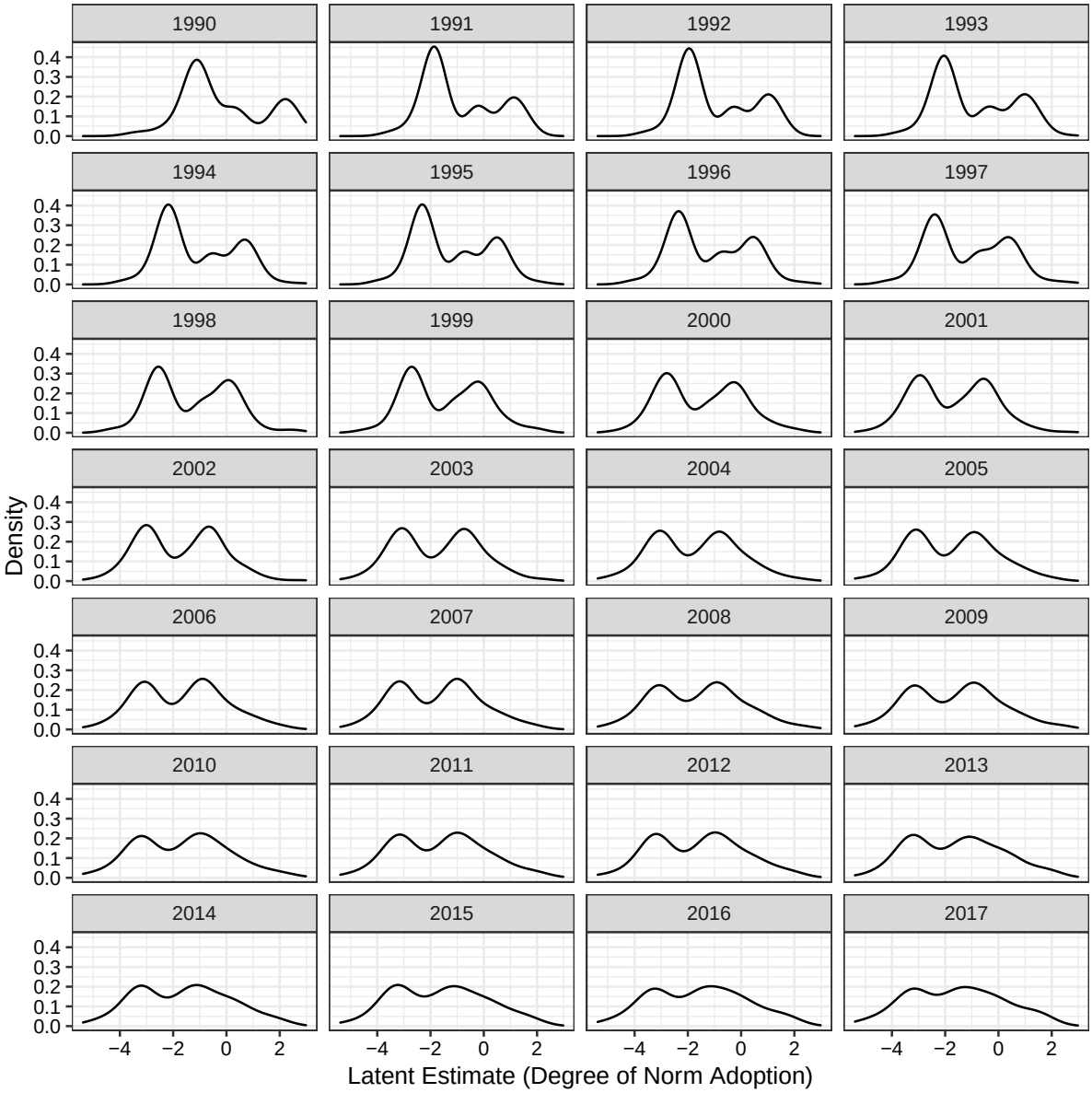

Figure 5: Distribution of latent estimate (norm adoption) median posterior sample for all countries in a given year.

3.2 Global Adoption, 2001/2009/2017 (World Maps)

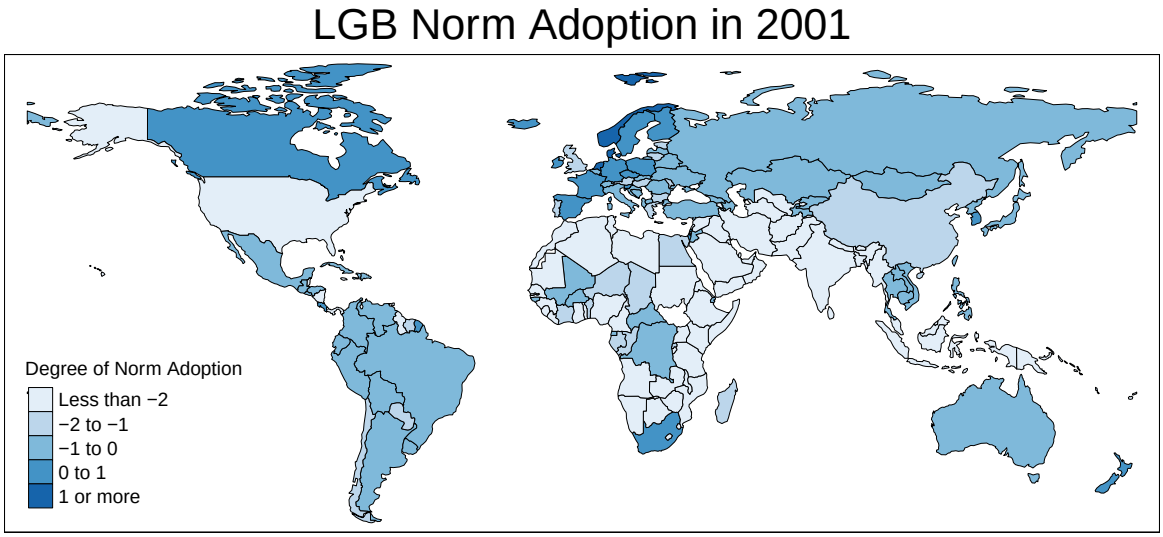

Figure 6: Median of the posterior distribution of latent estimates (norm adoption).

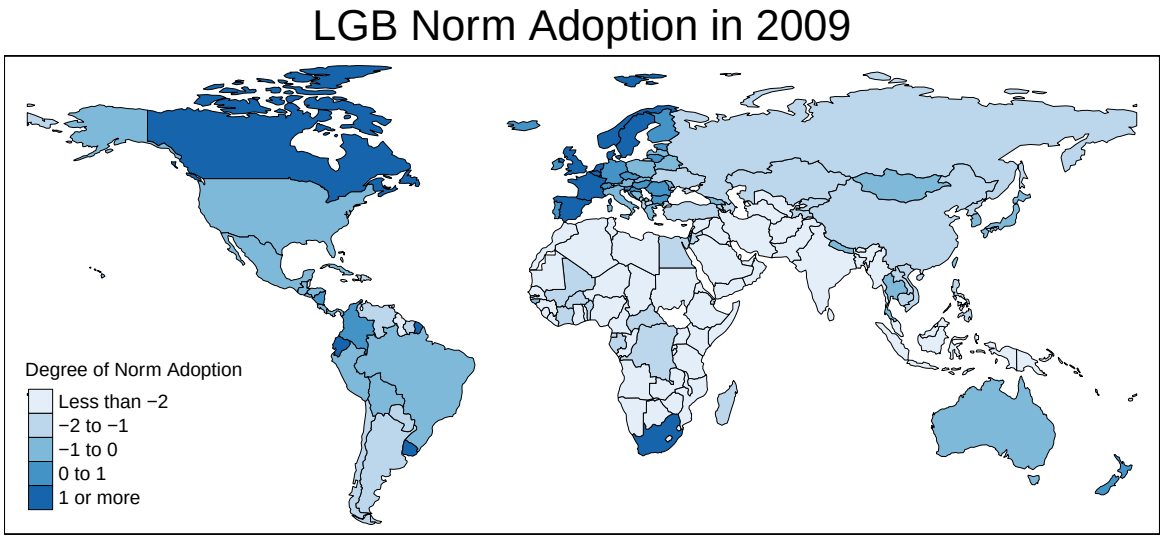

Figure 7: Median of the posterior distribution of latent estimates (norm adoption).

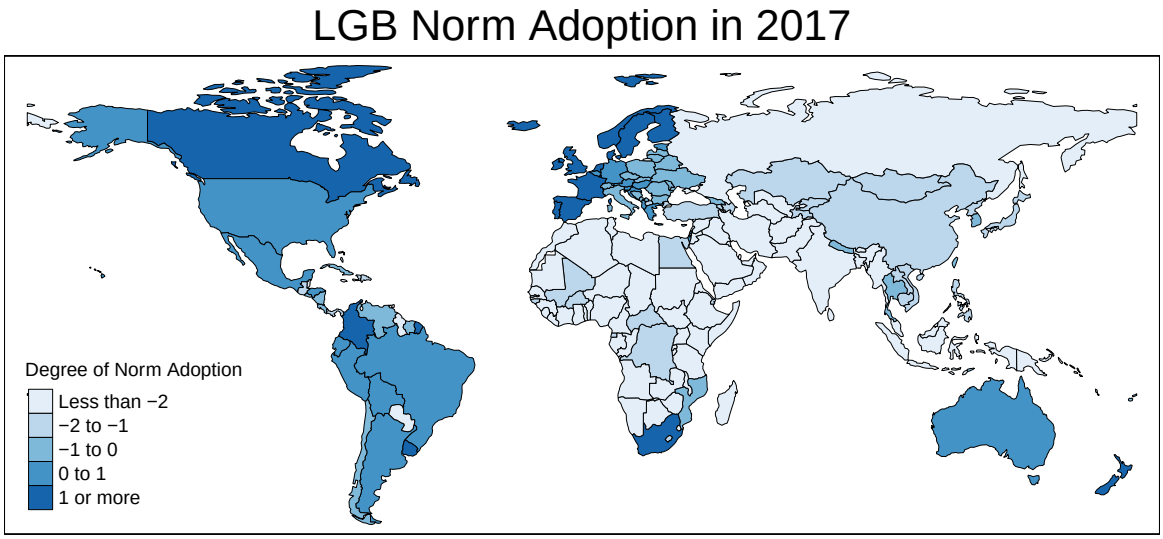

Figure 8: Median of the posterior distribution of latent estimates (norm adoption).

4 Results - Difficulty Parameter Estimates

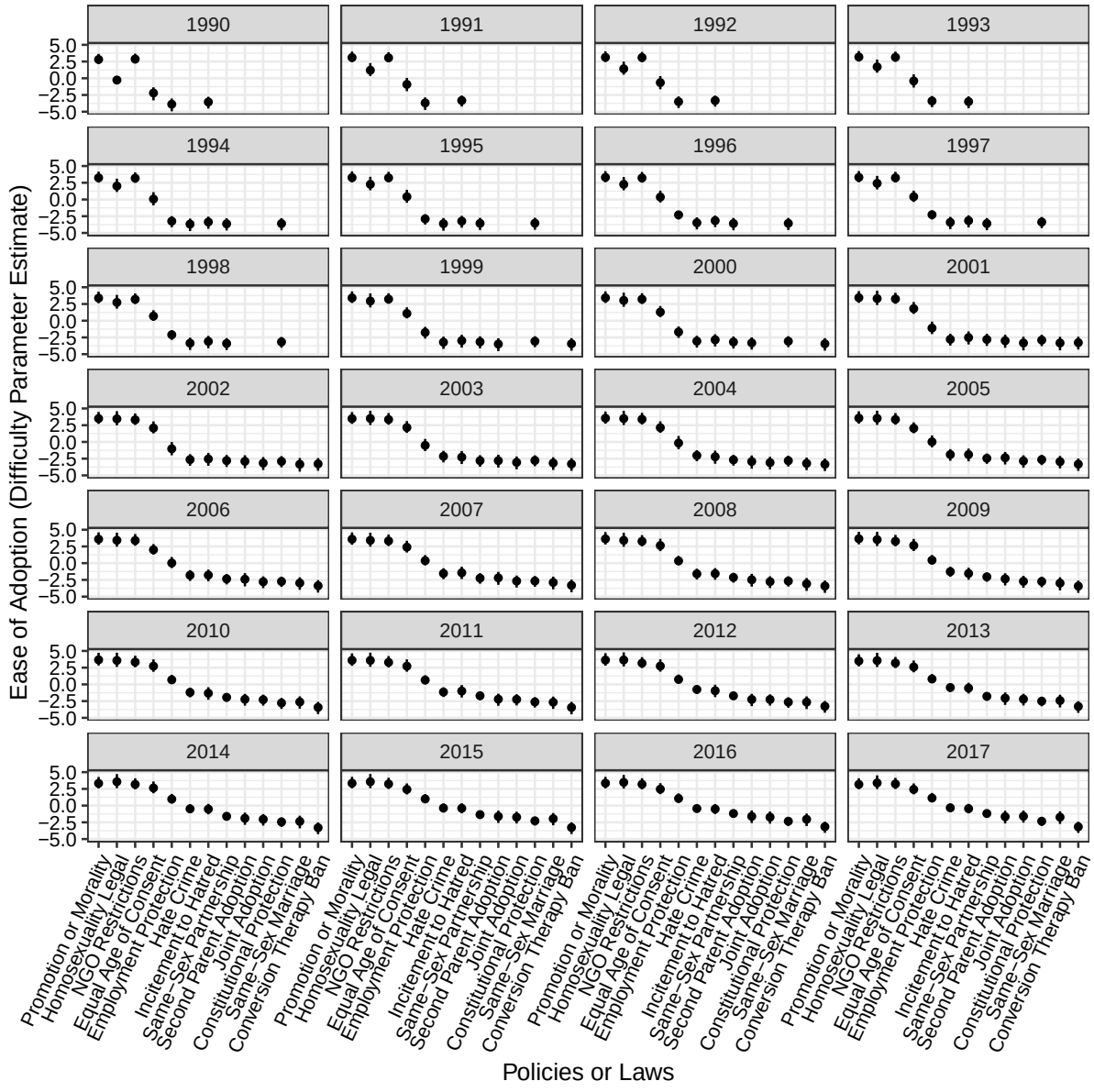

Figure 9: Difficulty parameter estimates with 95% credible intervals for each policy/law. Larger estimates correspond with policies/laws that are "easier" to adopt.

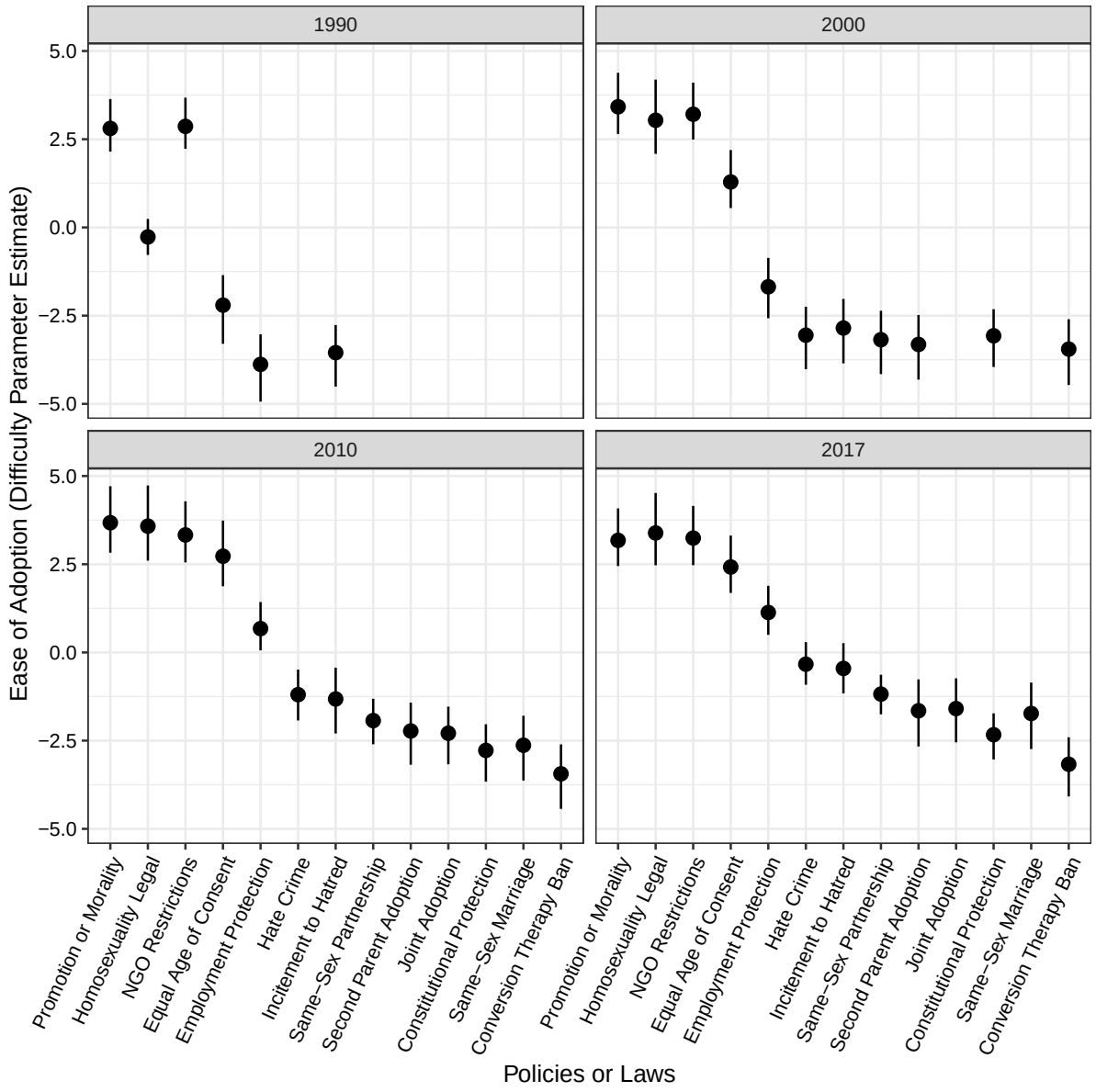

Figure 10: Difficulty parameter estimates with 95% credible intervals for each policy/law. Larger estimates correspond with policies/laws that are "easier" to adopt.

References

Armstrong II, David A, Ryan Bakker, Royce Carroll, Christopher Hare, Keith T Poole and Howard Rosenthal. 2014. *Analyzing Spatial Models of Choice and Judgment with R*. Boca Raton, FL: Chapman and Hall/CRC.

Bailey, Michael A, Anton Strezhnev and Erik Voeten. 2017. "Estimating Dynamic State Preferences from United Nations Voting Data." *Journal of Conflict Resolution* 61(2):430–456.

Baker, Frank B. 2001. *The Basics of Item Response Theory*. College Park, MD: ERIC.

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.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Correlates of War. 2017. "State System Membership List, v2016.".
URL: *http://correlatesofwar.org*

Fariss, Christopher J. 2014. "Respect for Human Rights Has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–318.

Fariss, Christopher J. 2018. "The Changing Standard of Accountability and the Positive Relationship Between Human Rights Treaty Ratification and Compliance." *British Journal of Political Science* 48(1):239–271.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Hierarchical/Multilevel Models*. New York, NY: Cambridge University Press.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

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

Martin, Andrew D and Kevin M Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153.

Patz, Richard J and Brian W Junker. 1999. "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models." *Journal of Educational and Behavioral Statistics* 24(2):146–178.

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.

Reuning, Kevin, Michael R Kenwick and Christopher J Fariss. 2019. "Exploring the Dynamics of Latent Variable Models." *Political Analysis* 27(4):503–517.

Schnakenberg, Keith E and Christopher J Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Treier, Shawn and D Sunshine Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73(4):679–703.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.