Estimating Smooth Country–Year Panels of Public Opinion

Christopher Claassen

School of Social and Political Sciences, University of Glasgow, Glasgow G12 8QQ, UK.
Email: christopher.claassen@glasgow.ac.uk

Abstract

At the microlevel, comparative public opinion data are abundant. But at the macrolevel—the level where many prominent hypotheses in political behavior are believed to operate—data are scarce. In response, this paper develops a Bayesian dynamic latent trait modeling framework for measuring smooth country–year panels of public opinion even when data are fragmented across time, space, and survey item. Six models are derived from this framework, applied to opinion data on support for democracy, and validated using tests of internal, external, construct, and convergent validity. The best model is reasonably accurate, with predicted responses that deviate from the true response proportions in a held-out test dataset by 6 percentage points. In addition, the smoothed country–year estimates of support for democracy have both construct and convergent validity, with spatiotemporal patterns and associations with other covariates that are consistent with previous research.

Keywords: latent variables, time series, hierarchical modeling, Bayesian estimation, public opinion, support for democracy

1 Introduction

Social scientists are awash in public opinion data. Over a dozen cross-national survey projects are now in existence, regularly asking nationally representative samples of respondents in all continents and regions their opinions on a diverse range of social and political topics. Few countries have not been surveyed at one time or another, and many countries have been polled numerous times, sometimes by several of these survey projects. At the dawn of cross-national public opinion research, when Almond and Verba (1963) completed their pioneering five-country study, researchers could hardly have dreamed of such a vast trove of public opinion survey data.

Yet, by another standard, public opinion data are scarce. Many theories of political behavior propose country-level relationships between aggregate opinion and political outcomes. For example, the theory of social capital proposes that social trust bolsters the quality of governance (Putnam 1993); the literature on policy-making argues that preferences shape policy choices (Stimson 1991); scholars of political tolerance claim that intolerance leads to the repression of dissent (Sullivan, Piereson, and Marcus 1982); and studies of democratization hypothesize that support for democracy helps sustain a democratic regime (Lipset 1959). When aggregated to the country level, however, a typical survey sample of one to two thousand respondents diminishes to a single data point. Thus, although we may have millions of respondents' opinions on a particular topic, we might only have a few hundred nationally aggregated opinions. While such a quantity of aggregate opinion data may be sufficient to assemble a cross-section of countries, comparing Sweden to Slovakia to Somalia at one point in time does not allow us to test the dynamic, causal hypotheses that animate much of our research.

Author's note: I am grateful for the helpful comments provided by Devin Caughey, Roberto Stefan Foa, Duncan Lee, Anthony J. McGann, Jamie Monogan, and Richard Traunmüller on earlier versions of this paper. I acknowledge the financial support of the Carnegie Trust for the Universities of Scotland and the Adam Smith Research Foundation at the University of Glasgow. Finally, I appreciate the research assistance provided by Jose Ricardo Villanueva Lira and Bryony MacLeod. Replication materials are provided in the *Political Analysis* dataverse (Claassen 2018).

Political Analysis (2019)
vol. 27:1–20
DOI: 10.1017/pan.2018.32

Published
4 July 2018

Corresponding author
Christopher Claassen

Edited by
Jonathan N. Katz

Instead, the fact that several cross-national public opinion survey projects have been running since the 1990s, if not before, offers a tantalizing possibility of measuring a panel of public opinion that varies across both space and time. Such country–year panels of public opinion would not only be of descriptive interest, they would also allow scholars to incorporate public opinion in studies of comparative political behavior and comparative political economy; in some cases, for the first time.

Unfortunately, however, aggregate public opinion data are not distributed neatly or evenly across space and time. Cross-national surveys are clustered in certain places and times, with fragmented time series and large spatial gaps. To make matters considerably worse, major public opinion concepts are typically measured in multiple ways, with the wording of questions or the nature of response sets varying widely, both within and across survey projects. Any analyst seeking to assemble a country–year panel of aggregate public opinion would thus appear to have to rely only on a single survey question. As a consequence, any panel that is assembled out of available data will be highly fragmented, with sparse coverage over space and time.

This paper proposes a method for estimating smooth panels of aggregate public opinion using all available survey data. The idea is to harness existing data to estimate latent country–year opinion, adjusting for the biases induced by different survey items and differential item functioning across countries, and smoothing over time. While a number of scholars have developed methods for smoothing single-country time series of aggregate opinion (e.g., Beck 1989; Stimson 1991; Jackman 2005; Voeten and Brewer 2006; McGann 2014), none have focused as yet on *cross-national panels* of opinion. The contribution of this paper is the development and validation of such a method. This method will be of interest to scholars of comparative political behavior and comparative political economy who would benefit from access to country–year panels of opinions on policy mood, social values, political culture, and so on.

2 Existing Research on Smoothing Aggregate Public Opinion

Political scientists have long been interested in smoothing estimates of aggregate public opinion over time. A pioneer in this area is Stimson (1991), who estimated ideology, or "policy mood," in the United States between 1956 and 1988. To accomplish this goal, he developed an ingenious dyad-ratios algorithm. This algorithm rests on the realization that while the level of respondent agreement varies idiosyncratically across survey items, the change over time in respondent agreement can be compared across items. The dyad-ratios algorithm thus uses the ratio of change over time to standardize survey items. These ratios are then combined using a factor analysis type procedure that weighs each item's ratio of change by the degree to which it correlates with the latent variable. Finally, the estimates are smoothed over time using an exponential smoother.

The dyad-ratios algorithm has been phenomenally popular. Stimson and colleagues have used it in several major studies of public opinion (e.g., Stimson, Mackuen, and Erikson 1995; Erikson, Mackuen, and Stimson 2002) and scholars continue to use it to this day to estimate smooth time series of aggregate opinion (e.g., Baumgartner, De Boef, and Boydstun 2008; Enns 2016). However, two years before Stimson, Beck (1989) provided an alternative, arguably more theoretically grounded, approach to smoothing aggregate opinion. He described a model of dynamic opinion that included a Kalman filter to smooth over time and a measurement model to combine multiple items into one opinion series. Indeed, Beck went even further by showing how the latent opinion estimates could be modeled using a set of covariates. Although he managed to fit and run such a sophisticated model using GAUSS software on a "386-based microcomputer," it appears that Beck was somewhat ahead of his time. It was the much simpler dyads ratio algorithm that became popular.

In recent years, methodologists interested in measuring latent aggregate opinion have proposed similar dynamic measurement models to Beck's (e.g., Green, Gerber, and De Boef 1999).

However, it is really with the rise of Bayesian methods—which not only include prior information to help estimate and identify complex models, but also provide a highly intuitive framework for understanding hierarchical and dynamic models—that smoothing aggregate opinion takes off.

Within this Bayesian approach, Jackman (2005) provides an early, seminal contribution, smoothing opinion over an electoral campaign by modeling observed polling marginals as true opinion plus random error. True opinion is additionally adjusted for biases induced by the methods used by survey firms, and is furthermore allowed to evolve over the campaign using a random walk error process. Voeten and Brewer (2006) estimate US public approval of the 2003 war in Iraq using a model developed from Jackman's framework. In addition, following Beck (1989), they include item intercepts and slopes to allow opinion to be combined from different survey questions.

While Voeten and Brewer (2006) and Jackman (2005) use linear models of the proportion of respondents who offer a particular opinion, Linzer (2013) instead uses a binomial specification to model the *number* of respondents offering a particular opinion. This neatly allows sampling error to be included in the estimates, but also allows survey items where almost all or almost no respondents agree (and thus proportions close to 0 or 1) to be more accurately modeled. McGann (2014) follows suit, but additionally includes a two-parameter item response theory (IRT) measurement model to estimate item effects. He further includes a beta prior on the binomial probability parameter to capture the overdispersion in survey data introduced by idiosyncrasies peculiar to survey data such as varying survey modes, methods of sampling, and so on.

This literature has focused on estimating an opinion time series within a single country. No researchers have attempted to extend these models to measure opinion across countries as well over time. Caughey and Warshaw (2015), however, have extended methods of smoothing opinion time series in a related direction by developing a "dynamic group IRT model" (DGIRT) for estimating opinions over time in small subnational groups. To do so, they combine a binomial IRT model, the method of multilevel regression and poststratification (MRP; Park, Gelman, and Bafumi (2004)), and a dynamic linear model of the latent opinion. Their model allows one to estimate opinion within small demographic and geographic groups and over time. It also allows for differing survey questions and surveys that are fragmented over time and space. The model is very powerful, but also very complicated.[1] Although the DGIRT model could be used to measure opinion across countries using complete national samples, much of the complexity comes from allowing the analyst to estimate subnational opinion with small and unrepresentative samples.

This paper instead focuses specifically on estimating country–year panels of opinion. The assumption is that nationally aggregated survey marginals are drawn from representative samples—or have been weighted to approximate representativity. The challenge is then to accurately measure opinion despite gaps in time, space, and survey item. The models for doing so are developed and presented in the next section.

3 Modeling Cross-National, Time-Series Latent Opinion

What do we require of a model of cross-national, time-series opinion? There are four guiding principles from existing research. First, we should treat opinion as an unobserved, latent trait, with observed survey responses being a function of these latent country–year traits. In effect we should set up a measurement model with latent estimates of country by time, as well as item-specific parameters to adjust the location and scale of the link between observed responses and aggregate opinion (Beck 1989; Voeten and Brewer 2006; McGann 2014; Caughey and Warshaw 2015).

Second, while classic measurement models—whether in the factor analytic or IRT traditions—can be thought of as estimating latent variables by smoothing over (for example) survey items,

1 Indeed Caughey and Warshaw (2015) note in a footnote that one of their models took "several weeks" of computing time to fit. In contrast, the models presented here all converge in one to three hours on a desktop computer.

Table 1. Models.

Model number	Response distribution	Item intercepts (λ)	Item–country intercepts (δ)	Item slopes (γ)
1	Binomial	✓		
2	Binomial	✓	✓	
3	Binomial	✓	✓	✓
4	Beta-binomial	✓		
5	Beta-binomial	✓	✓	
6	Beta-binomial	✓	✓	✓

Beck (1989), Voeten and Brewer (2006), and Caughey and Warshaw (2015) extend these models by additionally smoothing over time. I will follow suit by incorporating a model of temporal dynamics.

Third, we should model the number of respondents—rather than the derived proportion or percentage—offering an affirmative (or dissenting) opinion. This implies a binomial model linking observed responses to the measurement model (Linzer 2013; Caughey and Warshaw 2015). Such a specification allows for sampling error to be included. We can also extend this formulation by using a beta-binomial link, which includes an additional dispersion parameter (McGann 2014). This includes additional uncertainty in the estimates beyond mere sampling error.

Finally, since we are interested in modeling opinions across countries, we ought to adjust for heterogeneous item functioning, which is unfortunately quite prevalent in cross-national public opinion (Stegmueller 2011). I consider ways of accomplishing this below.

Following these principles, I develop six models of cross-national, time-series opinion (a summary of the six models is provided in Table 1). These models are tested using a real-world application: estimating support for democracy. Using both internal and external validation, I test the accuracy of the six sets of point estimates and variance estimates, and select a preferred model.[2]

3.1 Distributions

The observed number of respondents y_{ikt} offering an affirmative opinion (e.g., in support of democracy) for each country i, year t, and survey item k, is modeled as a binomial distributed count:

$$y_{ikt} \sim \text{Binomial}(s_{ikt}, \pi_{ikt}). \tag{1}$$

There are then two ways of proceeding. In the simpler binomial specification I model the probability π_{ikt} of offering an affirmative opinion directly, as a function of item and country–time effects (e.g., Linzer 2013; Caughey and Warshaw 2015). However one could also follow McGann (2014) in utilizing a beta prior on the probability parameter. This allows for some additional dispersion in the observed survey responses, which captures sources of error over and above simple sampling error. Indeed, since public opinion survey data are afflicted by numerous sources of errors—including methods of questionnaire translation and respondent selection, survey mode, and interviewing style (e.g., Weisberg 2005)—allowing for overdispersion in survey responses appears to be a prudent course of action.

I thus use the simpler binomial specification for three of the six models, and the binomial with beta prior, or beta-binomial, for the other three. The beta-binomial specification then also

2 Replication data and code are available on the *Political Analysis* dataverse. See Claassen (2018).

includes the following step:

$$\pi_{ikt} \sim \text{Beta}(\alpha_{ikt}, \beta_{ikt}). \tag{2}$$

The two shape parameters of the beta distribution can be reparameterized to an expectation parameter, η, and a dispersion parameter, ϕ:

$$\alpha_{ikt} = \phi\eta_{ikt} \tag{3}$$

$$\beta_{ikt} = \phi(1 - \eta_{ikt}). \tag{4}$$

3.2 Item and country parameters

In the case of the binomial specification, the probability parameters are modeled directly as a function of the latent country–year estimates and item parameters; in the case of the beta-binomial, the beta expectation parameter receives this measurement model. I utilize three variations of this measurement model. The first simply includes country–year latent effects and item intercepts (the beta-binomial version is shown here):

$$\eta_{ikt} = \text{logit}^{-1}(\lambda_k + \theta_{it}) \tag{5}$$

$$\lambda_k \sim \text{N}(\mu_\lambda,\ \sigma^2_\lambda). \tag{6}$$

The item intercepts λ adjust the location of the latent opinions for the idiosyncrasies of each survey item. They can thus be thought of as item bias effects. These intercepts are modeled hierarchically, with an expectation μ_λ and variance σ^2_λ estimated from the data. This hierarchical specification shrinks the item intercepts toward the mean to the extent that data are scarce, which guards against small within-item samples producing extreme estimates.

Survey items are, moreover, likely to have differing effects in different countries, a problem known as lack of equivalence (Stegmueller 2011). For example, one method of measuring support for democracy is to ask respondents for their opinions about having the army govern the country. Respondents in countries with a history of military rule are likely to respond quite differently than respondents in countries without such a history.

Fortunately, each item is asked multiple times in a given country. When replicates of items across units (respondents or countries) are available, analysts may also include parameters capturing item by unit bias (Skrondal and Rabe-Hesketh 2004). The second version of the measurement model thus includes a set of item by country effects δ to capture the heterogeneity in item bias across countries:

$$\eta_{ikt} = \text{logit}^{-1}(\lambda_k + \delta_{ik} + \theta_{it}) \tag{7}$$

$$\delta_{ik} \sim \text{N}(0,\ \sigma^2_\delta). \tag{8}$$

These item–country intercepts are also modeled hierarchically, which is helpful as the observed data are especially likely to be sparse when divided by country as well as item. By treating both the item and item–country effects as varying intercepts, or random effects, they can be interpreted as error terms (McGraw and Wong 1996). This lends an intuitive understanding to their role in the measurement equation: the λ effects can be seen as the item-level residuals, while the δ effects can be seen as the item–country-level residuals, leaving θ as the item and item–country adjusted estimates of latent support for democracy.

Finally, measurement models often also include item slopes, known as factor loadings in the factor analysis framework and discrimination parameters within the IRT approach. Whatever the name, item slopes allow the strength of the relationship between observed responses and latent

traits to vary across the items. Where an item shows a weaker relationship with the latent variable, it "loads" to a lesser extent than items showing a stronger relationship. I extend the second model by incorporating such item slopes γ:

$$\eta_{ikt} = \text{logit}^{-1}(\lambda_k + \delta_{ik} + \gamma_k \theta_{it}). \tag{9}$$

With both varying intercepts and varying slopes, this is a type of hierarchical (generalized) linear model, with observed responses nested within both items and countries (ignore time for the moment). As such, it is desirable to model the item-level equations jointly using a bivariate normal (Skrondal and Rabe-Hesketh 2004; Gelman and Hill 2007). This allows item intercepts and slopes to be correlated, with the ρ parameter capturing the degree of covariation.

$$\begin{pmatrix} \lambda_k \\ \gamma_k \end{pmatrix} \sim \text{N}\left[\begin{pmatrix} \mu_\lambda \\ \mu_\gamma \end{pmatrix}, \begin{pmatrix} \sigma_\lambda^2 & \rho\sigma_\lambda\sigma_\gamma \\ \rho\sigma_\lambda\sigma_\gamma & \sigma_\gamma^2 \end{pmatrix} \right]. \tag{10}$$

With three versions of the measurement model coupled with the two response distributions (binomial and beta-binomial), there are six models in total. These are outlined in Table 1.

3.3 Dynamic effects

Finally, for all six models, the latent opinion estimates are allowed to evolve over time. Doing so smooths over any gaps in each national time series. Following previous research on modeling dynamic latent traits (e.g., Jackman 2005; Caughey and Warshaw 2015), the temporal evolution of latent opinion is specified as a simple local-level dynamic linear model (Durbin and Koopman 2012), where the current level of latent opinion is a function of the previous year's level plus some random noise:

$$\theta_{it} \sim \text{N}(\theta_{i,t-1}, \sigma_\theta^2). \tag{11}$$

The variance of the noise term, σ_θ^2, is held constant across countries and estimated from the data.

3.4 Identification and priors

To identify latent trait models, analysts must impose restrictions on the location, scale, and perhaps also the direction (or sign) of the parameters (Bafumi *et al.* 2005). Models without item loadings are simpler in this respect as they only require location restrictions. These models were identified by fixing the first item intercept λ_1 at a value of 1. Models with item loadings (models 3 and 6) additionally require that the scale and direction of the parameters be identified. To do so, I fix the expectation of the item intercepts μ_λ to 0.5, and the expectation of the item slopes μ_γ to 1.[3] The direction of the item slopes γ is then identified by constraining these to be positive.

The estimated variances are given weakly informative half-Cauchy priors: $\sigma_\lambda \sim \text{C}^+(0, 2)$ (and similarly for σ_δ, σ_γ, and σ_θ). For the models including item slopes as well as intercepts, the variance–covariance matrix of item intercepts and slopes is decomposed into the product of the variances for each vector of parameters and a 2×2 correlation matrix, with ρ being the estimated correlation (Stan Development Team 2017). This correlation matrix is given an LKJ prior (Lewandowski, Kurowicka, and Joe 2009).

The expectation of the item intercepts μ_λ (for models 1, 2, 4, and 5) is given a N(1, 2) prior while the dispersion parameter ϕ (for the beta-binomial models), is given a $\Gamma(4, 0.1)$ prior. Finally, for all models, the initial value of latent opinion for each country, θ_{i1}, receives a N(0, 1) prior.

3 For these models λ_1 was not constrained.

4 Application: Support for Democracy

4.1 The concept of support for democracy

Political theorists since Aristotle have long argued the presence or absence of a democratic political system is somehow related to the attitudes and orientations of the citizenry. Interest in this theory continued into the modern era, leading to the development of the concept of political culture. According to scholars such as Inglehart and Welzel (2005) and Lipset (1959) democracy is legitimate—and stable—when it is "congruent" with the political culture. Put another way, democracy requires a democratic political culture.

There are in fact two distinct conceptualizations of democratic political culture. According to the first, citizens provide explicit support for democracy when they prefer a democratic regime to some nondemocratic alternative (e.g., Rose, Mishler, and Haerpfer 1998; Norris 1999; Mattes and Bratton 2007; Fuchs-Schündeln and Schündeln 2015). Here, democracy is legitimate because it is believed to be preferable by the public. According to the second conceptualization, citizens provide implicit support when they subscribe to a broader set of values emphasizing trust, tolerance, and freedom (Inglehart 2003; Inglehart and Welzel 2005). Here, democracy is legitimate because it is consistent with citizen's deeper values and strivings. Although both kinds of democratic political culture have been advocated as providing support for democracy, or perhaps even spurring democratization, the focus of this paper is on the first kind, explicit support for democracy, often simply referred to as "support for democracy."[4]

4.2 Data on support for democracy

I collected all available nationally aggregated responses to questions on support for democracy that were gathered by cross-national survey projects utilizing representative national samples of citizens. Data are collected from eleven survey projects including the World Values Survey and all the Global Barometer projects (see online supplementary materials available online at https://doi.org/10.1017/pan.2018.32). Surveys with relevant items were fielded in 144 countries over a 24 year period between 1992 and 2015. There are 3,014 nationally aggregated responses, obtained from 1,165 separate national survey samples.[5]

These data epitomize the challenges of measuring cross-national time-series opinion. First, they are sparse over time and space. If the focus is restricted to the 132 countries that were surveyed at least twice on explicit support over the years from 1992 to 2015, there is a potential dataset of 3,168 country–years. However, surveys were conducted in just over a third of these country–years. I present a visualization of the sparseness of this data in Figure 1. The top panel shows the fragmented supply of support for democracy measures across time for eight countries, selected for variance across regions and in availability of data.

To take the example of South Africa, questions on explicit support for democracy were asked in 11 national surveys over the period in question: by the World Values Survey in 1996, 2001, 2006, and 2013, the AfroBarometer project in 1999, 2003, 2005, 2008, and 2012, and Pew Global Attitudes in 2002 and 2013. Data on democratic support are thus only available for ten out of the 24 years in South Africa—and this is a case that has above average survey coverage.

Second, to compound the problem, researchers have not settled on standard survey questions for measuring democratic political culture. Indeed, there is an extraordinary diversity of approaches to measuring support for democracy: as many as 37 different survey questions

4 Scholars such as Canache, Mondak, and Seligson (2001) and Norris (1999) demonstrate that support for democracy must be distinguished from the concept of "satisfaction with democracy," measures of which are also widespread in comparative survey projects. I follow suit.

5 World Values Survey data were excluded for some countries and items due to known problems in some cases, and suspicious response levels in others. See Kurzman (2014) for further discussion and the online supplementary materials for details.

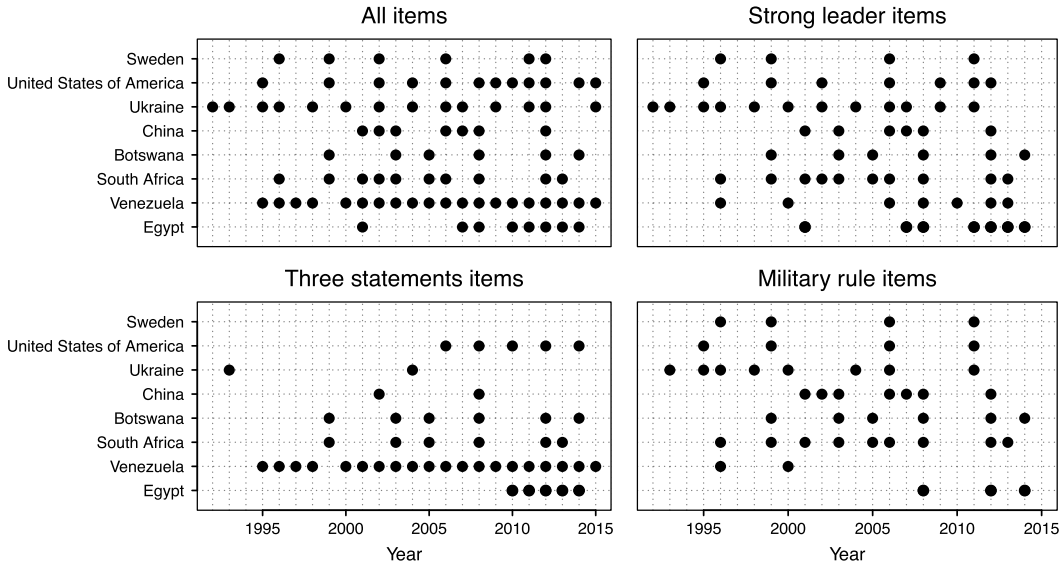

Figure 1. Sparseness of Aggregate Support for Democracy By Country, Year, and Survey Item. The first panel shows the availability of at least one survey item across a selection of eight countries and all 24 years. The other three panels indicate the availability of data for the three most common question themes. The wording of all survey items is included in the online supplementary materials.

clustered within nine broad approaches to question wording.[6] The lower panels of Figure 1 show the supply of support for democracy data within the three most prominent of these measurement approaches. Once disaggregated in this way, the data are clearly even more sparse across the country–year matrix.

To continue the South African example, although 11 surveys fielded questions on support for democracy in this country, each used several different items, resulting in 41 data points in total. These 41 data points are however fragmented across seven different questions. If one were to focus on a single survey question to obtain a meaningful time series, most of the data would have to be discarded. Even the most popular survey item, the question asking respondents the extent to which they support or oppose having a strong but undemocratic leader, is asked only in ten out of the 24 years.

As such, once available data on support for democracy are divided by survey item as well as country and year, they begin to look very sparse indeed. If survey questions from every one of the nine major approaches to question wording were asked annually in each country, there would be 28,512 country-level data points. Yet, only slightly more than 10% of these potential item–country–year cells actually feature data. Classical methods of measuring public opinion using multiple items, such as factor analysis of the complete cases, are simply not an option here. The data are too sparsely scattered across survey item, country, and year. Indeed, there are no complete cases unless one ignores the temporal dimension.

Given these limitations, analysts have tended to abandon any temporal variation in support for democracy (and other measures of cross-national opinion), focusing instead on cross-sectional variation. I aim to overcome these limitations by estimating support for democracy over 132 countries and over 24 years, creating a full time-series, cross-sectional dataset of 3,168 observations.

6 I took a conservative route in categorizing survey items by always classing two items as distinct when they were fielded by different projects even if their wording appeared to be identical. Doing so allows the item effect parameters to capture variation induced both by question wording and by idiosyncrasies in the methodology of the various survey projects.

4.3 Estimation

The six models are estimated using Bayesian Markov-Chain Monte Carlo (MCMC) methods via Stan software, which implements Hamiltonian Monte Carlo sampling (Carpenter *et al.* 2017; Stan Development Team 2017). Four parallel chains were run for 1,000 samples each, with the first 500 samples in each chain used for warm up, and discarded, and the remaining 2,000 samples of the posterior density thinned by half and analyzed further. This number of iterations proved to be more than sufficient for convergence, with the $\hat{R}$ diagnostic reaching a value of between 0.95 and 1.05 for all parameters in all models.

4.4 Preliminary results

Before testing the accuracy of the six models and verifying the validity of the estimates, I include a table (Table 2) showing parameter estimates (from model 5) and observed response proportions for three countries—the US, South Africa, and China—over five years—2005 to 2009. This table includes labels for country, year, and survey item; parameter estimates for item bias, item–country bias, and the country–year latent opinion; and both observed and simulated response proportions.

By employing a "dataset" perspective, Table 2 helps explicate the modeling framework employed in this paper. In particular, in certain countries and years (e.g., the first row in the table, corresponding to the United States in 2005), no public opinion surveys asking support for democracy questions were fielded. The modeling framework, however, can estimate opinion even in years where no survey measures were available. The table also shows that other country–year combinations (such as the US the following year, in 2006) benefit from having several observed survey responses.

Although the main focus of the paper will be on the latent country–year estimates θ, Table 2 also allows readers to compare observed response proportions (y_{ikt}/s_{ikt}) with those simulated from the model, contingent on the other parameters ($\tilde{y}_{ikt}/s_{ikt}$). Doing so helps demonstrate that the modeling framework can be compared to simple linear or generalized linear models, where an outcome (here, y_{ikt}) is modeled as a function of some parameters and/or data. Simulated response proportions are also used to test and compare the accuracy of the six models, which is the task we turn to next.

5 Validation Tests

5.1 Internal and external validation

In this section, I compare the accuracy of the latent opinion estimates obtained from each model, both in comparison to an absolute standard, and in comparison to each other. The first test is a test of the predictive accuracy of the models when using the same data that were used to fit the model, which Hastie, Tibshirani, and Friedman (2009) refer to as internal validation. In particular, the mean absolute error (MAE) is used to measure the average discrepancy between the observed proportions of respondents offering a prodemocratic attitude y_{ikt}/s_{ikt}, and the simulated proportions $\tilde{y}_{ikt}/s_{ikt}$:

$$\text{MAE} = \frac{1}{J}\sum_{j\in ikt}^{J}\left|\frac{y_{ikt}}{s_{ikt}} - \frac{\tilde{y}_{ikt}}{s_{ikt}}\right|. \tag{12}$$

Internal validation is simple to conduct, but favors more complex models. Reliance on metrics of internal validation could therefore lead to the selection of a model that overfits the dataset at hand. Analysts thus instead use information criteria, which attempt to estimate out-of-sample predictive error by penalizing models as their parameters increase in number. A good choice of information criterion for Bayesian MCMC methods is the "Leave-One-Out" information criterion

Table 2. Estimated parameters and observed data for three countries and five years.

Project & item type (k)	Country (i)	Year (t)	Country latent est. (θ_{it})	Item bias (λ_k)	Item-country bias (δ_{ik})	Obs'd. prop. $\left(\frac{y_{ikt}}{s_{ikt}}\right)$	Sim'd. prop. $\left(\frac{\tilde{y}_{ikt}}{s_{ikt}}\right)$
	USA	2005	0.36				
WVS–Eval. dem.		2006	0.29	1.74	−0.53	0.82	0.83
LAPOP–Churchill		2006	0.29	0.83	0.32	0.93	0.82
WVS–Army rule		2006	0.29	1.13	0.24	0.84	0.85
WVS–Strong leader		2006	0.29	0.32	−0.02	0.65	0.66
WVS–Import. dem.		2006	0.29	2.09	−0.46	0.85	0.88
LAPOP–3 statements		2006	0.29	1.00	0.10	0.75	0.81
		2007	0.21				
LAPOP–Churchill		2008	0.18	0.83	0.32	0.77	0.80
LAPOP–3 statements		2008	0.18	1.00	0.10	0.79	0.79
Pew–Strong leader		2009	0.22	0.55	−0.48	0.53	0.56
AfroB–1 party rule	South	2005	−0.26	1.06	−0.15	0.66	0.65
AfroB–Army rule	Africa	2005	−0.26	1.00	0.25	0.72	0.72
AfroB–Strong leader		2005	−0.26	1.30	−0.23	0.64	0.68
AfroB–3 statements		2005	−0.26	0.70	0.20	0.65	0.65
WVS–Eval. dem.		2006	−0.32	1.74	−0.10	0.86	0.80
WVS–Army rule		2006	−0.32	1.13	−0.34	0.60	0.63
WVS–Strong leader		2006	−0.32	0.32	0.02	0.50	0.52
WVS–Import. dem.		2006	−0.32	2.09	0.14	0.91	0.88
		2007	−0.38				
AfroB–1 party rule		2008	−0.38	1.06	−0.15	0.63	0.63
AfroB–Army rule		2008	−0.38	1.00	0.25	0.67	0.70
AfroB–Strong leader		2008	−0.38	1.30	−0.23	0.63	0.67
AfroB–3 statements		2008	−0.38	0.70	0.20	0.67	0.63
		2009	−0.38				
	China	2005	−0.52				
AsiaB–Army rule		2006	−0.75	0.52	−0.09	0.46	0.48
AsiaB–Strong leader		2006	−0.75	0.52	0.92	0.83	0.71
AsiaB–Eval. dem.		2006	−0.75	2.03	0.75	0.93	0.90
WVS–Eval. dem.		2007	−0.74	1.74	−0.35	0.61	0.66
WVS–Army rule		2007	−0.74	1.13	−0.60	0.38	0.45
WVS–Strong leader		2007	−0.74	0.32	0.14	0.42	0.43
WVS–Import. dem.		2007	−0.74	2.09	0.01	0.78	0.79
AsianB–3 statements		2008	−0.67	0.54	0.26	0.54	0.52
AsianB–Army rule		2008	−0.67	1.38	−0.25	0.57	0.59
AsianB–Desire for dem.		2008	−0.67	2.05	−0.45	0.65	0.71
AsianB–Dem. suitable		2008	−0.67	1.77	−0.19	0.72	0.69
AsianB–Strong leader		2008	−0.67	1.17	0.03	0.61	0.61
		2009	−0.61				

Parameter estimates are drawn from those obtained using Model 5 and are unstandardized.

(LOO-IC), which Vehtari, Gelman, and Gabry (2017) argue to be superior to alternatives such as the Deviance and Watanabe–Akaike information criteria. I follow their advice in including Stan code for estimating the LOO-IC for each of the six models, and report these results below.

Better still, however, is to select models using data that are not also used to fit the model, which is known as external validation (Hastie, Tibshirani, and Friedman 2009). To do so, the dataset of national opinions is randomly split into a 75% training set and a 25% test set. The six models are fit to the training set, and the resulting parameter estimates are used to predict the national proportions offering a supportive (i.e., prodemocratic) response for each of the 744 survey items comprising the test dataset. I again calculate the MAE, but now in predicting the error in the held-out test dataset.

Finally, I examine the accuracy of the estimates of uncertainty produced by each models by calculating their credible interval coverage (CIC). To do so, I find—for each model—the percentage of the $J = 744$ observed survey proportions that are included in the 80% credible interval of the corresponding simulated survey proportions ($\tilde{y}_j / s_j$):

$$\text{CIC} = \frac{100}{J} \sum_{j=1}^{J} \left[\frac{y_j}{s_j} \in \text{CI}_{80} \left(\frac{\tilde{y}_j}{s_j} \right) \right]. \tag{13}$$

To provide a baseline comparison for the validation tests, I also fit Caughey and Warshaw's (2015) DGIRT model to the training dataset and used it to predict responses on the test set.[7] Three naïve methods of estimating the out-of-sample proportions are also included as additional baselines. In the first of these, I use the country-mean proportions from the training dataset to predict the response proportions in the test set. Second, I use the item mean proportions, and third, I use the grand mean response proportion across the entire training dataset.

The results of these internal and external validation tests are displayed in Table 3. Beginning with the tests of internal validation, three findings are apparent. First, all six models offer a substantially better fit to the observed proportions supporting democracy than the baseline estimates. Indeed, taking the country means as a baseline, the most accurate models offer up to 79% reduction in MAE in tests of internal validation. This is hardly surprising given the difficulty in estimating the proportion responding affirmatively to a particular support for democracy survey item in a particular country and year knowing nothing other than the average proportion responding affirmatively in that country across all years and items. Yet it does show already that these models are adding value.

Perhaps more interesting is the second finding, which is that models 1 and 4—which include item intercepts but not item slopes or item–country intercepts—offer substantially worse fit than the other, more complex models. The error rate is roughly halved when adding item–country intercepts (which are incorporated in models 2, 3, 5, and 6). This result is confirmed by the LOO-IC measures. Although the LOO-IC penalizes the log-likelihood for the number of estimated parameters, models 2 and 3 offer a better fit than model 1, as do models 5 and 6 when compared with model 4.[8] There is however little to distinguish between the models with item slopes (3 and 6) and the models without (2 and 5).

I now turn to the external validation tests. These offer a better means of gauging model fit because the models are estimated and tested using different datasets. Focusing first on the MAE results, one can see that the six models continue to offer improvements in accuracy when compared with the baseline, country-mean estimates. The difference, however, has

7 I used the `dgirt()` function provided in the `dgo` package for R, which is created by Caughey and Warshaw and allows analysts to run the DGIRT model without having to delve into Stan.

8 The use of different distributions means that one cannot compare the LOO-IC across the binomial and beta-binomial specifications.

Table 3. Internal and External Validation Tests.

	Internal Validation Tests			External Validation Tests		
Model	Mean Absolute Error (MAE)	% Improvement in MAE	Leave-One-Out Information Criterion	Mean Absolute Error (MAE)	% Improvement in MAE	80% Credible Interval Coverage
1	0.050	52.4	143676	0.082	26.2	17.1
2	0.023	78.3	60561	0.070	37.1	39.4
3	0.022	79.0	58849	0.072	34.4	37.9
4	0.062	40.7	35956	0.070	36.3	38.2
5	0.032	69.5	34375	0.061	44.9	60.3
6	0.032	69.9	34354	0.062	44.3	60.8
DGIRT				0.088	20.0	17.5
Country means	0.105			0.110		
Item means	0.095	8.9		0.094	15.3	
Grand mean	0.129	−22.9		0.125	−13.2	

Internal validation uses the same data for model fitting and validation. External validation creates two separate datasets: models are fit to the 75% training set and validated using the 25% test (or hold-out) set. Percent improvement in MAE is a comparison between model MAE and country-mean MAE. The DGIRT model is proposed by Caughey and Warshaw (2015) and implemented in the dgo R package.

diminished when compared with the internal validation MAE results. Part of the gap between the model estimates and the baseline estimates were thus due to the models overfitting the data. Nevertheless, any of the six models offers a gain in accuracy over the baseline fit, with up to 45% reduction in MAE.

In addition, any of the models developed in this paper perform at least as well in predicting out-of-sample survey responses, and usually better, than Caughey and Warshaw's (2015) DGIRT model. The DGIRT model is about as accurate as model 1, which uses a binomial specification and includes only item intercepts. As mentioned, although the DGIRT model can, in principle, be used to estimate cross-national opinion, it was developed instead for estimating subnational opinion. Analysts face differing challenges in these two contexts: when estimating subnational opinion (but not cross-national opinion), samples are small and unrepresentative; when estimating cross-national opinion (but not subnational opinion), items may not be equivalent across countries. These results demonstrate that analysts should use models designed for the idiosyncrasies of cross-national opinion when estimating cross-national opinion.[9]

These external validation tests also confirm that the simpler, item-intercept only models (models 1 and 4) are the least accurate. The additional complexity added by including item–country intercepts (models 2 and 5) produces a meaningful reduction in MAE of around 1 percentage point when compared with the item-intercept only models. In contrast, adding item slopes or factor loadings does not improve predictive accuracy at all.

In addition, the external validation tests show that the beta-binomial (models 4–6) specifications are slightly more accurate than the corresponding binomial models (1–3), which contrasts with the results of the internal validation tests. The additional dispersion added by the beta-binomial enhances the accuracy of the estimates, perhaps by capturing some of the nonsampling error endemic in public opinion data.

9 An additional consideration is the length of time that is required to estimate such models. As Caughey and Warshaw (2015) point out, fitting the DGIRT model is time-consuming. It took me 56 hours to fit this model to the training dataset, compared with between one and two hours for each of my six models.

Finally, on to tests of CIC. These tests measure the accuracy of the estimates of uncertainty produced by each model. A model with accurate uncertainty estimates should have similar empirical and nominal levels of CIC. Since I use 80% credible intervals, one should expect 80% empirical coverage. Coverage substantially below this nominal level shows overly optimistic standard errors; coverage substantially above this level indicates overly conservative standard errors.[10]

The results indicate that the uncertainty estimates generated by the six models are all too optimistic. The standard errors, in other words, are too small. None of the rates of empirical CIC come appreciably close to the nominal level of 80%. There are of course, many sources of error in cross-national public opinion data (e.g., Weisberg 2005), and I have explicitly modeled only a few. Moreover, some sources of error—such as the translation problems in the World Values Survey identified by Kurzman (2014)—are impossible to model.

However, the beta-binomial specification, which includes an overdispersion parameter, ϕ, proves to have substantially more accurate uncertainty estimates than the simpler binomial specification. While the binomial CICs are very poor, ranging from 17 to 39%, the beta-binomial CICs are much closer to nominal 80% level, ranging from 38 to 61%. Including item–country bias effects also produces substantially better CIC, in both the binomial and beta-binomial specifications. Models 5 and 6, with both item–country intercepts and beta-binomial distributions, have the most accurate estimates of uncertainty of all.

Weighing up all the evidence from these tests of external validation, it would appear that the models which use beta-binomial specifications and include item–country effects (models 5 and 6) are the most accurate. These models show similarly low levels of error in predicting survey responses in the test dataset and have similar CIC. Model 5 has the advantage of being simpler to code and run than model 6, as it does not require the covariance of the item effects to be estimated. Model 6, however, might be useful for situations where analysts are unable to benefit, as we have here, from prior research investigating the reliabilities (or factor loadings) of potentially relevant survey items.

5.2 Construct and convergent validation

The external validation tests have demonstrated that my modeling framework—and in particular, the beta-binomial specification with item–country intercepts—is fairly accurate in predicting observed survey responses in a hold-out sample. To further build confidence in this approach I examine, in this subsection, whether the latent estimates correspond to the theoretical construct of support for democracy. I examine, in other words, whether the country–year measures of support for democracy behave as previous scholars have suggested this variable *should* behave.

In particular, I consider the cross-national distribution of estimated support for democracy at certain points in time and the over-time evolution of support for democracy for certain countries. These analyses will permit some discussion of the construct validity of the estimates. I also examine the correlation, at certain points in time, between cross-national point estimates of support for democracy and cross-national experience with democracy. These correlations constitute a test of the convergent validity of the measures. The analyses that follow utilize estimates obtained from model 5.

To assess the convergent validity of the model, I examine the covariation between support for democracy and cumulative experience with democracy at two points in time, 2015 and 2005. Cumulative experience with democracy is the sum of the democracy scores for a given country between the year in question and 1950, with each year's score discounted by 2%.[11] Scholars have

10 For example, an estimated interval between negative and positive infinity would show 100% coverage but would otherwise be completely uninformative.

11 I use the "Liberal democracy index" from the Varieties of Democracy project (Lindberg *et al.* 2014).

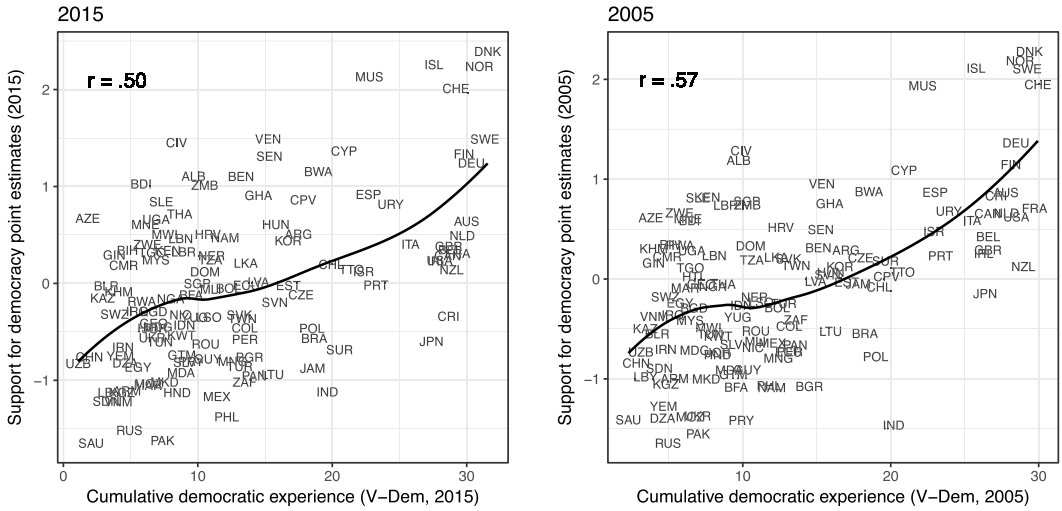

Figure 2. The Relationship Between Support for Democracy and Democratic Experience. Estimates of support for democracy, from Model 5, for all countries, plotted against cumulative democratic experience, in 2005 (top) and 2015 (bottom). Cumulative democratic experience is the sum of annual democracy scores discounted by 2% a year. A LOWESS line is added to each plot.

previously demonstrated that supportive attitudes toward democracy are linked with the length of time a country has been democratic (Mattes and Bratton 2007; Fuchs-Schündeln and Schündeln 2015). Scholars have also argued—although perhaps not yet empirically demonstrated—that support for democracy helps democratic institutions to survive (Lipset 1959; Rose, Mishler, and Haerpfer 1998; Norris 1999). Since either of these processes would lead to a correlation between support for democracy and democratic experience, this can be interpreted as a test of the convergent validity of the estimates. To carry out this test, I plot point estimates of support for democracy against democratic experience in both 2015 and 2005 (Figure 2).

In both 2005 and 2015, a robust and positive relationship is evident between the two variables (in 2005 the Pearson's correlation is 0.57; in 2015, 0.50). The more extensive a country's experience with democracy, the higher the support that its citizens express for a democratic versus an autocratic system. These correlations show that the estimates of latent opinion do in fact behave as theories of democratic political culture have suggested.

Moving on to construct validity, I examine the estimated levels of support for democracy for a selection of eight countries, which were selected as representing a range of levels of support as well as some interesting dynamics. These are displayed in Figure 3: each plot shows the latent estimates for a particular country over 24 years.[12] The darker line indicates the mean value of θ in each year; the lighter lines indicate 200 random draws from the posterior density of θ, and collectively show uncertainty in the latent estimates. Finally, the observed data are also displayed on these plots using points.[13]

First, when data are abundant in a particular country (e.g., Venezuela), the estimates are fairly precise (the y-axis is calibrated on the z-score scale). When data are scarce (e.g., Egypt and China before 2000), the estimates are noisier, and become increasingly so the larger the gap there is in the data. The beta-binomial model is fairly aggressive in smoothing across time compared with the binomial specification (not shown), which produces a more jagged, rapidly changing

12 Plots displaying the full set of opinion time series for all 132 countries are included in the online supplementary materials.

13 The observed data (which are national proportions offering support for democracy) are measured on a different scale to the latent estimates (which are unit-normal standardized). I thus standardized the observed responses by centering by survey item and dividing by the standard deviation of all responses. This places the observed data on approximately the same scale as the latent estimates.

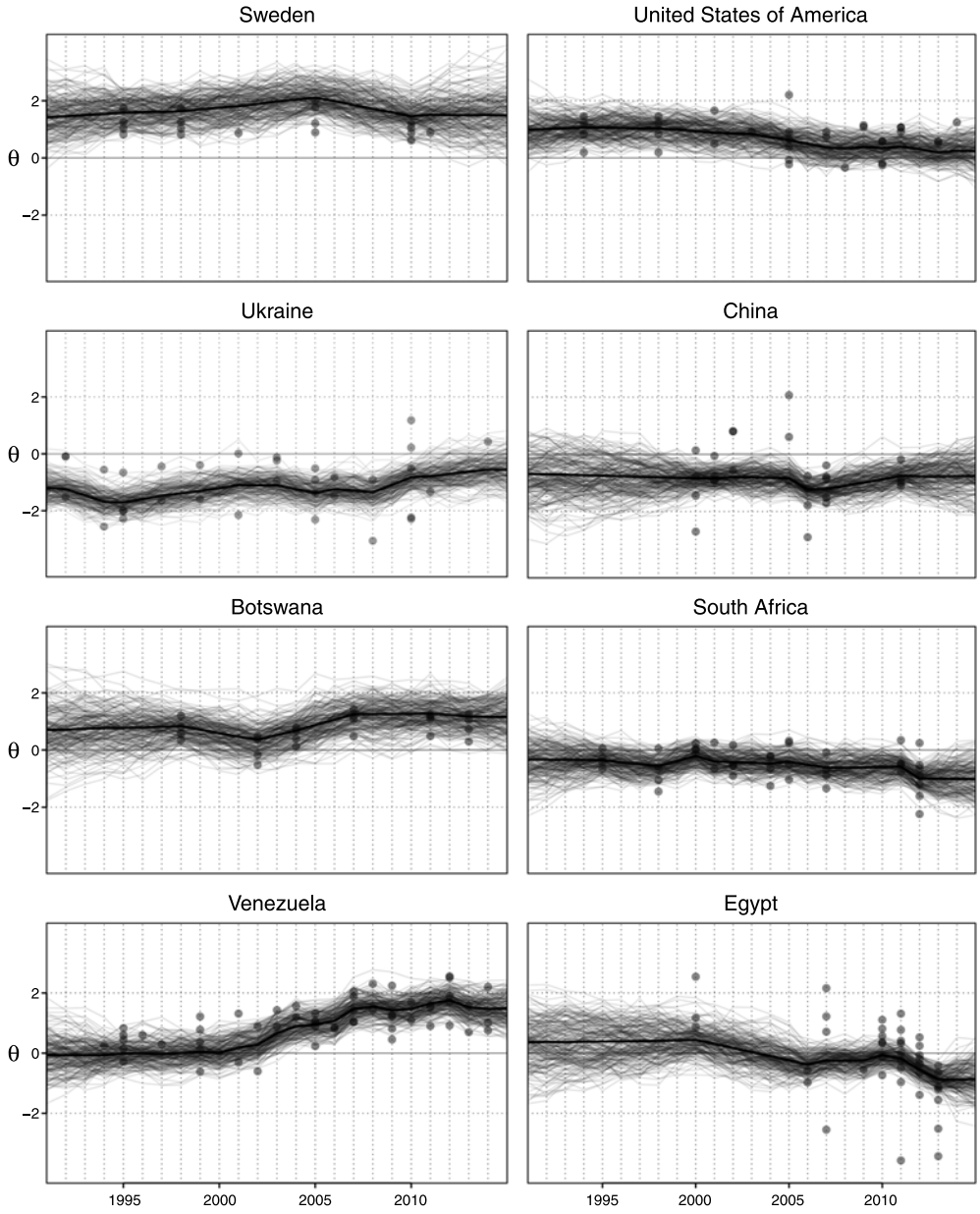

Figure 3. Estimated Support for Democracy for 8 Countries over 24 Years. Estimates of support for democracy, from Model 5, for eight selected countries over 24 years. Each plot shows 200 random draws from the posterior distribution of θ for a particular country. The posterior means are indicated using bold lines. Observed survey responses for each country are plotted using points; these are unit-normal standardized within survey item so as to display them on roughly the same scale as the latent estimates.

pattern of opinion. Such a pattern is not particularly plausible for a slow-moving orientation such as democratic political culture (Inglehart and Welzel 2005), which suggests again that the beta-binomial specification is preferable.

Second, established democracies such as Sweden show high levels of support for democracy. This is consistent with previous research focusing on particular subsets of the available survey data (e.g., Klingemann 1999). An important exception to this pattern is the Anglophone democracies, such as the United States, where declining support for democracy is evident. This is consistent with recent research by Foa and Mounk (2016).

Third, newer democracies show divergent trends. I examine a pair of cases from Southern Africa: Botswana has high and increasing support for democracy, which echoes previous case

study research (Hjort 2009). In contrast, Botswana's neighbor, South Africa, shows fairly low (and declining) support, which resonates with earlier survey research (Gibson 2003).

Finally, one can see that countries with a long history of autocratic rule, such as China and the Ukraine, have low levels of support, as existing research on authoritarian legacies would lead us to expect (Rose, Mishler, and Haerpfer 1998; Fuchs-Schündeln and Schündeln 2015). Moreover, in Egypt and Venezuela one can see public support for democracy reacting to political events, albeit in divergent ways. Venezuelan support for democracy steadily increased after Chavez began dismantling democratic checks and balances in 2000. In contrast, Egyptians reacted to the tumult of the Arab Spring in 2010 by turning away from democracy in the years that followed.

5.3 Item analysis

Finally, I examine the item parameters in more detail. Doing so will permit further discussion of the validity of the estimates. Moreover, the fact that item analysis is even possible illustrates another advantage of my modeling approach. I use estimates from model 6, which includes item slopes, for this section.

This analysis will focus on the item characteristic curves (ICCs), which are plotted in Figure 4. ICCs display the relationship between the proportion of a national sample responding supportively toward democracy (y-axis) and the latent estimates of support (x-axis). The vertical alignment of the curves is governed by the item intercepts λ, while the steepness of the curves is governed by the item slopes γ. To aid in interpretation, I group the items by their survey project, and use varying shades of gray and line types to identify the main question wording approach.

Turning to Figure 4, one can see that all 37 items display a positive relationship between the latent quantity and the observed responses. All items, in other words, have positive slopes. In addition, most items have slopes of similar magnitude. These are welcome findings, as they indicate that the included survey items do indeed measure the latent construct. It is not a particularly surprising finding, however, as items were selected based on the results of previous analyses of microlevel survey data (e.g., Rose, Mishler, and Haerpfer 1998; Klingemann 1999; Mattes and Bratton 2007). Items that bore some superficial resemblance to support for democracy, but which did not display a deeper empirical relationship with this latent variable were not included in the analysis in the first place.[14]

There are nonetheless a few items with weaker slopes and therefore more tenuous relationships with latent support for democracy. First is the "army rule" item from the New Democracies Barometer and second is the "evaluate democratic political system" item from the AsiaBarometer. Another three come from the World Values Surveys: the items asking respondents to rate the "importance of living in a democracy," and to evaluate a "strong leader" and a "democratic political system." This latter survey question has previously been criticized as offering only "lip service" to democracy, rather than deeply rooted support (Inglehart 2003). In two out of the three survey projects in which it is employed, this type of question does indeed show a weaker relationship with latent support for democracy.

In contrast, two widely used approaches for measuring support for democracy—the "three statements" and "evaluate army rule" survey questions—perform well across regions and survey projects. Both have pronounced positive slopes, indicating that such questions allow researchers to discriminate between respondents who favor democracy and those who do not. Indeed, national samples show widely varying levels of agreement with these items as their underlying support for democracy increases: at low levels of support for democracy (two standard deviations below the mean), 25% to 40% of respondents tend to offer the democratic responses to the three statements questions; at high levels (two standard deviations above the mean), around 85% do so.

14 The generally similar magnitudes of the item slopes also might explain why the inclusion of these parameters in Models 3 and 6 did not improve their accuracy.

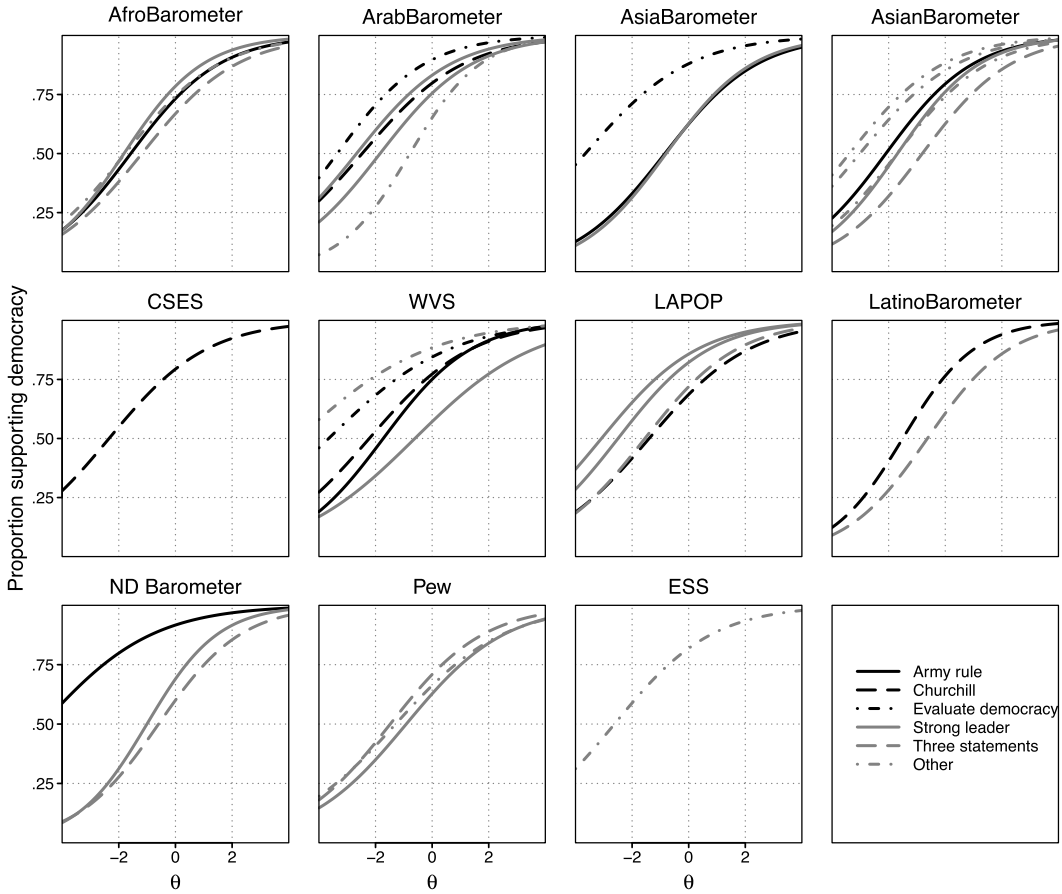

Figure 4. Item Characteristic Curves for All Items, Grouped By Project. Lines show the item characteristic curves (ICCs) for each item, with parameter estimates drawn from Model 6. ICCs depict the relationship between the proportion of a national sample responding supportively toward democracy (y-axis) and the latent estimates of support for democracy (θ, x-axis). The vertical alignment of the curves is governed by the item intercepts λ; the steepness of the curves is governed by the item slopes γ. The ICCs are grouped by survey project. The combination of line type and line shade indicates question theme. The legend is in the lower right corner. All parameters (θ, λ, and γ) are standardized such that θ has a unit-normal scale.

The main finding from this item analysis is that all included items show a marked, positive relationship between the latent variable and the observed responses. Indeed, most of the items show a strong relationship, indicating that they are sound measures of the latent construct of support for democracy. In addition, the ability to conduct such an item analysis illustrates another advantage of the proposed modeling framework.

6 Conclusion

Smooth panels of cross-national public opinion would be of great interest to scholars of comparative politics and comparative political economy. Yet assembling and estimating such panels is far from straightforward because public opinion data are fragmented over space and time and fractured across the numerous survey items that are used to measure any given opinion construct. To make matters worse, cross-national opinion data are gathered by a variety of survey projects using a variety of methodologies and in dozens of languages and countries, threatening their equivalence across countries.

This paper has proposed, developed, and validated a dynamic Bayesian latent variable framework for extracting smooth panels from such disparate cross-national opinion data. Despite the challenges of this task, the six models perform fairly well in tests of external validation.

The most accurate models predict aggregate survey responses that, on average, deviate by 6 percentage points from the observed percentages. These models also provide modestly accurate estimates of measurement uncertainty, with empirical CIC falling 20 percentage points short of the nominal level. I furthermore find that the estimated panel of opinions on support for democracy displays spatiotemporal patterns and associations with other variables that are consistent with previous research, suggesting both construct and convergent validity. Given the problems endemic in such data, I think that these results warrant optimism.

They also warrant further application, refinement, and testing. For scholars interested in applying these models to other contexts and to other opinions, I find, first, that a beta-binomial specification should be selected rather than the simpler binomial. Although the binomial provides out-of-sample predictions that are only slightly less accurate than the beta-binomial, the associated estimates of uncertainty are far too optimistic. Indeed, as the literature on survey error (e.g., Weisberg 2005) has cautioned—and recent election polling failures have demonstrated—public opinion data include numerous sources of error beyond that due only to sampling. Whether the beta-binomial more generally allows for better smoothing of opinion time series than the simpler binomial is a challenge for future research.

Second, I find that item slopes or factor loadings do not increase the accuracy of my models appreciably, but they do have a diagnostic utility, as demonstrated in the item analysis. Moreover, this particular analysis benefitted from a substantial literature which had already established the microlevel reliabilities of the included items. Item slopes would be a helpful model feature for analysts interested in estimating opinions for topics where such a literature is lacking.

Finally, adding item by country bias parameters increases the accuracy of both point and uncertainty estimates. This is hardly surprising because scholars have long warned of the dangers of assuming that a particular survey item operates in the same fashion across national contexts (e.g., Stegmueller 2011). However, analysts should note that it is only possible to include such item–country parameters when item–country replicates—particular items repeated in particular countries over time—are available.

Supplementary materials

For supplementary materials accompanying this paper, please visit https://doi.org/10.1017/pan.2018.32.

References

Almond, Gabriel A., and Sidney Verba. 1963. *The civic culture: Political attitudes and democracy in five nations*. Boston: Little & Brown.

Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13(2):171–187.

Baumgartner, Frank R., Suzanna L. De Boef, and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. New York: Cambridge University Press.

Beck, Nathaniel. 1989. Estimating dynamic models using Kalman filtering. *Political Analysis* 1:121–156.

Canache, Damarys, Jeffery J. Mondak, and Mitchell A. Seligson. 2001. Meaning and measurement in cross-national research on satisfaction with democracy. *Public Opinion Quarterly* 65(4):506–528.

Caughey, Devin, and Christopher Warshaw. 2015. Dynamic estimation of latent opinion using a hierarchical group-level IRT model. *Political Analysis* 23(2):197–211.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1):1–32.

Claassen, Christopher. 2018 Replication data for: Estimating smooth country-year panels of public opinion, https://doi.org/10.7910/DVN/A47LUM, Harvard Dataverse, V3, UNF:6:EbIp/yft64tlqvHHOIx3GA==.

Durbin, James, and Siem Jan Koopman. 2012. *Time series analysis by state space methods*. 2nd edn. Oxford: Oxford University Press.

Enns, Peter K. 2016. *Incarceration nation: How the United States became the most punitive democracy in the world*. New York: Cambridge University Press.

Erikson, Robert S., Michael MacKuen, and James A. Stimson. 2002. *The macro polity*. New York: Cambridge University Press.
Foa, Roberto Stefan, and Yascha Mounk. 2016. The democratic disconnect. *The Journal of Democracy* 27(3):5–17.
Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2015. On the endogeneity of political preferences: Evidence from individual experience with democracy. *Science* 347(6226):1145–1148.
Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
Gibson, James L. 2003. The legacy of apartheid: Racial differences in the legitimacy of democratic institutions and processes in the New South Africa. *Comparative Political Studies* 36(7):772–800.
Green, Donald P., Alan Gerber, and Suzanna L. De Boef. 1999. Tracking opinion over time: A method for reducing sampling error. *Public Opinion Quarterly* 63:178–192.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd edn. New York: Springer.
Hjort, Jonas. 2009. Pre-colonial culture, post-colonial economic success? The Tswana and the African Economic Miracle. *The Economic History Review* 63(3):688–709.
Inglehart, Ronald. 2003. How solid is mass support for democracy – and how can we measure it? *PS, Political Science and Politics* 36(1):51–57.
Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, cultural change, and democracy: The human development sequence*. New York: Cambridge University Press.
Jackman, Simon. 2005. Pooling the polls over an election campaign. *Australian Journal of Political Science* 40(4):499–517.
Klingemann, Hans-Dieter. 1999. Mapping political support in the 1990s: A global analysis. In *Critical citizens: Global support for democratic governance*, ed. Pippa Norris. Oxford: Oxford University Press.
Kurzman, Charles. 2014 World values lost in translation. *Washington Post*. September 2, online only. https://www.washingtonpost.com/news/monkey-cage/wp/2014/09/02/world-values-lost-in-translation/.
Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100(9):1989–2001.
Lindberg, Staffan I., Michael Coppedge, John Gerring, Jan Teorell, Daniel Pemstein, Eitan Tzelgov, Yi-ting Wang, Adam Glynn, David Altman, Michael Bernhard, Steven Fish, Alan Hicken, Matthew Kroenig, Kelly McMann, Pamela Paxton, Megan Reif, Svend-Erik Skaaning, and Jeffrey Staton. 2014. V-Dem: A new way to measure democracy. *Journal of Democracy* 25(3):159–169.
Linzer, Drew. 2013. Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association* 108(501):124–134.
Lipset, Seymour Martin. 1959. Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review* 53(1):69–105.
Mattes, Robert, and Michael Bratton. 2007. Learning about democracy in Africa: Awareness, performance, and experience. *American Journal of Political Science* 51(1):192–217.
McGann, Anthony J. 2014. Estimating the political center from aggregate data: An item response theory alternative to the Stimson Dyad ratios algorithm. *Political Analysis* 22(1):115–129.
McGraw, Kenneth O., and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1(1):30–46.
Norris, Pippa. 1999. Institutional explanations for political support. In *Critical citizens: Global support for democratic governance*, ed. Pippa Norris. Oxford: Oxford University Press.
Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis* 12(4):375–385.
Putnam, Robert D. 1993. *Making democracy work: Civic traditions in modern italy*. Princeton: Princeton University Press.
Rose, Richard, William Mishler, and Christian Haerpfer. 1998. *Democracy and its alternatives: Understanding post-communist societies*. Baltimore: Johns Hopkins University Press.
Skrondal, Anders, and Sophia Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: CRC Press.
Stan Development Team. 2017. *Stan modeling language: User's guide and reference manual: Stan 2.15.0.* Stan Development Team.
Stegmueller, Daniel. 2011. Apples and oranges? The problem of equivalence in comparative research. *Political Analysis* 19(4):471–487.
Stimson, James A. 1991. *Public opinion in America: Moods, cycles, and swings, transforming american politics*. Boulder, CO: Westview Press.
Stimson, James A., Michael B. Mackuen, and Robert S. Erikson. 1995. Dynamic representation. *American Political Science Review* 89(3):543–565.

Sullivan, John L., James E. Piereson, and George E. Marcus. 1982. *Political tolerance and american democracy*. Chicago: University of Chicago Press.
Vehtari, Aki, Andrew Gelman, and Johan Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5):1413–1432.
Voeten, Erik, and Paul R. Brewer. 2006. Public opinion, the war in Iraq, and presidential accountability. *Journal of Conflict Resolution* 50(6):809–830.
Weisberg, Herbert F. 2005. *The total survey error approach*. Chicago: University of Chicago Press.