Advance Access publication February 4, 2015

Political Analysis (2015) 23:197–211
doi:10.1093/pan/mpu021

Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model

Devin Caughey
Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4301, USA
e-mail: caughey@mit.edu (corresponding author)

Christopher Warshaw
Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4301, USA
e-mail: cwarshaw@mit.edu

Edited by Jonathan Katz

Over the past eight decades, millions of people have been surveyed on their political opinions. Until recently, however, polls rarely included enough questions in a given domain to apply scaling techniques such as IRT models at the individual level, preventing scholars from taking full advantage of historical survey data. To address this problem, we develop a Bayesian group-level IRT approach that models latent traits at the level of demographic and/or geographic groups rather than individuals. We use a hierarchical model to borrow strength cross-sectionally and dynamic linear models to do so across time. The group-level estimates can be weighted to generate estimates for geographic units. This framework opens up vast new areas of research on historical public opinion, especially at the subnational level. We illustrate this potential by estimating the average policy liberalism of citizens in each U.S. state in each year between 1972 and 2012.

1 Introduction

Since the advent of public opinion polling in the 1930s, millions of Americans—not to mention residents of other countries—have been surveyed on their political opinions, yielding a wealth of information on citizens' political attitudes over the past eight decades.[1] Scholars' ability to take full advantage of this information, however, has been hampered by two limitations of the data. First, until recently, each survey typically included only a handful of political questions, thus ruling out individual-level scaling techniques such as item response models. Second, few questions have been asked in comparable fashion across many polls, making it difficult to measure opinion change over time. This sparseness of the survey data has forced required scholars to restrict their focus either to the small number of academic surveys with many questions or to the few question series asked consistently over time. These difficulties are particularly acute when the target of inference is subnational opinion, for which small sample sizes present an additional problem.

Because of these challenges, scholars of subnational politics have often been forced to rely on measures that crudely proxy for their quantity of interest and that ignore variation over time and within subnational units. For example, the most widely used state-level measures of citizens' policy liberalism are based either on time-invariant measures of respondents' ideological self-identification

Authors' note: We are grateful to Kevin Quinn, Simon Jackman, and Teppei Yamamoto for their advice on the model derivation and validation, and to Bob Carpenter and Alex Storer for their assistance with coding and computation. We also received excellent feedback from Stephen Jessee, Bob Erikson, Mike Alvarez, John Jackson, and others at PolMeth 2013. Adam Berinsky, Eric Schickler, and Tom Clark were kind enough to share their data with us. We appreciate the research assistance of Stephen Brown, Justin de Benedictis-Kessner, and Melissa Meek. Supplementary materials for this article are available on the Political Analysis Web site.

[1] Replication materials for all of the results in this article are provided in the online dataverse archive associated with this article (Caughey and Warshaw 2014).

© The Author 2015. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

(Erikson, Wright, and McIver 1993) or on a weighted average of the scaled roll call votes of elected officials (Berry et al. 1998). The recent advent of online polls with large samples and many policy questions has made it possible to derive more direct and accurate measures of subnational policy liberalism (Tausanovitch and Warshaw 2013), but of course these measures cannot be extended back to earlier eras.

Our goal in this article is to introduce a measurement strategy that overcomes these obstacles. Our approach allows researchers to use survey indicators to estimate the distribution of latent traits such as policy liberalism at the subnational level, within a framework that simultaneously accounts for cross-sectional and over-time variation.

The model we develop builds upon several recent statistical advances in modeling public opinion. The first is item response theory (IRT), a statistical framework for dichotomous votes or survey responses that can be interpreted as an operationalization of the spatial utility model (Clinton, Jackman and Rivers 2004). We depart from the conventional IRT framework by modeling opinion not at the individual level, but rather at the level of subpopulation groups defined by demographic and geographic characteristics (Mislevy 1983; Lewis 2001). Second, we embed the group-level IRT model in a multilevel framework, modeling the group means hierarchically so as to borrow strength from demographically and geographically similar groups (Fox and Glas 2001). Third, to accommodate opinion change over time, we allow the hierarchical parameters to evolve according to a dynamic linear model, thus borrowing strength across time as well (Martin and Quinn 2002). Finally, the time-specific estimates of average group opinion may then be weighted and aggregated to produce dynamic opinion estimates for states or other geographic units (Park, Gelman, and Bafumi 2004).

Our approach has substantial advantages over existing methods. It shares with individual-level IRT models the benefits of using many indicators for a given construct, but a group-level approach is much more computationally efficient. In addition, because a given respondent need answer only a single question, it opens up vast amounts of opinion data to IRT modeling. Our model's dynamic structure naturally accommodates opinion change over time, a central focus of public opinion research (e.g., Stimson 1991). Unlike Stimson's "mood" algorithm, however, our approach is derived from an explicit individual-level model that accommodates cross-sectional and over-time differences as well as sampling variability in a unified framework (cf. Enns and Koch 2013; McGann 2014).

This modeling framework is quite general and can be applied to a wide variety of time-varying latent constructs. It is most valuable when computational or data limitations make an individual-level model difficult or impossible to estimate. We demonstrate the usefulness of our framework in Section 4, using it to estimate the average policy liberalism of U.S. state publics in each year between 1972 and 2012. The Supplementary Materials contain additional applications to state-level support for the New Deal in the 1936–52 period and state-level confidence in the U.S. Supreme Court in 1965–2010. Many other applications are also possible, including modeling cross-national opinion dynamics, inferring voter preferences from electoral data, and estimating the average ideal points of party caucuses in Congress or state legislatures.

This article proceeds as follows. We first review existing approaches to modeling public opinion, with a particular focus on the problem of estimating latent quantities from sparse survey data. Next, we derive and explain our dynamic group-level IRT model of latent opinion. We then illustrate and validate our model using the example of policy liberalism in U.S. states between 1972 and 2012. The final section concludes.

2 Modeling Latent Quantities in the Mass Public

Many of the most important constructs in public opinion research are latent quantities. Policy liberalism, political knowledge, racial resentment, trust in government—none of these traits can be measured well with a single question. Rather, scholars must infer individuals' latent scores from multiple indicators, typically with the aid of latent-variable models such as factor analysis or IRT models. The accuracy and precision of these inferences generally improves as the number of indicators increases (e.g., Ansolabehere, Rodden, and Snyder 2008).

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

The requirement that multiple indicators be observed radically restricts the proportion of available survey data amenable to individual-level scaling. For example, since the advent of mass opinion polling in the 1930s, millions of American citizens have been asked hundreds of different policy questions. But the vast majority of polls pose just one or two policy questions to each respondent, forcing scholars interested in citizens' policy liberalism to rely on a small number of academic surveys that contain more than a few questions per respondent.

The problem is even more acute for scholars interested in the geographic distribution of policy liberalism, because the sample sizes of surveys like the American National Election Studies (ANES) are too small for reliable subnational inferences.[2] Only since the 2000s have scholars had access to the Cooperative Congressional Election Study (CCES) and other surveys with both large subnational samples and multiple policy questions per respondent. With these data, political scientists have used Bayesian IRT models to estimate the subnational distribution of liberalism and thus to examine important topics such as spatial voting and representation (Jessee 2009; Bafumi and Herron 2010).

Even polls with sample sizes as large as the CCES's can contain small or even empty samples for at least some subnational units. In such cases, opinion estimates for subnational units can be improved through the use of multilevel regression and poststratification (MRP; Park, Gelman, and Bafumi 2004). The idea of MRP is to first derive Bayesian model-based estimates of opinion in subpopulation groups defined by geographic and demographic characteristics. These estimates are then poststratified in proportion to the groups' composition of the population. Because the shrinkage estimates "borrow strength" from groups with similar characteristics, MRP is typically more accurate than aggregating raw samples (Lax and Phillips 2009; Warshaw and Rodden 2012; though see Buttice and Highton 2013). Tausanovitch and Warshaw (2013) demonstrate that IRT modeling and MRP can be combined to estimate the distribution of latent variables such as policy liberalism at the subnational level. But like all individual-level IRT models, their approach requires many issue questions per respondent, so it cannot be extended back in time.

Scholars interested in opinion change over time have been constrained to examine the nation as a whole, typically ignoring individual-level variation entirely. Stimson (1991), for example, uses the factor-analytic Dyad Ratios algorithm to estimated changes in the aggregate "policy mood" of the United States. Enns and Koch (2013) combine this approach with MRP to generate state-level estimates of policy mood. As McGann (2014) observes, however, an unappealing feature of the Dyad Ratios algorithm is its lack of grounding in a coherent individual-level model. As an alternative, he proposes a group-level IRT model for national mood that is similar to the approach we take in Section 3.[3] Unlike the Dyad Ratios algorithm, McGann's model has clear individual-level microfoundations, though like most studies of mood, he uses it to model only over-time changes in national opinion.

Finally, several works have exploited the time-series structure of their data to borrow strength across time using a dynamic linear model (DLM), the Bayesian analogue to the frequentist Kalman filter (Jackman 2009, 471–72). Martin and Quinn (2002), for instance, use a "random walk" DLM to smooth Supreme Court ideal points across sessions. Jackman (2005) and Linzer (2013) use somewhat more complex DLMs to model opinion over the course of a presidential campaign. Rather than interpreting DLMs as structural models whose parameters are of interest in themselves, these Bayesian applications use them as convenient means of smoothing estimates over time.

Our model, described in the following section, incorporates elements of several of the approaches outlined above. We relate observed responses to the latent trait using an IRT model, which we estimate at the group rather than individual level to avoid the problem of too few questions per respondent. To improve the accuracy of the estimates of group means, we borrow strength from similar groups via a hierarchical model. To borrow strength across time as well, we model the evolution of the hierarchical parameters using DLMs. As in MRP, we derive estimates for

[2]Moreover, the ANES samples are not representative of state populations due to their multi-stage area sampling design.
[3]We derived our model independently of McGann (2014), building instead on Mislevy's (1983) derivation of IRT models for grouped data.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

geographic aggregates by poststratifying the estimated group means to match the population distribution. All this is done in a Bayesian estimation framework.

3 A Dynamic Hierarchical Group-Level IRT Model

In this section, we describe our dynamic measurement model. Our aim is to use data from a large number of polls, each including as few as one survey question, to make inferences about opinion in demographically and/or geographically defined groups at a given point in time. The group estimates may be of interest in themselves, or their weighted average may be used to estimate opinion in states or other geographic units. To understand the logic of the model, it is helpful to derive it step by step, beginning with the group-level IRT model.

3.1 *Group-Level IRT Model*

In the two-parameter IRT model, each response $y_{ij} \in \{0, 1\}$ is a function of subject i's score on the latent trait (θ_i), item j's *difficulty* (α_j) and *discrimination* (β_j), and an error term (e_{ij}). The response $y_{ij} = 1$ if and only if $\beta_j\theta_i - \alpha_j + \epsilon_{ij} > 0$. If ϵ_{ij} is assumed to be i.i.d. standard normal, then the probability of answering correctly is given by the normal ogive IRT model,

$$\Pr[y_{ij} = 1] = p_{ij} = \Phi(\beta_j\theta_i - \alpha_j), \tag{1}$$

where Φ is the standard normal CDF (Fox 2010, 10).

Accurate estimation of θ_i requires data on many subjects, each of whom answers many items (Lewis 2001, 277). Unfortunately, only a small minority of public opinion surveys contain enough items to make estimation of θ_i remotely plausible. As Bailey (2001), Lewis (2001), and others have noted, however, it is often possible to make inferences about the distribution of θ_i even when individual-level estimation is not feasible. We rely particularly on Mislevy (1983), who derives group-level representations of various IRT models that permit group means to be estimated even if each individual answers only a single question. The essential idea is to model the θ_i in group g as distributed normally around the group mean $\overline{\theta}_g$ and then marginalize over the distribution of θ_i.

To derive the group-level representation of the normal ogive model, it is helpful to reparameterize it as

$$p_{ij} = \Phi[(\theta_i - \kappa_j)/\sigma_j], \tag{2}$$

where $\kappa_j = \alpha_j/\beta_j$ and $\sigma_j = \beta_j^{-1}$ (Fox 2010, 11). The item *threshold* κ_j is the value of θ_i at which a subject has a 50% probability of answering item j correctly.[4] The *dispersion* σ_j represents the magnitude of the measurement error for item j. Given the normal ogive IRT model and the assumption that θ_i is normally distributed within subpopulation groups, the probability that a randomly sampled member of group g correctly answers item j is

$$p_{gj} = \Phi[(\overline{\theta}_g - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}], \tag{3}$$

where $\overline{\theta}_g$ is the mean of θ_i in group g, and σ_θ is the within-group standard deviation (SD) of θ_i (Mislevy 1983, 278). See Supplementary Materials for a formal derivation of equation (3).

Rather than modeling the individual responses y_{ij}, as in a typical IRT model, we instead model $s_{gj} = \sum_i^{n_{gj}} y_{i[g]j}$, the total number of correct answers to item j out of the n_{gj} responses of subjects in group g. Assuming that responses are independent conditional on θ_i, κ_j, and σ_j and that each subject answers one question, $s_{gj} \sim \text{Binomial}(n_{gj}, p_{gj})$. In Section 3.4, we relax the assumption of one item per subject.

[4] In terms of a spatial model, κ_j is the *cutpoint*, or point of indifference between two choices.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

3.2 *Hierarchical Model for Group Means*

The number of groups whose opinion can be estimated using the group-level IRT model in equation (3) is very limited due to the sparseness of survey data, which leads to unstable or even undefined group-level estimates. In fact, the only previous political science application of a group-level IRT model, McGann (2014), considers only a single group (the British public) and is based entirely on aggregate rather than individual-level data. In addition to precluding subnational opinion estimation, modeling a single group means that the model considers only over-time opinion variation and ignores cross-sectional variation, which tends to be much larger.

A natural way to deal with the sparseness problem is to smooth the group-level estimates by modeling them hierarchically using a multilevel model. The hierarchical model for the group means can be written as

$$\overline{\theta}_g \sim \mathrm{N}(\xi + \boldsymbol{x}'_{g.}\boldsymbol{\gamma},\ \sigma^2_{\overline{\theta}}), \tag{4}$$

where ξ is an intercept term, $\boldsymbol{x}_{g.}$ is a vector of P observed attributes of group g, and $\boldsymbol{\gamma}$ is a P-vector of hierarchical coefficients. If we let it vary by time period, the intercept ξ_t captures opinion dynamics common to all units. The vector $\boldsymbol{x}_{g.}$ may include geographic identifiers, demographic predictors, or interactions thereof. For example, if groups are defined by the intersection of *State*, *Race*, and *Gender*, the groups' means could be modeled as an additive function of intercepts for each state as well as for each race and gender category.

To the extent that there are many members of group g in the data, the estimate of $\overline{\theta}_g$ will be dominated by the likelihood. In the opposite case of an empty cell, $\overline{\theta}_g$ will be shrunk all the way to the linear predictor $\boldsymbol{x}_{g.}'\boldsymbol{\gamma}$. In this sense, the hierarchical model imputes responses for groups for which data are missing, either because they were not sampled or because their responses were coded as NA (e.g., "don't knows"). The model thus automatically generates estimates for all groups, even those with no observed respondents.

3.3 *Dynamic Linear Model for Hierarchical Parameters*

To estimate opinion change across time, we could simply estimate equation (4) anew in each time period. This is essentially what Enns and Koch (2013) do to generate the state-specific question marginals they feed into the Dyad Ratios algorithm (which then smooths the data over time). Such an approach is not feasible for an IRT model, which would perform very poorly in years with few questions and force years with no data to be dropped from the analysis. At the other extreme, we could constrain the hierarchical coefficients in equation (4) to be constant across time. This approach too is unappealing, especially for long time series, because it is insensitive to changes in the predictiveness of group characteristics (e.g., the emergence of a gender gap in opinion over time).

A third alternative is to smooth the hierarchical coefficients across time, thus striking a balance between the strong assumption that the model is constant over time and the equally strong assumption that opinion is independent across periods (Martin and Quinn 2002, 140). DLMs provide a natural way to borrow strength across time. Smoothing latent-variable models reduces temporal variability, leading to more efficient estimates and generally improving the predictive performance of the model (Armstrong et al. 2014, 303–4).

For the purposes of pooling information across time, it is usually sufficient to use a simple *local-level* DLM, which treats a parameter's value in one period as its expected value in the next (a so-called "random walk" prior). We use the following local-level transition model for the intercept ξ_t:

$$\xi_t \sim \mathrm{N}(\xi_{t-1},\ \sigma^2_\gamma), \tag{5}$$

where the innovation variance σ^2_γ determines the weight of the data in period t relative to $t-1$.[5] If there are no new data in period t, then equation (5) acts as a predictive model, imputing an

[5]Unlike many applications (e.g., Martin and Quinn 2002), which treat the innovation variance as a tuning parameter, the innovation variances in our model are estimated from the data.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

estimated value for ξ_t (Jackman 2009, 474). We use an analogous local-level DLM to model the evolution of hierarchical coefficients that correspond to *demographic* predictors such as gender or race:

$$\gamma_{pt} \sim \mathrm{N}(\gamma_{p,t-1},\ \sigma^2_\gamma). \tag{6}$$

For coefficients corresponding to *geographic* predictors such as state, we write the model more generally so as to permit the optional inclusion of geography-level attributes. Including geographic covariates pools information cross-sectionally among similar geographic units, which can improve the efficiency of geographic effect estimates (Park, Gelman, and Bafumi 2004). The transition equation for geographic intercepts is

$$\gamma_{pt} \sim \mathrm{N}(\gamma_{p,t-1}\delta_t + \boldsymbol{z}'_{p.,t}\boldsymbol{\eta}_t,\ \sigma^2_\gamma), \tag{7}$$

where δ_t is a scalar lag coefficient, $\boldsymbol{z}_{p.,t}$ is a vector of geography-level attributes (e.g., *Per Capita Income*), and $\boldsymbol{\eta}_t$ is a vector of coefficients (Jackman 2009, 471–72). Geographic intercepts are thus modeled as a weighted combination of their value in the previous period ($\gamma_{p,t-1}$) and the attributes contained in $\boldsymbol{z}_{p.,t}$, with weights δ_t and $\boldsymbol{\eta}_t$, respectively. To the extent that geographic effects are temporally stable, the over-time pooling represented by δ_t will tend to dominate the cross-sectional pooling captured by $\boldsymbol{\eta}_t$.

We use local-level transition models for all other time-varying parameters: the coefficients in equation (7) (δ_t and $\boldsymbol{\eta}_t$), the SD of θ_i within groups ($\sigma_{\theta,t}$), and the residual SD of group means ($\sigma_{\bar{\theta},t}$). We model the SDs on the log scale, as in

$$\sigma_{\theta,t} \sim \mathrm{LN}(\log(\sigma_{\theta,t-1}),\ \sigma^2_\sigma), \tag{8}$$

where LN indicates the lognormal distribution and σ^2_σ is a hyperparameter to be estimated.

3.4 *Respondent Weights*

Before we present the complete model, we add one further extension, which is to allow for respondent-level weights. Weights may be required for two reasons. The first is to adjust for unequal response probabilities within groups. Many surveys include such weights, derived from known sampling probabilities and/or from a post hoc weighting method such as poststratification. Second, weights may also be used to account for multiple responses per survey respondent. If not accounted for, such respondent-level clustering leads to underestimates of the uncertainty surrounding the group means.

We deal with both kinds of weights using the following procedure, which builds on that proposed by Ghitza and Gelman (2013). We first estimate a "design effect" d_{gt} for each group g and period t,

$$d_{gt} = 1 + \left(\frac{\mathrm{sd}_{gt}(w_{i[gt]})}{\mathrm{ave}_{gt}(w_{i[gt]})}\right)^2, \tag{9}$$

where $w_{i[gt]}$ is the survey weight of individual i, and the average and SD are taken across respondents in group g in period t. Second, we calculate an adjusted sample size $n^*_{gjt} \leq n_{gjt}$, using the formula

$$n^*_{gjt} = \lceil \sum_{i=1}^{n_{gjt}} \frac{1}{r_{i[gt]} d_{gt}} \rceil, \tag{10}$$

where $r_{i[gt]}$ is the number of questions respondent i answered and $\lceil \cdot \rceil$ is the ceiling function.[6] If each individual in group g answers one question ($r_{i[gt]} = 1, \forall i$) and there is no within-group variation in

[6]We round to conform to the binomial probability distribution, and use the ceiling function to avoid a sample size of 0. Ghitza and Gelman (2013) do not round because their non-Bayesian approach allows for quasi-likelihood functions such as non-integer binomials.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

weights ($d_{gt} = 1$), then $n^*_{gjt} = \sum_{i=1}^{n_{gjt}} 1 = n_{gjt}$. But if subjects answer multiple items or there is within-group variation in survey weights, then (rounding aside) $n^*_{gjt} < n_{gjt}$. Third, we take the weighted mean of each group's responses to item j:

$$\overline{y}^*_{gjt} = \sum_{i=1}^{n_{gjt}} \frac{w_{i[gt]} y_{i[g]jt}}{r_{i[gt]}} \Big/ \sum_{i=1}^{n_{gjt}} \frac{w_{i[gt]}}{r_{i[gt]}}. \tag{11}$$

Finally, we replace the raw sums s_{gjt} with the weighted sums $s^*_{gjt} = \lceil n^*_{gjt} \overline{y}^*_{gjt} \rfloor$, where $\lceil \cdot \rfloor$ is the nearest integer function.

As equation (11) shows, each response $y_{i[g]jt}$ is weighted by i's survey weight ($w_{i[gt]}$) divided by the number of questions i answered ($r_{i[gt]}$). As a consequence, i's total weight across all $r_{i[gt]}$ items i answered is $r_{i[gt]} \times \frac{w_{i[gt]}}{r_{i[gt]}} = w_{i[gt]}$. In other words, each respondent's total contribution to the estimate of $\overline{\theta}_{gt}$ is determined by their survey weight, not by how many questions they answered. Further, group g's total sample size across all items j in period t is $\sum_i w_{i[gt]}$, the weighted sum of the period-specific number of respondents in the group.

3.5 *The Full Model*

We are now in a position to write down the complete model. Adding the indexing by t, the group-level IRT model is

$$s^*_{gjt} \sim \text{Binomial}(n^*_{gjt}, p_{gjt}), \tag{12}$$

where

$$p_{gjt} = \Phi[(\overline{\theta}_{gt} - \kappa_j)/\sqrt{\sigma^2_{\theta,t} + \sigma^2_j}]. \tag{13}$$

The time-indexed hierarchical model for the vector of group means is

$$\overline{\theta}_{gt} \sim \text{N}(\xi_t + \boldsymbol{x}'_{g,,t}\gamma_t, \sigma^2_{\overline{\theta},t}). \tag{14}$$

The only parameters in the model that are not indexed by t are the item parameters κ_j and σ_j, which are constrained to be constant across time. Substantively, this corresponds to the requirement that the item characteristic curves mapping item responses to the latent θ space do not change over time. This constraint bridges the model across time, putting the $\overline{\theta}_{gt}$ from different periods on a common metric. In many contexts, however, it may make sense to allow item parameters to evolve over time, a possibility explored in the Supplementary Materials.

3.6 *Identification, Priors, and Estimation*

One-dimensional IRT models must be identified by fixing the direction, location, and scale of the latent dimension (Clinton, Jackman, and Rivers 2004). We fix the direction of the metric by coding all question responses to have the same polarity (e.g., 1 indicates the liberal response) and restricting the discrimination parameters β_j to be positive for all items. We identify the location and scale by rescaling the item parameters α_j and β_j (Fox 2010, 88–89). In each iteration m, we set the location by transforming the J difficulties to have a mean of 0: $\tilde{\alpha}^{(m)}_j = \alpha^{(m)}_j - J^{-1}\sum_{j=1}^{J} \alpha^{(m)}_j$. Similarly, we set the scale by transforming the discriminations to have a product of 1: $\tilde{\beta}^{(m)}_j = \beta^{(m)}_j (\prod_{j=1} \beta^{(m)}_j)^{-1/J}$. The transformed parameters $\tilde{\alpha}_j$ and $\tilde{\beta}_j$ are then re-parameterized as κ_j and σ_j, which enter into the group-level response model in equation (3). For most parameters, we employ weakly informative priors that are proper but provide relatively little

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

information.[7] We estimated the model using the program `Stan`, as called from `R` (Stan Development Team 2013; R Core Team 2013).[8] Supplementary Materials include the `Stan` code used to estimate the model.

3.7 *Weighting Group Means to Estimate Geographic Opinion*

The estimates of the yearly group means $\overline{\theta}_{gt}$ may be of interest in themselves, but they are also useful as building blocks for estimating opinion in geographic aggregates (Park, Gelman, and Bafumi 2004). A major advantage of simulation-based estimation is that it facilitates proper accounting of the uncertainty surrounding functions of the estimated parameters. For example, the estimated mean opinion in a given state is a weighted average of mean opinion in each demographic group, which is itself an estimate subject to uncertainty. The uncertainty in the group estimates can be appropriately propagated to the state estimates via the distribution of state estimates across simulation iterations. Posterior beliefs about average opinion in the state can then be summarized via the means, SDs, and so on of the posterior distribution. We adopt this approach in presenting the results of the model in the application that follows.

4 Application and Validation: U.S. Policy Liberalism, 1972–2012

Having derived and explained our model, we now turn to demonstrating its usefulness and validity. In this application, we use our model to estimate state domestic policy liberalism in each year between 1972 and 2012. This quantity of interest is very similar to the concept of "public policy mood" modeled by Stimson (1991), Enns and Koch (2013), and McGann (2014), among others. The primary difference is that mood is a relative concept—should the government be doing "more" or "less" *than it currently is* (e.g., Stimson 2012, 31). By contrast, we conceive of policy liberalism as a construct that can be compared in absolute terms over time, independent of the policy status quo. The main practical consequence of this definitional distinction is that we include only data based on questions that refer to specific policy outcomes (e.g., Should the government guarantee health care to all citizens?) rather than policy changes (e.g., Should access to government-provided health care be expanded?). Given different data, however, our model could also be used to estimate policy mood.

Policy liberalism is also related to ideological identification, which is typically measured with a categorical question asking respondents to identify themselves as "liberal," "moderate," or "conservative." Because they have been asked in standardized form in a very large number of polls, ideological identification questions have been the most widely used survey-based measures of state liberalism (Erikson, Wright, and McIver 1993, 2006). While an important construct in its own right, "symbolic" ideological identification is conceptually and empirically distinct from "operational" ideology expressed in the form of policy preferences (Ellis and Stimson 2012). That our model generates dynamic survey-based estimates of policy liberalism is thus an important advance over existing approaches.

Our data for this application consist of survey responses to forty-seven domestic policy questions spread across 350 public opinion surveys fielded between 1972 and 2012. The questions cover traditional economic issues such as taxes, social welfare, and labor regulation, as well as topics

[7] The first-period priors for all SD parameters are half-Cauchy with a mean of 0 and a scale of 2.5 (Gelman 2007). The difficulty and discrimination parameters are drawn, respectively, from N(0, 1) and LN(0, 1) prior distributions and then transformed as described above. All coefficients not modeled hierarchically are drawn from distributions centered at 0 with an estimated SD, except $\delta_{t=1}$ and $\boldsymbol{\eta}_{t=1}$, which are modeled more informatively as N(0.5, 1) and N(0, 10), respectively. Note, however, that δ_t does not enter into the model until $t = 2$ (when the first lag becomes available), and thus its value in $t = 1$ serves only as a starting point for its dynamic evolution between the first and second periods.

[8] Stan is a `C++` library that implements the No-U-Turn sampler (Hoffman and Gelman forthcoming), a variant of Hamiltonian Monte Carlo that estimates complicated hierarchical Bayesian models more efficiently than alternatives such as `BUGS`. In general, four thousand iterations (the first two thousand used for adaptation) in each of ten parallel chains proved sufficient to obtain satisfactory samples from the posterior distribution. Computation time depends on the number of groups, items, and time periods; run times for the models reported in this article ranged between a day and several weeks.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

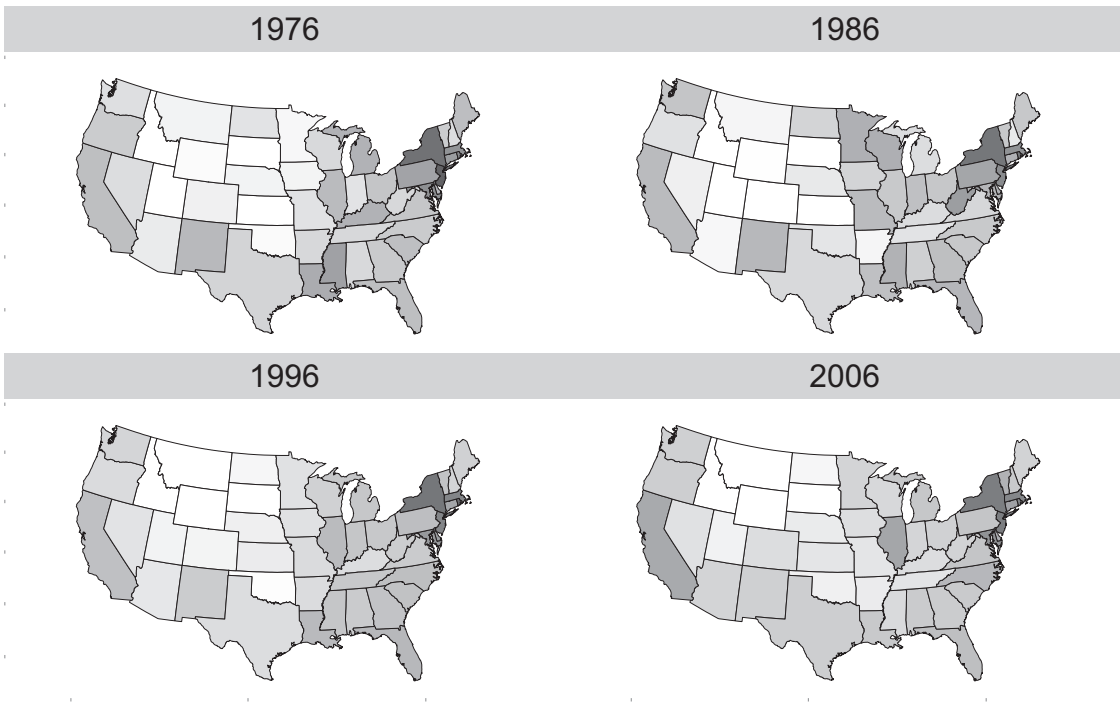


Fig. 1 Average citizen policy liberalism by state, 1976–2006. Darker shade indicates more liberal. The estimates have been centered and standardized in each year to accentuate the grayscale contrasts.

like gun control, immigration, and environmental protection. For conceptual clarity and comparability with policy mood, this application includes only questions for which the "liberal" answer involved greater government spending or activity.[9] The responses of over 570,000 different Americans are represented in the data.

We model opinion in groups defined by states and a set of demographic categories (e.g., race and gender). To mitigate sampling error for small states, we model the state effects in the first time period as a function of state *Proportion Evangelical/Mormon*. The inclusion of state attributes in the model partially pools information across similar geographical units in the first time period, improving the efficiency of state estimates (Park, Gelman, and Bafumi 2004). We drop *Proportion Evangelical/Mormon* after the first period because we found that the state intercept in the previous period tends to be much more predictive than state attributes.

To generate annual estimates of average opinion in each state, we weighted the group estimates to match the groups' proportions in the state population, based on data from the U.S. Census (Ruggles et al. 2010). Figure 1 maps our estimates of state policy liberalism in 1976, 1986, 1996, and 2006. The cross-sectional patterns are generally quite sensible—the most conservative states are in the Great Plains, while New York, California, and Massachusetts are always among the most liberal states. Moreover, Fig. 1 confirms that the states have remained generally stable in their relative liberalism, consistent with Erikson, Wright, and McIver's (2006) finding that state publics have been stable in terms of ideological identification. According to our estimates, only a few states' policy liberalism have shifted substantially over time. Southern states such as Mississippi and Alabama have become somewhat more conservative over time, while states in New England have become somewhat more liberal.

[9]For example, questions about restricting access to abortion were not included. Stimson (1999, 89–91) notes that the temporal dynamics of abortion attitudes are distinct from other issues, at least before 1990.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

4.1 *Cross-Validation*

The use of multilevel modeling to smooth subnational opinion estimates across cross-sectional units has been well validated (Lax and Phillips 2009; Warshaw and Rodden 2012; Tausanovitch and Warshaw 2013). A more innovative aspect of our model is the DLM for the parameters of the hierarchical model, which pools information across time in addition to cross-sectionally. Although a number of political science works have employed similar temporal smoothing methods (e.g., Martin and Quinn 2002; Jackman 2005; Park 2012; Linzer 2013; Wawro and Katznelson 2013), their application to dynamic public opinion has not been validated as extensively as multilevel modeling has. One noteworthy potential concern about our approach to dynamics is that even though the $\overline{\theta}_{gt}$ are re-estimated in each period, smoothing the hierarchical coefficients across periods dampens the estimates' sensitivity to rapid opinion changes (e.g., a sharp conservative turn in a specific state), especially in years when the data are thin.

To investigate this possibility, we designed a cross-validation study that compared the performance of our approach (the *pooled* model) to one in which the intercept and coefficients of the hierarchical model are estimated separately in each period (the *separated* model).[10] Specifically, we took a validation set approach (James et al. 2013, 176–8) in which 25% of respondents in each group-year were sampled to create a training data set.[11] We used the training data to estimate both the pooled model and the separated model. Based on the parameter estimates from each model, we calculated the predicted proportion of liberal responses to each item in each group-year:

$$\hat{p}_{gjt} = \Phi[(\hat{\overline{\theta}}_{gt} - \hat{\kappa}_j)/\sqrt{\hat{\sigma}^2_{\theta,t} + \hat{\sigma}^2_j}]. \tag{15}$$

To evaluate the out-of-sample performance of each model, we compared each predicted proportion with the proportion of liberal responses in the other 75% of the data, generating the prediction error for each of the N item–group–year triads:

$$\hat{e}_{gjt} = \frac{s^*_{gjt}}{n^*_{gjt}} - \hat{p}_{gjt}. \tag{16}$$

We contrasted the two models in terms of three metrics: bias ($N^{-1}\sum \hat{e}_{gjt}$), mean absolute error ($N^{-1}\sum |\hat{e}_{gjt}|$), and root-mean-square error ($\sqrt{N^{-1}\sum \hat{e}^2_{gjt}}$). We replicated the whole process ten times, thus producing ten out-of-sample estimates of bias, mean absolute error (MAE), and root-mean-square error (RMSE) for each model.

As Table 1 indicates, the pooled model is clearly superior to the separated model in terms of bias, MAE, and RMSE. Though the differences (expressed in percentage points) are not large, the pooled model strictly dominates the separated in every replication but one. The improvement in efficiency is to be expected given that the pooled model borrows strength from adjacent periods. That the pooled model exhibits less bias—in fact, is nearly unbiased when averaged across replications, in contrast to the liberal bias of the separated model—is perhaps more surprising, given that Bayesian smoothing shrinks estimates away from the (unbiased) maximum likelihood estimate. The explanation is that the coefficient estimates in the separated model are shrunk as well, but toward the cross-sectional mean rather than toward their value in the previous period.

In summary, the cross-validation results corroborate the value of pooling the hierarchical coefficients over time via a dynamic linear model. Temporal smoothing results not only in greater efficiency but also in less bias than estimating the hierarchical model separately by period, at least in this application. Thus, for the general purpose of measuring opinion over time, pooling

[10] To keep the comparison transparent and minimize computation time (which was still very lengthy), we defined groups by state only, with no demographic covariates. We also restricted the time period covered to 1976–2010.

[11] We sampled 25% rather than splitting the sample equally because we wanted to compare the models' performance when data are relatively sparse, and second to leave enough out-of-sample data to generate precise estimates of bias, MAE, and RMSE.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

Table 1 Out-of-sample bias, MAE, and RMSE of the separated and pooled models across 10 cross-validation replications

	Separated model			*Pooled model*			*Difference in magnitude*		
Rep.	*Bias*	*MAE*	*RMSE*	*Bias*	*MAE*	*RMSE*	*Bias*	*MAE*	*RMSE*
1	0.50	13.37	18.71	0.21	13.08	18.36	0.29	0.29	0.35
2	0.43	13.31	18.58	0.11	13.02	18.23	0.31	0.29	0.35
3	0.26	13.46	18.80	0.00	13.17	18.44	0.26	0.29	0.36
4	0.13	13.44	18.76	−0.26	13.13	18.40	−0.13	0.31	0.36
5	0.53	13.35	18.64	0.24	13.07	18.32	0.28	0.29	0.32
6	0.36	13.37	18.79	0.01	13.11	18.46	0.35	0.26	0.33
7	0.22	13.50	18.86	−0.10	13.20	18.49	0.12	0.30	0.37
8	0.50	13.43	18.76	0.28	13.17	18.44	0.22	0.26	0.32
9	0.15	13.40	18.78	−0.14	13.11	18.42	0.01	0.30	0.35
10	0.42	13.36	18.72	0.21	13.06	18.37	0.21	0.30	0.35
Mean	0.35	13.40	18.74	0.06	13.11	18.39	0.19	0.29	0.35

Note. The rightmost panel reports the difference in magnitude between the models (e.g., $|\text{Bias}_{separated}| - |\text{Bias}_{pooled}|$). All values are expressed in terms of percentage points.

appears to be the better choice. Nevertheless, the separated model may be preferable in certain circumstances, such as when one wishes to estimate abrupt opinion changes within a demographic group or geographic unit.

4.2 *Construct Validation*

The split-sample validation approach shows that our pooled model dominates a separated model where the intercept and coefficients of the hierarchical model are estimated separately in each period. However, it speaks only partially to the ability of our model to accurately estimate state- and national-level policy liberalism. To further assess our estimates' validity as a measure of policy liberalism, we examine their correlation with measures of several theoretically related constructs (a procedure (Adcock and Collier 2001) refer to as "construct validation").

First, we examine the cross-sectional correlation between our measure of policy liberalism and Democrats' presidential vote share. While presidential election results are not a perfect measure of citizens' policy preferences (Levendusky, Pope, and Jackman 2008; Kernell 2009), a variety of previous scholars have used presidential election returns to estimate state and district preferences. Thus, to the extent that policy attitudes predict presidential partisanship, a high correlation with Democratic presidential vote share would suggest that our estimates are accurate measures of states' policy preferences. Figure 2 shows that there is indeed a strong cross-sectional relationship between our estimates of state policy liberalism and presidential vote share between 1972 and 2012.[12] Moreover, the relationship increases in strength over time, mirroring the growing alignment of policy preferences with partisanship and presidential voting at the individual level (Fiorina and Abrams 2008, 577–82).

While the strong relationship with presidential vote share demonstrates the cross-sectional validity of our measure, it does not provide information about the ability of our model to detect *changes* in the mass public's preferences over time. Presidential votes are ill suited for this task since partisan vote shares ebb and flow for reasons unrelated to changes in the policy liberalism of the American public. For instance, parties could nominate a low-valence candidate, or there could be an incumbency advantage for presidents running for a second term. To validate the over-time validity of our estimates, we compare it to the dynamics of policy mood, which is explicitly

[12]We find a similarly strong relationship between our estimates of state policy liberalism and estimates of state ideology from exit polls.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

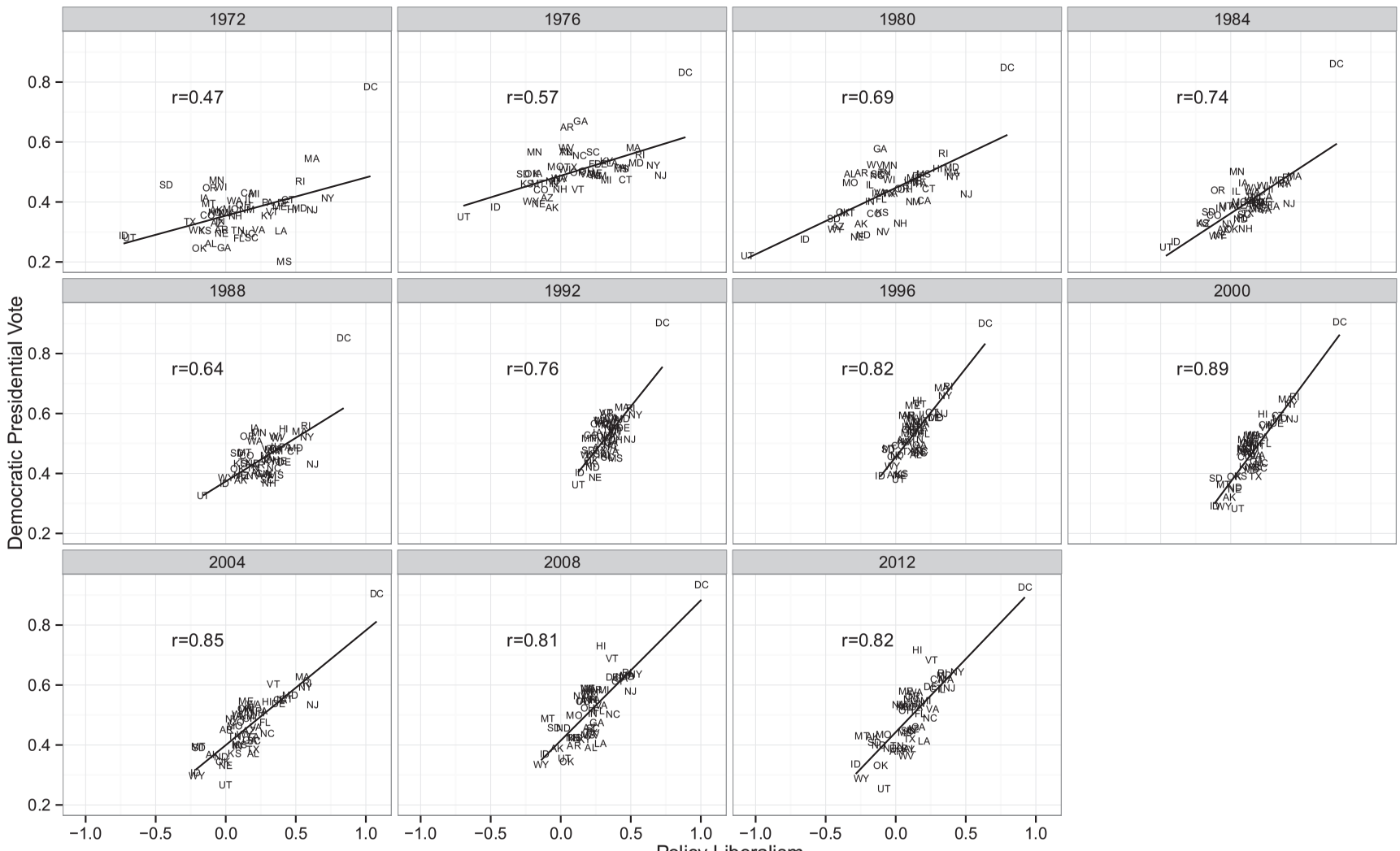


Fig. 2 Relationship between policy liberalism and Democratic presidential vote share, 1972–2012.

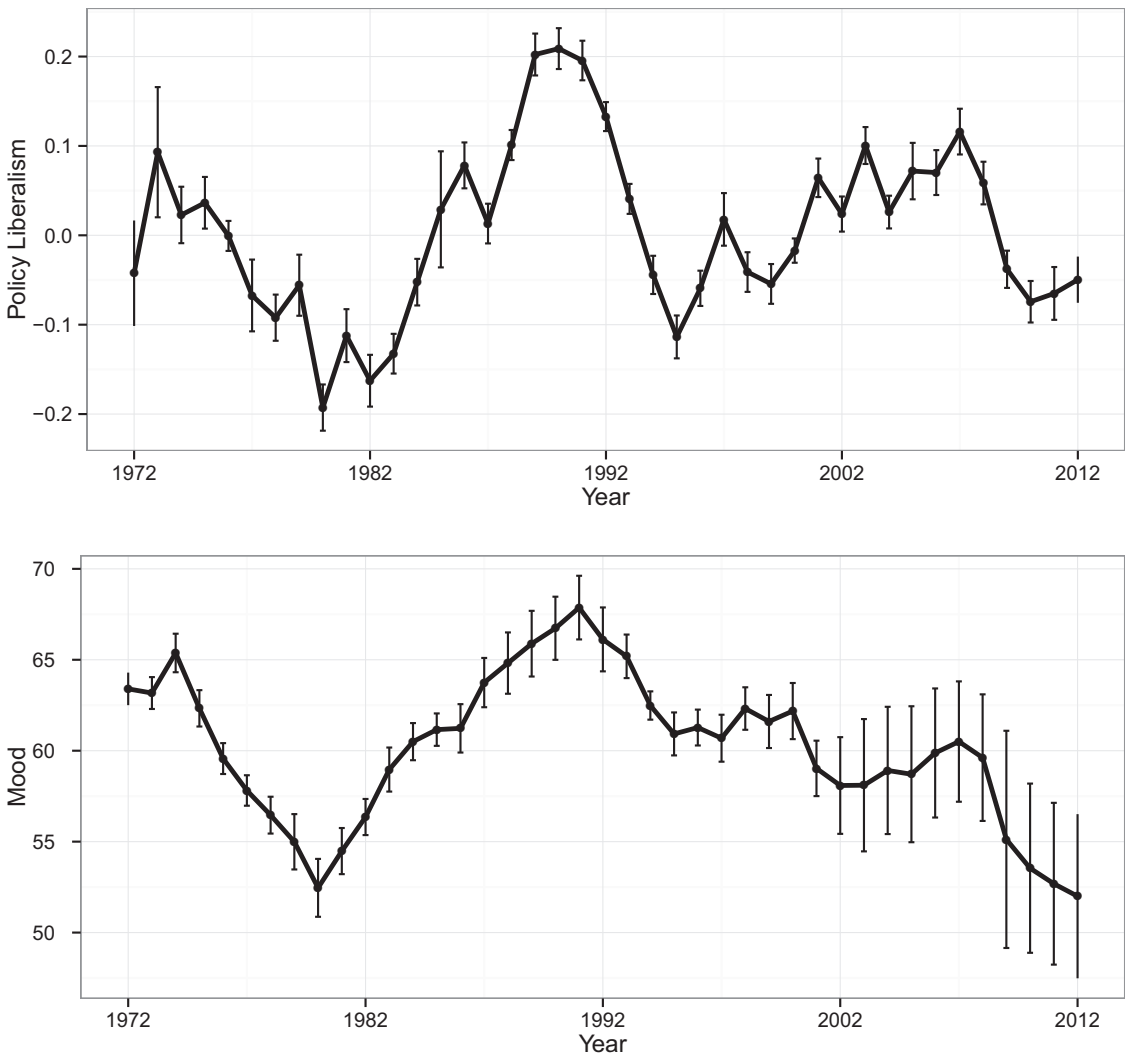


Fig. 3 Relationship between national policy liberalism and policy mood, 1972–2012.

designed to measure changes in the mass public's policy preferences over time (Stimson 1991). Of course, we should not expect a perfect correlation because mood is defined relative to the policy status quo where policy liberalism is not.

Figure 3 shows that despite the theoretical differences and limited data overlap between policy liberalism and mood, the national trends in both measures look very similar. The most liberal period for both mood and policy liberalism was around 1990, while the most conservative period was around 1980. Moreover, both mood and our measure of policy liberalism show a marked shift to the ideological right after 2006. The only major divergence between the two scales is in the early 2000s. However, note that Stimson's mood estimates are quite imprecisely estimated during this period. Overall, the correlation between policy liberalism and Stimson's mood is 0.67, which further validates the ability of our model to detect over-time changes in latent public opinion.

5 Conclusion

Recent advances in the modeling of public opinion have dramatically improved scholars' ability to measure the public's preferences on important issues. However, it has been difficult to extend these techniques to a broader range of applications due to computational limitations and problems of

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

data availability. For instance, it has been impossible to measure the public's policy preferences at the state or regional level over any length of time.

In this article, we develop a new group-level hierarchical IRT model to estimate dynamic measures of public opinion at the subnational level. We show that this model has substantial advantages over an individual-level IRT model for the measurement of aggregate public opinion. It is much more computationally efficient and permits the use of sparse survey data (e.g., where individual respondents only answer one or two survey questions), vastly increasing the applicability of IRT models to the study of public opinion.

Our model has a large number of potential substantive applications for a diverse range of topics in political science. For instance, we have shown how it could be used to generate a dynamic measure of the public's policy preferences in the United States at the level of states or congressional districts. These advances in the measurement of the public's policy preferences have the potential to facilitate new research agendas on representation and the causes and effects of public opinion more generally.

Our approach could be used for a wide variety of applications in comparative politics, where survey data are generally quite sparse. Our approach enables scholars to contract sensible measures of public opinion at the national or subnational level in both industrialized countries and emerging democracies. These new measures of public opinion could be used to examine how variation in political institutions affects the link between public opinion and policy outcomes.

Finally, our approach has implications for applications beyond the study of ideology and representation. Our model could be used to measure changes in political knowledge at both the national and subnational levels. It could also be used to measure preferences regarding specific issues or institutions. For instance, our approach could be used to measure the public's latent approval of Congress, the Supreme Court, the president or the media at the state and national levels.

Funding

The data collection and preparation for this article was supported by a grant from the MIT School of Humanities, Arts, and Social Sciences.

References

Adcock, Robert, and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3):529–46.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder, Jr. 2008. The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review* 102(2):215–32.

Armstrong, 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: CRC Press.

Bafumi, Joseph, and Michael C. Herron. 2010. Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress. *American Political Science Review* 104(3):519–42.

Bailey, Michael. 2001. Ideal Point Estimation with a Small Number of Votes: A Random-Effects Approach. *Political Analysis* 9(3):192–210.

Berry, William D., Evan J. Ringquist, Richard C. Fording, and Russell L. Hanson. 1998. Measuring Citizen and Government Ideology in the American States, 1960–93. *American Journal of Political Science* 42(1):327–48.

Buttice, Matthew K., and Benjamin Highton. 2013. How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys? *Political Analysis* 21(4):449–67.

Caughey, Devin, and Christopher Warshaw. 2014. Replication Data for: Dynamic Estimation of Latent Opinion from Sparse Survey Data Using a Group-Level IRT Model. http://dx.doi.org/10.7910/DVN/27899. Dataverse [Distributor] V1 [Version].

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The Statistical Analysis of Roll Call Data. *American Political Science Review* 98(2):355–70.

Ellis, Christopher, and James A. Stimson. 2012. *Ideology in America*. New York: Cambridge University Press.

Enns, Peter K., and Julianna Koch. 2013. Public Opinion in the U.S. States: 1956 to 2010. *State Politics and Policy Quarterly* 13(3):349–72.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. New York: Cambridge University Press.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press

———. 2006. Public Opinion in the States: A Quarter Century of Change and Stability. In *Public Opinion in State Politics*, ed. Jeffrey E. Cohen, 229–53. Palo Alto, CA: Stanford University Press.
Fiorina, Morris P., and Samuel J. Abrams. 2008. Political Polarization in the American Public. *Annual Review of Political Science* 11(1):563–88.
Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer (PDF ebook).
Fox, Jean-Paul, and Cees A. W. Glas. 2001. Bayesian Estimation of a Multilevel IRT Model Using Gibbs Sampling. *Psychometrika* 66(2):271–88.
Gelman, Andrew. 2007. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis* 1(3):515–33.
Ghitza, Yair, and Andrew Gelman. 2013. Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups. *American Journal of Political Science* 57(3):762–76.
Hoffman, Matthew D., and Andrew Gelman. Forthcoming. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*.
Jackman, Simon. 2005. Pooling the Polls over an Election Campaign. *Australian Journal of Political Science* 40(4):499–517.
———. 2009. *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: Wiley.
James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer (PDF ebook).
Jessee, Stephen A. 2009. Spatial Voting in the 2004 Presidential Election. *American Political Science Review* 103(1):59–81.
Kernell, Georgia. 2009. Giving Order to Districts: Estimating Voter Distributions with National Election Returns. *Political Analysis* 17(3):215–35.
Lax, Jeffrey R., and Justin H. Phillips. 2009. How Should We Estimate Public Opinion in The States? *American Journal of Political Science* 53(1):107–21.
Levendusky, Matthew S., Jeremy C. Pope, and Simon D. Jackman. 2008. Measuring District-Level Partisanship with Implications for the Analysis of US Elections. *Journal of Politics* 70(3):736–53.
Lewis, Jeffrey B. 2001. Estimating Voter Preference Distributions from Individual-Level Voting Data. *Political Analysis* 9(3):275–97.
Linzer, Drew A. 2013. Dynamic Bayesian Forecasting of Presidential Elections in the States. *Journal of the American Statistical Association* 108(501):124–34.
Martin, Andrew D., and Kevin M. Quinn. 2002. Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10(2):134–53.
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):115–29.
Mislevy, Robert J. 1983. Item Response Models for Grouped Data. *Journal of Educational Statistics* 8(4):271–88.
Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis* 12(4):375–85.
Park, Jong Hee. 2012. A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models. *American Journal of Political Science* 56(4):1040–54.
R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. R Foundation for Statistical Computing, http://www.R-project.org/.
Ruggles, Steven, J., Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota.
Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 1.3. http://mc-stan.org/.
Stimson, James A. 1991. *Public Opinion in America: Moods, Cycles, and Swings*. Boulder, CO: Westview.
———. 1999. *Public Opinion in America: Moods, Cycles, and Swings*. 2nd ed. Boulder, CO: Westview.
———. 2012. On the Meaning & Measurement of Mood. *Daedalus* 141(4):23–34.
Tausanovitch, Chris, and Christopher Warshaw. 2013. Measuring Constituent Policy Preferences in Congress, State Legislatures and Cities. *Journal of Politics* 75(2):330–42.
Warshaw, Christopher, and Jonathan Rodden. 2012. How Should We Measure District-Level Public Opinion on Individual Issues? *Journal of Politics* 74(1):203–19.
Wawro, Gregory J., and Ira Katznelson. 2013. Designing Historical Social Scientific Inquiry: How Parameter Heterogeneity Can Bridge the Methodological Divide between Quantitative and Qualitative Approaches. *American Journal of Political Science* 58(2):526–46.

https://doi.org/10.1093/pan/mpu021 Published online by Cambridge University Press