

Dynamic Estimation of Latent Opinion from Sparse Survey Data Using a Group-Level IRT Model

Devin Caughey* & Christopher Warshaw†

Department of Political Science
Massachusetts Institute of Technology

First draft: 2013/06/04

This draft: 2013/07/12

Abstract

Recent advances in the modeling of public opinion have dramatically improved scholars' ability to measure the public's views on important issues. For instance, Bayesian item-response theory (IRT) models provide a flexible framework for placing survey respondents in a low-dimensional space, while the combination of multilevel modeling and poststratification (MRP) improves small-area estimation of public opinion. However, it has been difficult to extend these techniques to a broader range of applications due to computational limitations and problems of data availability. In this paper, we develop a new group-level Bayesian IRT model that overcomes these limitations. Rather than estimating opinion at the individual level, we propose a hierarchical IRT model that estimates mean opinion in groups defined by demographic and geographic characteristics. Opinion change over time is accommodated with a dynamic linear model for the parameters of the hierarchical model. The group-level estimates from this model can be re-weighted to generate estimates for geographic units. This approach 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 and representation. We demonstrate the advantages of this approach for the study of the American public's policy preferences in both the modern and mid-20th century periods. We also demonstrate a potential application of our model for the study of judicial politics.

This paper was prepared for presentation at the Summer Meeting of the Society for Political Methodology, University of Virginia, Charlottesville, VA, July 18–20, 2013. We are grateful to Bob Carpenter, Kevin Quinn, Alex Storer, and Teppei Yamamoto for their helpful advice on this paper. We also appreciate valuable research assistance from Stephen Brown and Justin de Benedictis-Kessner.

*caughey@mit.edu

†cwarshaw@mit.edu

1 Introduction

Recent advances in the modeling of public opinion have dramatically improved scholars' ability to measure the public's views on important issues. Two of the most important advances in recent years are Bayesian item-response theory (IRT) models (Jessee, 2009; Treier and Hillygus, 2009) and the combination of multilevel modeling and poststratification (MRP) (Park, Gelman and Bafumi, 2004, 2006). IRT models provide a flexible framework for placing survey respondents in a low-dimensional space, and MRP improves the accuracy of opinion estimates in geographic and/or demographic subpopulations. IRT and MRP have been jointly applied to surveys in which each respondent is asked a large number of questions (Tausanovitch and Warshaw, 2013). But it has been difficult to extend these techniques to a broader range of applications due to computational limitations and problems of data availability. In particular, scholars have not been able to use these methods to examine opinion *change* over time, a central focus of public opinion research (e.g., Stimson, 1991; Page and Shapiro, 1992).

In this paper, we develop a dynamic group-level Bayesian IRT model designed to overcome these limitations. Rather than estimating opinion at the individual level, our group-level IRT model instead estimates average opinion in subpopulations defined by demographic and geographic characteristics (Mislevy, 1983). The group means are themselves modeled hierarchically (Fox and Glas, 2001), and opinion change over time is accommodated by allowing the parameters of the hierarchical model to evolve according to a dynamic linear model (Martin and Quinn, 2002). In the spirit of MRP, the group-level estimates generated by the model can be weighted and aggregated to produce time-specific opinion estimates for

states or other geographic units (Park, Gelman and Bafumi, 2004, 2006).

Our approach has substantial advantages over an individual-level IRT model for the measurement of aggregate public opinion. First, because the number of parameters is determined primarily by the number of groups rather than by the number of respondents, our group-level approach is much more computationally efficient than individual-level IRT models. Indeed, estimating an individual-level IRT model with the number of survey respondents we have (as many as 1 million) would far exceed the memory capacity of the typical personal computer. A second major advantage of our approach is that it permits the use of sparse surveys (i.e., where each respondent answers only a few questions), which describes the vast majority of available poll data, especially historically. This dramatically expands the applicability of IRT models to the study of public opinion.

We demonstrate the advantages of this approach using three substantive applications. First, we use the model to estimate the mean policy liberalism in U.S. states for each year between 1981 and 2012 (cf. Stimson, 1991; Enns and Koch, In Press). We show that our estimates are highly correlated with existing measures of state ideology that were estimated with substantially more data. In addition, the temporal dynamics of our estimates of state-level liberalism are more sensible than existing measures. As a result, our model will enable scholars to examine a wide variety of questions on representation in the modern era. For instance, scholars could re-examine whether changes in state-level ideology are causing other changes in state policy or political outcomes.

In our second application, we estimate regional support for New Deal liberalism from commercial opinion polls conducted between 1936–45. This application showcases two advantages of our approach. First, the survey data are extremely sparse in this period, making

it impossible to estimate an individual-level IRT model. Second, quota sampling rather than probability sampling was used to select respondents, and thus the poll samples are highly unrepresentative of the American public (Berinsky et al., 2011). However, by weighting our group estimates to match the groups’ distribution in the population, we obtain substantially less biased as well as more efficient regional opinion estimates. We show that our estimates are sensible and reveal novel dynamics in mass opinion during these years.

Third, we use our model to estimate state-level approval of the Supreme Court over a nearly fifty-year span, 1963 to 2010. A wide variety of theories in the judicial politics literature depend on strong measures of judicial approval or confidence. Yet, it has been difficult to develop accurate measures of Supreme Court approval due to the sparseness of survey data on the Supreme Court. We show that our measure varies sensibly overtime and is relatively highly correlated with previous measures. We also discuss several potential applications of these new estimates of judicial approval to explain variation in judicial decision making and congressional “court curbing” behavior.

Our paper proceeds as follow. First, we discuss previous approaches to modeling public opinion. We discuss both the strengths and weaknesses of the existing approaches. Next, we describe our dynamic hierarchical group-level IRT model of latent opinion. Then, we describe several substantive applications of our model. Finally, we briefly conclude.

2 Existing Approaches to Modeling Public Opinion

The measurement model we expound in this paper draws upon three important approaches to modeling public opinion: item response theory, multilevel modeling and poststratification, and dynamic measurement models. In this section, we treat each approach in turn, briefly

summarizing the literature and our model’s relationship to it.

Item response theory (IRT) was originally developed as a means of estimating subjects’ ability (or other latent trait) from their responses to categorical test questions (Lord and Novick, 1968). In the field of public opinion, IRT models have been used to generate measures of political knowledge (Delli Carpini and Keeter, 1993) and, more recently, to estimate the respondents’ latent positions in ideological space. Notwithstanding the notorious lack of constraint in the issue attitudes of mass publics (Converse, 1964), IRT models have been shown to generate useful low-dimensional summaries of citizens’ political preferences that are highly predictive of other important political attitudes and behavior (Treier and Hillygus, 2009; Tausanovitch and Warshaw, 2013). IRT models have also been used to estimate the policy ideal points of legislators and other political elites (Bailey, 2001; Martin and Quinn, 2002; Clinton, Jackman and Rivers, 2004; Shor and McCarty, 2011), sometimes in the same ideological space as ordinary citizens (Jessee, 2009; Bafumi and Herron, 2010).

Like other dimension-reduction methods, such as additive scales or factor analysis, IRT models benefit from the reduction in measurement error that comes from using multiple indicators of a single latent concept (Ansolabehere, Rodden and Snyder, 2008).¹ Yet IRT models also offer a number of methodological advantages over alternative methods. In particular, IRT models can be motivated by an explicit spatial utility model appropriate for dichotomous data (Clinton, Jackman and Rivers, 2004, 356), a feature not shared by factor analysis, which assumes multivariate normality of the responses.² The growing accessibility

¹Accurate estimation of individual-level ability parameters requires that each subject answer many questions, typically at least 15 (see, e.g., Jessee, 2009).

²When this is not an appropriate approximation (e.g., dichotomous or ordinal variables), conventional factor analysis can produce biased preference estimates (Kaplan, 2004). For a comparison of the utility models underlying factor analysis and ideal-point estimation, see Brady (1990).

of Bayesian simulation methods has further increased the range of IRT models, allowing, for example, easy characterization of the uncertainty around any parameter estimates or functions thereof.³

The second methodological approach we draw upon in this paper is multilevel regression and poststratification (Park, Gelman and Bafumi, 2004, 2006). MRP was developed as a method for estimating subnational (e.g., state) opinion from national surveys. The idea behind MRP is to model respondents' opinion hierarchically based on demographic and geographic predictors, partially pooling respondents in different states to an extent determined by the data. The smoothed estimates of opinion in each demographic cell are then weighted to match the cells' proportion in the population, yielding estimates of average opinion in each state. Subnational opinion estimates derived from this method have been shown to be more accurate than alternatives, such as aggregation across polls (Lax and Phillips, 2009*b*; Warshaw and Rodden, 2012).

MRP was originally developed to estimate average opinion on particular questions (e.g., support for gay marriage; see Lax and Phillips, 2009*a*), but it can also be applied to latent constructs, such as those measured by IRT models. Tausanovitch and Warshaw (2013) do just this, combining IRT and MRP models to estimate the ideology of states, legislative districts, and cities over the past decade. Their approach, however, has several weaknesses that limits its broader applicability. First, it requires a large number of issue questions to be asked to individual survey responses. This means that it would not be applicable to earlier eras where most surveys only asked individual respondents a handful of policy

³By contrast, classical factor analysis does not provide uncertainty estimates for the factor scores, though see Quinn (2004) and Jackman (2009, 438–53) for Bayesian implementations of factor analysis.

questions. Second, it requires substantial computational resources to estimate the latent ideology of hundreds of thousands of individuals. Finally, their approach does not directly model changing public opinion over time. Our approach offers a much more efficient way to model public opinion at the state or district level. Moreover, it allows us to easily model the evolution of public opinion at the subnational level.

The third strand of scholarship that we build on in this paper is that on dynamic measurement models, a broad class of models designed to make inferences about one or more dynamic latent variables. Early political-science applications by Beck (1990) and Kellstedt, McAvoy and Stimson (1996) modeled such aggregate constructs as presidential approval and U.S. monetary policy. Several more recent applications have taken an explicitly Bayesian approach to dynamic measurement using dynamic linear models (DLMs), including Martin and Quinn’s (2002) dynamic estimation of the ideal points of Supreme Court justices and Jackman’s (2005) dynamic model of vote intention over the course of a campaign. Of particular relevance for our purposes is Linzer’s (2013) model of state-level U.S. presidential vote intention, which employs a hierarchical specification that allows the model to borrow strength both across states and, through the use of random-walk priors, over time.

Finally, it is worth noting the connection between our work and the literature on “public policy mood” that originated with Stimson (1991). The mood literature is an important reference point for us because it too focuses on the dynamics of a low-dimensional summary of public opinion. Our work bears a particularly close connection to Enns and Koch (In Press), who model the state-level dynamics of mood based on state opinion estimates generated by an MRP-based approach. This is a promising method that performs very well on various metrics of validity.

Despite the substantive overlap in our approaches, however, they differ in key respects. Most importantly, although we are too interested in aggregate opinion, our estimates are based on an explicit individual-level IRT model of the survey response, whereas mood is fundamentally a global construct, a property of the public as a whole rather than the individuals who constitute it.⁴ One value of having an individual-level model is that it accounts for cross-sectional variation across individuals within the same framework as over-time variation. By contrast, the mood calculation algorithm only takes into account the over-time variation in question levels, and the individual-level model that underlies mood is heuristic rather than formal.

3 A Dynamic Hierarchical Group-Level IRT Model

In this section, we describe our dynamic public-opinion model, which builds on several of the approaches described in Section 2. Our aim is to use data from 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. As Figure 1 illustrates, our model has three primary components: a group-level IRT model, a hierarchical model for the group means, and a dynamic model for the hierarchical parameters. To understand the logic of the model, it is helpful to derive

⁴See Stimson (2002), however, for an insightful exploration of the micro-foundations of mood. In this piece, Stimson factor-analyzes the General Social Survey, which between 1973 and 1996 asked each respondent a number of spending questions. Using the average of the factor scores in each year as a micro-level measure of mood, Stimson shows that the over-time correlation between micro-level and aggregate measures of mood is remarkably high. Stimson’s investigation is similar in spirit to ours, but his approach is fundamentally limited by the sample size of the GSS, its unavailability before 1973, and the limited range of policy issues it covers.

it step by step, beginning with the group-level IRT model.

3.1 The Group-Level IRT Model

The conventional two-parameter IRT model characterizes each response $y_{ij} \in \{0, 1\}$ as a function of individual i 's latent *ability* (θ_i), the *difficulty* (α_j) and *discrimination* (β_j) of item j , and an error term (e_{ij}), where:

$$y_{ij} = \begin{cases} 1, & \text{if } \beta_j \theta_i - \alpha_j + e_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

If e_{ij} is assumed to be i.i.d. standard normal, then the probability of answering correctly is given by the normal ogive IRT model:

$$\Pr[y_{ij} = 1] = p_{ij} = \Phi(\beta_j \theta_i - \alpha_j) \quad (2)$$

where Φ is the standard normal CDF (Jackman, 2009, 455; Fox, 2010, 10).

Rather than modeling individual responses to each question, as in a typical IRT model, we instead model the total number of correct responses in each group g : $s_{gj} = \sum_i^{n_{gj}} y_{igj}$. Writing the model at the level of the group allows us to estimate the mean ability in each group ($\mu_{gt}^{(\theta)}$) without having to estimate the individual abilities, substantially reducing the computational burden. The essential idea is to model the θ_i as distributed normally around the group means and marginalize over the distribution of abilities.⁵

⁵For our purposes, the individual abilities are mere nuisance parameters because our real interest is estimating the *average* opinion in different demographic groups. In this respect, we share similarities with Bailey (2001) and especially Lewis (2001), who propose methods of estimating ideal points from relatively few responses that involve marginalizing over the distribution of individual abilities. These methods, however, require at

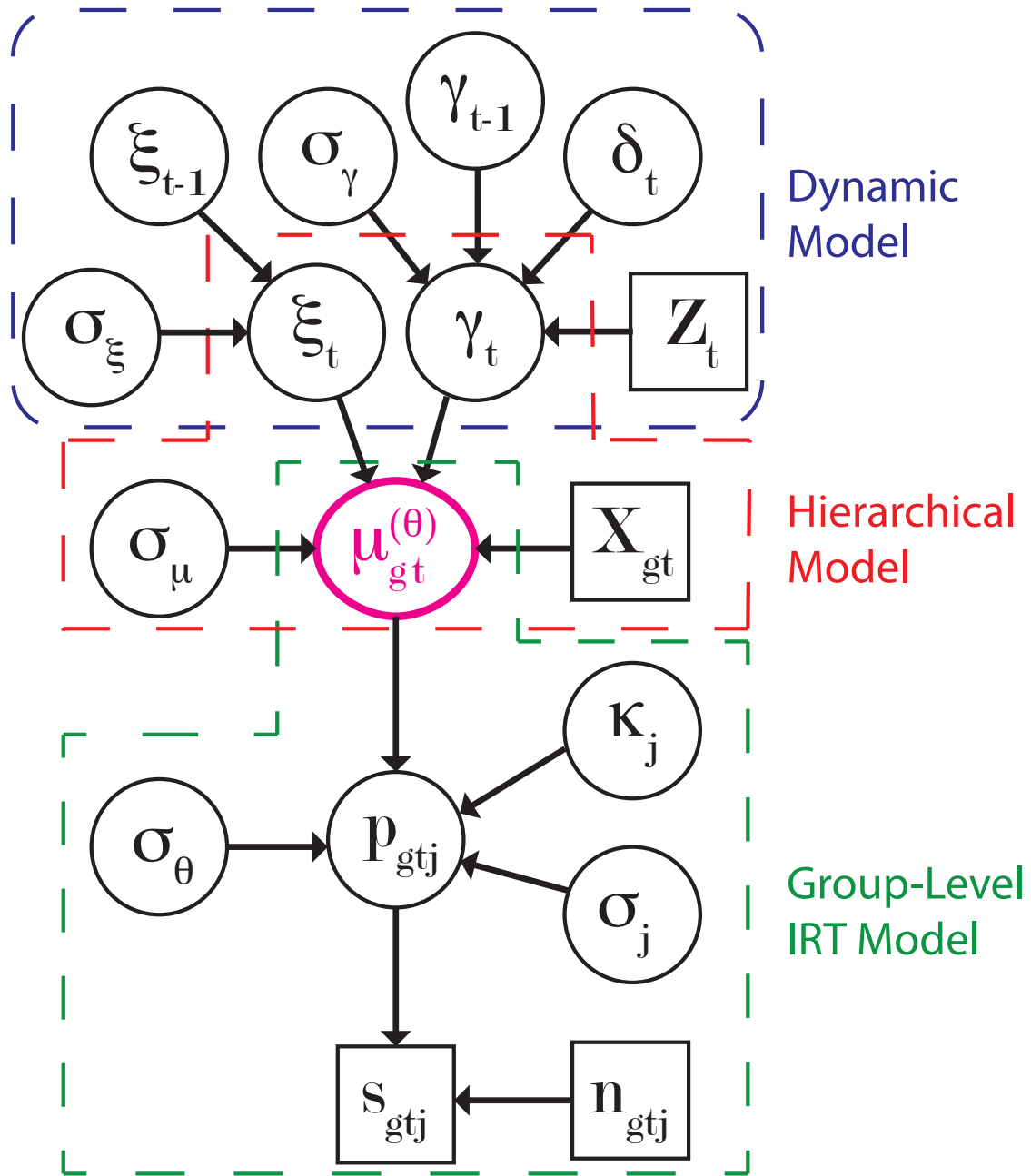


Figure 1: Directed acyclic graph of the dynamic hierarchical group-level IRT model (priors omitted). Squares and circles indicate, respectively, observed and unobserved nodes. Groups are indexed by g , items by j , and time periods by t . The target of inference is $\mu_{gt}^{(\theta)}$: mean latent opinion in each group in each year.

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

$$p_{ij} = \Phi[(\theta_i - \kappa_j)/\sigma_j] \quad (3)$$

where $\kappa_j = \alpha_j/\beta_j$ and $\sigma_j = \beta_j^{-1}$ (Fox, 2010, 11). In this formulation, the item *threshold* κ_j represents the ability level at which a respondent has a 50% probability of answering question j correctly.⁶ The *dispersion* σ_j , which is the inverse of the discrimination β_j , represents the magnitude of the measurement error for item j . Given the normal ogive IRT model and normally distributed group abilities, the probability that randomly sampled member of group g correctly answers item j is:

$$p_{gj} = \Phi[(\mu_g^{(\theta)} - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \quad (4)$$

where $\mu_g^{(\theta)}$ is the mean of the θ_i in group g , σ_θ is the within-group standard deviation of abilities, and κ_j and σ_j are the threshold and dispersion of item j (Mislevy, 1983, 278). Assuming that each respondent answers one question and each response is independent conditional on θ_i , κ_j , and σ_j , the number of correct answers to item j in each group, s_{gj} , is distributed Binomial(n_{gj} , p_{gj}), where n_{gj} is the number of non-missing responses.⁷ See Appendix A for a formal derivation of Equation 4.

least a half-dozen responses per individual, whereas in the survey data we use respondents are often asked only a single question.

⁶In terms of a spatial model, κ_j is the midpoint, or point of indifference between two choices.

⁷Multiple responses by a single respondent would introduce unobserved clustering in the data, leading to underestimates of the uncertainty surrounding the group means. To avoid this problem, we randomly sample one question for each respondent who answered more than one, coding all the other responses as missing.

3.2 The Hierarchical Model for Group Means

As stated in Equation 4, the group-level IRT model gets us what we want: estimates of the average ability in each group. The number of groups whose opinion can be effectively estimated using this model, however, is limited due to the resulting sparseness in the survey data, which leads to unstable and, in the extreme, undefined group estimates. Given this problem, it makes sense to smooth out the group IRT estimates by modeling them hierarchically (Bailey, 2001; Fox and Glas, 2001; cf. Park, Gelman and Bafumi, 2004, 2006).

We employ the following hierarchical model for the vector $\boldsymbol{\mu}^{(\theta)}$ of group means (for now, we suppress the time index t):

$$\boldsymbol{\mu}^{(\theta)} \sim \mathcal{N}(\xi + \mathbf{X}\boldsymbol{\gamma}, \sigma_{\mu}^2) \quad (5)$$

where ξ an intercept common to all groups, \mathbf{X} is a matrix of observed group characteristics, $\boldsymbol{\gamma}$ is a vector of hierarchical coefficients, and the scalar σ_{μ}^2 is the variance of $\boldsymbol{\mu}^{(\theta)}$ around the linear predictor $\xi + \mathbf{X}\boldsymbol{\gamma}$. The intercept term ξ_t captures opinion dynamics that are common to all units. It is thus akin to Stimson’s (1991) concept of national “mood,” from which the dynamics of individual groups and states may deviate.

The matrix \mathbf{X} may include geographic identifiers, demographic predictors, or interactions thereof. For example, if groups are defined by the interaction of *State*, *Race*, and *Gender*, the groups means could be modeled as an additive function of intercepts for each state as well as each racial and gender category. Since the hierarchical coefficients are themselves modeled, there is no need to exclude base categories. It therefore may be convenient to

overparameterize the hierarchical model and include indicators for all levels of each variable.

To the extent that there are many members of group g in the data, the estimate of $\mu_g^{(\theta)}$ will be dominated by the likelihood. In the opposite case of an empty cell, $\mu_g^{(\theta)}$ will be estimated based solely on the hierarchical model. In this sense, the hierarchical model functions as an imputation model for groups for which data are missing. The model thus automatically generates estimates for all groups, even those with no observed respondents.

3.3 The Dynamic Model for Hierarchical Coefficients

The group means and hierarchical coefficients in Equation 5 are not indexed by t , implicitly constraining them to be constant over time. If we wish to examine opinion change over time, however, we must relax this requirement. To do so, we add a dynamic linear model (DLM) to the hierarchical group-level IRT model described in the previous sections (on Bayesian DLMs, see Martin and Quinn, 2002; Jackman, 2009, 271–2).

We use a local-level or “random walk” model for the evolution of the common intercept ξ_t :

$$\xi_t \sim \mathcal{N}(\xi_{t-1}, \sigma_\xi^2) \tag{6}$$

where σ_ξ^2 is a time-invariant scalar to be estimated. The estimated ξ_t in each time period thus serve as priors for the estimates in the subsequent period, with σ_ξ^2 determining the relative weight of the new data. If there is no new data in period t , then the transition model in Equation 6 acts as a predictive model, imputing an estimated value for t (Jackman, 2009, 274).

We specify a more general DLM for γ_t . Let $\gamma_{p,t}$ be the value of hierarchical coefficient

p in time t , let $\mathbf{Z}_{p,t}$ be a row vector of observed predictors of $\gamma_{p,t}$, and let $\delta_{p,t}^{(\gamma)}$ and $\boldsymbol{\delta}_t^{(\mathbf{Z})}$ be time-specific transition parameters. The transition equation for $\gamma_{p,t}$ is

$$\gamma_{p,t} \sim \mathcal{N}(\delta_{p,t}^{(\gamma)} \gamma_{p,t-1} + \mathbf{Z}_{p,t} \boldsymbol{\delta}_t^{(\mathbf{Z})}, \sigma_\gamma^2) \quad (7)$$

where σ_γ^2 is also a time-invariant scalar.⁸ In essence, each $\gamma_{p,t}$ is modeled as a weighted combination of its value in the previous period ($\gamma_{p,t-1}$) and the predictors in $\mathbf{Z}_{p,t}$, with weights $\delta_{p,t}^{(\gamma)}$ and $\boldsymbol{\delta}_t^{(\mathbf{Z})}$, respectively. The coefficients $\delta_{p,t}^{(\gamma)}$ and $\boldsymbol{\delta}_t^{(\mathbf{Z})}$ can either be estimated anew in each period or modeled as a function of their previous value.⁹

We specify the transition model in Equation 7 differently depending on whether coefficient $\gamma_{p,t}$ corresponds to a demographic attribute (e.g., race) or geographic unit (e.g., state). For demographic variables, we omit exogenous predictors $\mathbf{Z}_{p,t}$ and constrain $\delta_{p,t}^{(\gamma)}$ to equal 1, thus reducing Equation 7 to the same local-level transition model defined for ξ_t in Equation 6. By contrast, we model the geographic effects in γ_t as a function of aggregate characteristics, such as *Proportion Evangelical* in a state. The inclusion of aggregate characteristics pools information across similar geographical units, improving the accuracy of the geographic effect estimates (e.g., Park, Gelman and Bafumi, 2004).

We are now in a position to write down the entire model depicted in Figure 1. Adding the indexing by t , the group-level IRT model is:

$$s_{gjt} = \text{Binomial}(n_{gjt}, p_{gjt}) \quad (8)$$

⁸Technically, we allow the innovation variance σ_γ^2 to differ between demographic and geographic predictors.

⁹We model them as evolving according to a random walk with an estimated variance: $\boldsymbol{\delta}_t \sim \mathcal{N}(\boldsymbol{\delta}_{t-1}, \sigma_\delta^2)$.

where

$$p_{gjt} = \Phi[(\mu_{gt}^{(\theta)} - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \quad (9)$$

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

$$\boldsymbol{\mu}_t^{(\theta)} \sim \mathcal{N}(\xi_t + \mathbf{X}_t \boldsymbol{\gamma}_t, \sigma_\mu^2) \quad (10)$$

whose coefficients $\boldsymbol{\gamma}_t$ evolve over time according to the transition model in Equation 7.

It is crucial to note which parameters in the model are indexed by t and which are not. The hierarchical coefficients ξ_t and $\boldsymbol{\gamma}_t$, the group means $\mu_{gt}^{(\theta)}$, and the response probabilities p_{gjt} are all allowed to vary across time periods. By contrast, the scale parameters σ_θ , σ_μ , σ_ξ , and σ_γ are constant, implying homoskedasticity both across coefficients/units and over time. The item parameters κ_j and σ_j are also constrained to be constant, implying that the mapping between the latent θ space and the response probability for a given question does not change over time. Under this crucial bridging assumption, the latent opinion estimates can be compared on a common metric across periods.

3.4 Identification, Priors, and Estimation

The parameters in an IRT model cannot be identified without restrictions on the parameter space (e.g., Clinton, Jackman and Rivers, 2004). In the case of a one-dimensional model, the direction, location, and scale of the latent dimension must be fixed *a priori*. To fix the direction of the metric, we coded all question responses so that higher values were more liberal, and restrict the sign of the discrimination parameter β_j to be positive for all items. Following Fox (2010, 88–9), we identify the location and scale by rescaling the

item parameters α and β . In each iteration m , we set the location by transforming the J difficulties to have a mean of 0: $\tilde{\alpha}_j^{(m)} = \alpha_j^{(m)} - \bar{\alpha}^{(m)}$. Similarly, we set the scale by transforming the discriminations to have a product of 1: $\tilde{\beta}_j^{(m)} = \beta_j^{(m)} (\prod_j \beta_j^{(m)})^{-1/J}$. 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 (see Equation 4).

For most parameters, we employ weakly informative priors that are proper but provide relatively little information.¹⁰ We estimated the model using the program **Stan**, as called from R (Stan Development Team, 2013; R Core Team, 2013). Stan is a C++ library that implements the No-U-Turn sampler (Hoffman and Gelman, In Press), a variant of Hamiltonian Monte Carlo that estimates complicated hierarchical Bayesian models more efficiently than alternatives such as BUGS. In general, 4,000 iterations (the first 2,000 used for adaptation) in each of 10 parallel chains proved sufficient to obtain satisfactory samples from the posterior distribution, at least for the year-specific group means.

3.5 Weighting Group Means to Estimate Geographic Opinion

The estimates of the yearly group means $\mu_{gt}^{(\theta)}$ may be of interest in themselves, but they are also useful as building blocks for estimating opinion in geographic aggregates. As Park, Gelman and Bafumi (2004, 2006) demonstrated and others (Lax and Phillips, 2009*b*; Warshaw and Rodden, 2012) have confirmed, weighting model-based group opinion estimates to match

¹⁰All standard deviation parameters are modeled as half-Cauchy with a mean of 0 and a scale of 2.5 (Gelman, 2007; Gelman, Pittau and Su, 2008). The difficulty and discrimination parameters are drawn respectively from $\mathcal{N}(0, 1)$ and $\ln\mathcal{N}(0, 1)$ prior distributions and then transformed as described above. All coefficients not modeled hierarchically are drawn from distributions centered at 0 with an estimated standard deviation, except $\delta_{t=1}^{(\gamma)}$ and $\delta_{t=1}^{(\mathbf{Z})}$, which are modeled more informatively as $\mathcal{N}(0.5, 1)$ and $\mathcal{N}(0, 1)$ respectively. Note, however, that the $\delta_t^{(\gamma)}$ and $\delta_t^{(\mathbf{Z})}$ do not enter into the model until $t = 2$ (when the first lag becomes available), and thus their values in $t = 1$ only serve as starting points for their dynamic evolution between the first and second periods.

population targets can substantially improve estimates of average opinion in states, districts, and other geographic units. Our approach to estimating opinion in geographic units extends in several respects the multilevel regression and poststratification method described by these authors.

First, unlike previous implementations of MRP, we do not estimate an individual-level model of opinion and derive from it estimates of mean opinion in each group. Rather, we directly model the group means $\mu_{gt}^{(\theta)}$. Second, our model smooths the estimates not only across space, as MRP does, but also over time, via the dynamic model for the hierarchical parameters.

A major advantage of simulation-based estimation is that it facilitates proper accounting for uncertainty in functions of the estimated parameters. For example, 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, standard deviations, and so on of the posterior distribution. We adopt this approach in presenting the results of the model in the applications that follow.

4 Applications and Validation

Having derived and explained our model in detail, we now turn to demonstrating its usefulness and validity using three applications. The first application uses survey responses to domestic policy questions in the years 1981–2012 to estimate a state-level latent opinion dimension akin to Stimson’s (1991) concept of “public policy mood.” The second applica-

tion models regional support for the New Deal based on quota-sampled opinion polls fielded between 1937 and 1945. The third application focuses on a narrower opinion dimension, tracking state-level favorability toward the Supreme Court between 1963 and 2010.

4.1 State-Level Policy Preferences, 1981–2012

Previous scholars have used a variety of approaches to measure how the mass public’s ideological preferences have evolved. One approach is to use ideological self-placement data from a representative survey (e.g., Erikson, Wright and McIver, 1994).¹¹ But since the plurality of respondents list themselves as moderate, this measure lacks granularity. Further, there is great variation in political views within each ideological category (Treier and Hillygus, 2009). More importantly, the measure lacks validity as a measure of policy preferences, because the relationship between policy preferences and ideological identification is both noisy and biased (Free and Cantril, 1967; Ellis and Stimson, 2012). Stiglitz (2009) demonstrates that use of self-placement scales varies across states, while Jessee (2009) suggests that the use of self-placement scales may vary in idiosyncratic ways across individuals.

An alternative approach is to aggregate survey questions on specific issues to measure the policy preferences or mood of the population. For instance, Stimson (1991) uses a factor-analytic algorithm to make use of the broad range of survey questions that are asked about domestic policy across many overlapping years. He extracts the common variance among survey question responses to create an overall index of policy liberalism. This “mood” measure has been used as a variable to predict a variety of outcomes, including policy

¹¹ Ideological self-placement measures are based on a categorical question that asks respondents whether they consider themselves Very Liberal, Liberal, Somewhat Liberal, Moderate, Somewhat Conservative, Conservative, or Conservative.

outcomes, elections, Supreme Court decisions, and the partisanship of individuals (Erikson, MacKuen and Stimson, 2002).

A problem with both of these approaches is that national surveys generally do not have enough respondents in each state in a given year to develop accurate estimates of state-level policy preferences (Erikson, 1978). Park, Gelman and Bafumi (2004, 2006) and Lax and Phillips (2009*b*) overcome this problem using multilevel regression and poststratification (MRP). Several recent studies have found that MRP models yield accurate estimates of public opinion in states and congressional districts using national samples of just a few thousand respondents (Park, Gelman and Bafumi, 2004, 2006; Lax and Phillips, 2009*b*; Warshaw and Rodden, 2012).

The existing study most similar to ours is Enns and Koch (In Press), which combines these approaches to measure the public’s mood at the state level between 1956 and 2010. The authors first use MRP to model the proportion of respondents in each state who favor the liberal position on each of a wide variety of policy questions. They then use Stimson’s (1991) Wcalc algorithm to combine the smoothed state-level marginals into a single yearly measure of state mood. The resulting estimates perform well in a variety of validation checks.

Unlike these previous studies, our approach models changes in state-level policy preferences through a hierarchical model, where the parameters are allowed to evolve according to a dynamic linear model (Martin and Quinn, 2002). Similarly to other recent MRP-based studies, however, we weight the annual group estimates to match their proportion in the state populations. The weights reduce the variance due to sampling error, as well as any sampling biases, in our estimates (Little and Vartivarian, 2005). In the following sections, we show that our estimates show sensible movements in state-level policy preferences.

4.1.1 Data

In order to measure state-level policy preferences in the modern era, we use a wide array of survey questions on public policy issues from hundreds of individual surveys between 1981 and 2010. The responses to each question were dichotomized at a consistent threshold (typically some version of approve/disapprove or agree/disagree) and coded so that 1 indicates a liberal response. In the modern era, respondents answered more than one question in a number of surveys (e.g., the Cooperative Congressional Election Studies). For these respondents, we randomly sample their response to one question and coded the other responses as missing. There are a total of over 650,000 non-missing responses spread across over 250 polls. Despite our large aggregate dataset, however, we still have extremely sparse data in many years during the 1980s. For instance, we have only 3,288 respondents in 1983 and 4,702 in 1984. Our hierarchical model helps smooth our estimates during this period. We bridge our estimates together over time using common questions that have stayed stable over time, such as questions about gun control, the death penalty, and universal health care.¹²

4.1.2 Model

We model opinion in groups defined by states and race.¹³ In order to smooth sampling error for small states, we model the state effects as a function of aggregate demographic characteristics of states, including *Proportion Evangelical*, *Proportion in Urban areas*, and *Proportion in a Union* in each state. The inclusion of aggregate characteristics partially pools information across similar geographical units, improving the efficiency of our estimates

¹²Note that we do not assume that questions that are explicitly defined relative to the status quo stay stable over time. For instance, we do not assume that a question asking individuals about their preference for more or less government spending means that same thing in 1981 and 2012.

¹³We use three racial categories: white, black, and other.

of the policy preferences of each state (Park, Gelman and Bafumi, 2004). Overall, we have 153 group estimates for each of 32 years. Following the MRP method, we weight the annual group estimates to match their proportion in the state populations based on the IPUMS “5-Percent Public Use Microdata Sample” from the U.S. Census (Ruggles et al., 2010).

4.1.3 Estimates

Figure 2 plots our estimates of state policy preferences between 1981 and 2012. Overall, Figure 2 shows that the states have remained generally stable in their relative preferences. This supports the findings of Erikson, Wright and McIver (2006) who have argued that the order of states’ ideological preferences has stayed relatively constant in the modern era. Notwithstanding the states’ stability relative to one another, the average policy preference in the nation has shifted considerably over time. The nation became more liberal in the late-1980s, more conservative in the early 1990s, and more liberal in the late 2000s. These shifts generally correspond with changes in election outcomes in sensible ways.

Further, consistent with Enns and Koch (In Press), the over-time variation in the liberalism of a given state is almost as large as the variation across states at a given point in time. The -0.36 change in national liberalism between 1991 and 1995, for example, is about as large as the cross-sectional difference between California and Utah. It is almost as large as the cross-sectional standard deviation of latent policy preference across individuals, which is around 0.4. It should be noted that this sort of comparison between individual-level and temporal variation would be impossible were our approach not grounded on a model of individual opinion.

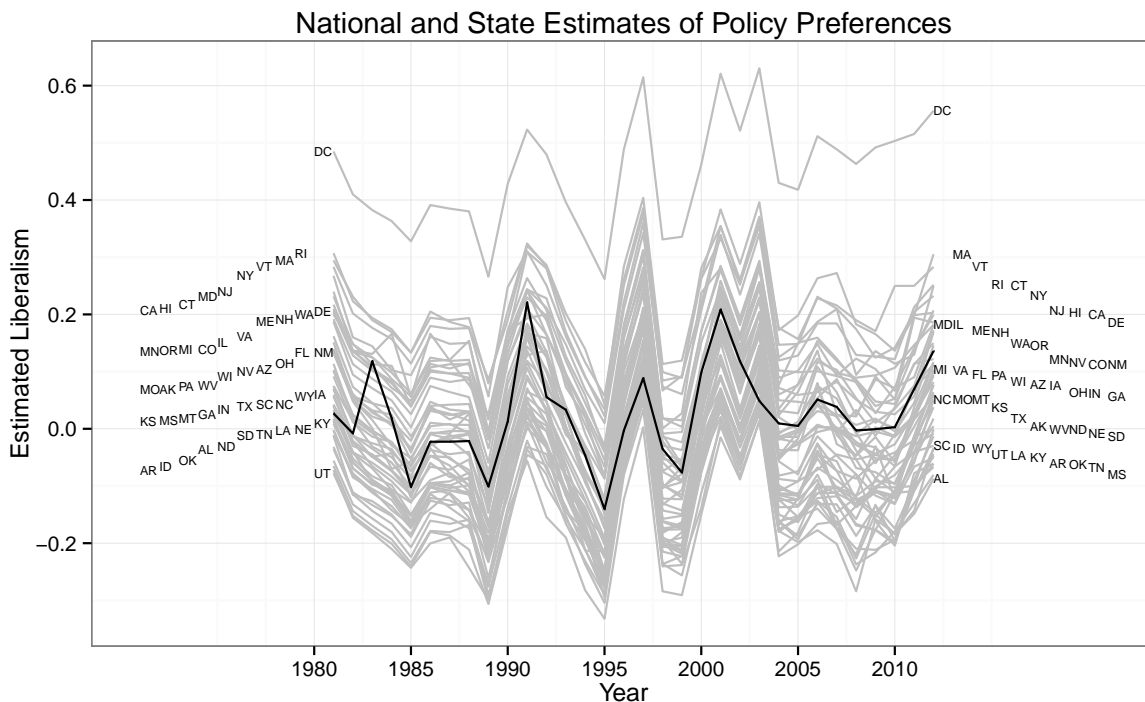


Figure 2: Policy preferences by state in the modern era.

4.1.4 Validation

How well does our measure of state policy preferences perform? One reasonable validation metric is to examine the correlation of our measure of policy preferences with presidential vote share (Tausanovitch and Warshaw, 2013). A variety of previous scholars have used election returns to estimate state and district preferences (Canes-Wrone, Brady and Cogan, 2002; Ansolabehere, Snyder and Stewart, 2001). Presidential election results are not a perfect measure of citizens' policy preferences (Levendusky, Pope and Jackman, 2008). But a high correlation with presidential vote shares would suggest our estimates are accurate measures of states' policy preferences.

The top row of Table 1 shows that over the past thirty years, the correlation between our estimates of policy preferences and presidential vote share is .79. The cross-sectional

Table 1: Correlation of Policy Preferences Measures with Presidential Vote Share

	All Years	1984	1988	1992	1996	2000	2004	2008	2012
Dynamic IRT Policy Pref.	.79	.66	.64	.79	.86	.91	.92	.93	.93
Enns and Koch Mood	.47	.82	.86	.76	.72	.87	.89	.72	NA
EWM Ideology	.42	.54	.14	.68	.47	.77	NA	NA	NA
Berry et al	.65	.65	.75	.65	.77	.79	.87	.82	NA

correlation between our estimates of policy preferences and presidential vote shares have steadily improved over the past thirty years. In the 1980s, the correlation between our measure of policy preferences and presidential vote share is about .65. The low correlation between our estimates and presidential vote share in the 1980s may reflect the sparseness of our survey data during this period. By the 2000s, when there is much more survey data available, the correlation with presidential vote share increases to above .9.

The next three rows of Table 1 show the correlation between the Democrat’s share of the two-party presidential vote in each state and previous, alternative measures of state ideology. The second row shows the correlation between Enns and Koch (In Press)’s state-level mood measure and presidential vote shares. The third row shows the correlation between Erikson, Wright and McIver (2007)’s ideology measure and presidential vote shares. Finally, the fourth row shows the correlation between Berry et al. (1998, 2007)’s citizen ideology measure and presidential vote shares. In general, our measure of state policy preferences has a higher correlation with presidential vote share than any of these alternative measures of ideology despite the fact that we have less data in many years than previous measures.¹⁴

Another reasonable validation metric is to examine how well our measure predicts depen-

¹⁴For instance, Enns and Koch (In Press) have 12,844 respondents in 1983 and 5,697 in 1984 compared to less than 5,000 respondents in our model for both of these years.

dent variables that should be partially caused by variation in state-level policy preferences. Table 2 compares the correlation between Shor and McCarty (2011)’s measure of the median voter in the lower chamber of each state’s legislature and both our measure as well as other previous measures of state ideology. In general, our measure has a higher correlation with the median voter in the state house than any of these previous measures.

Table 2: Correlation of Ideology Measures with Median Legislator in State House

	All Years	1996	1998	2000	2002	2004	2006	2008	2010
Dynamic IRT Policy Pref.	.5	.63	.67	.69	.65	.64	.72	.71	.81
Enns and Koch Mood	.38	.49	.43	.53	.57	.56	.56	.45	.67
EWM Ideology	.42	.37	.46	.59	.28	NA	NA	NA	NA
Berry et al	.60	.66	.69	.60	.50	.62	.65	.59	.71

Finally, Table 3 compares the correlation of the various ideology measures with the index of state policies in 1986 from Erikson, Wright and McIver (1994). Our measure has a substantially higher correlation with this policy index than any of the alternative measures of ideology.

Table 3: Correlation of Ideology Measures with EWM’s Policy Index (1986)

Dynamic IRT Ideology	.69
Enns and Koch Mood	.46
EWM Ideology	.50
Berry et al	.66

4.2 Regional Liberalism from Quota-Sampled Polls, 1936–1945

Existing measures of policy ideology, most notably Stimson’s policy mood, do not extend earlier than the 1950s, despite the fact that national opinion polls have existed since the

mid-1930s. A major reason for this time limitation is that until recently, individual-level data from polls conducted in the 1930s and 1940s were not easily accessible to researchers. Moreover, most polls from this era were based on quota samples rather than probability samples, and are thus unrepresentative of the American public in key respects (Berinsky et al., 2011). The resulting bias in the survey marginals renders them problematic as direct inputs into estimators such as Stimson’s Wcalc algorithm. Recently, however, a team led by Adam Berinsky and Eric Schickler has addressed these problems by converting the poll data to usable form and proposing techniques for weighting the data to population targets, thus ameliorating sample-selection bias.

4.2.1 Data

Building on the work of Berinsky and Schickler, we estimate regional liberalism based on questions drawn from quota-sampled polls conducted between 1937 and 1945. Our sample includes all 67 issue questions broadly related to the New Deal that were asked in more than one poll. Responses were dichotomized at a consistent threshold (typically some version of approve/disapprove) and coded so that 1 indicates a liberal response. For the few respondents who answered multiple question, their response to one question was randomly sampled and the other responses coded as missing. There are a total of 226,032 non-missing responses spread across nearly 100 polls.

4.2.2 Model

To reduce sample-selection bias as much as possible, we model opinion in groups defined by all demographic variables whose regional population distributions are known: *Female*, *Black*,

Farmer, *Urban*, *Professional*, *Age* (three categories), and *Phone in Household*.¹⁵ With an interaction with four-category *Region* (Midwest, Northeast, South, and West), there are a total of 768 groups observed over 9 years.

Next, we weight the yearly group estimates to match their proportion in the regional populations. Since we do not always know the complete joint distribution of auxiliary variables in the population, we employ a more general weighting framework, calibration estimation, of which poststratification and raking are special cases (Deville and Särndal, 1992; Bethlehem, 2002). Calibration estimation enables us to weight the group estimates so as to match all information available regarding the marginal and joint distributions of auxiliary variables in the population. To the extent that the weighting variables predict respondents' poll responses and their probability of being sampled, the weights reduce the variance as well as the bias of the regional estimates (Little and Vartivarian, 2005).

4.2.3 Estimates

Figure 3 plots the estimated average liberalism in four regions, with associated 68% error bands. Two interesting patterns emerge from these estimates. The first is the change in the relative position of the South, which switched from the most liberal region before 1941 to the most conservative region afterwards. This shift in the ideological ordering of regions, however, pales in comparison to a second pattern, the rightward turn among the public as a whole, which was apparently concentrated between 1940 and 1942.

¹⁵The IPUMS microsample of the 1940 U.S. Census (Ruggles et al., 2010) contains the joint distribution of all of these variables except *Phone in Household*, whose marginal distribution in each region was calculated from AT&T corporate records.

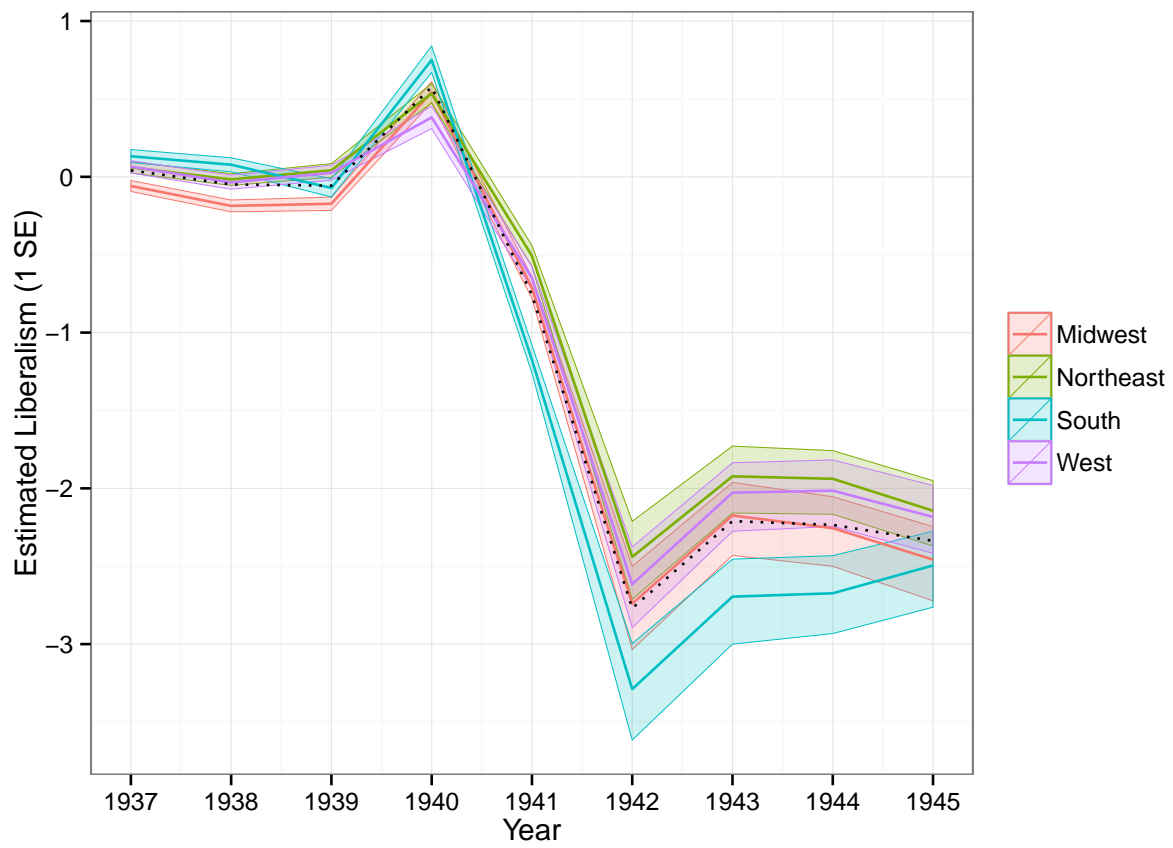


Figure 3: Estimated Regional Support for New Deal Liberalism, 1937–45

4.2.4 Validation

Both the regional and temporal patterns in Figure 3 are consistent with previous scholarship on the conservative reaction in this era. As a number of works have emphasized, this was a period of growing fatigue with liberal reform and of rising anti-labor sentiment, particularly among white Southerners (Patterson, 1967; Garson, 1974; Katznelson, Geiger and Kryder, 1993; Brinkley, 1995; Schickler and Caughey, 2011). Previous studies, however, have missed the magnitude and timing of the rightward shift in public opinion displayed in Figure 3, in part because of the difficulty of making over-time comparisons without a means of putting different questions on a common metric.¹⁶ Although our over-time estimates should be interpreted cautiously due to the limited number of questions asked both before and after 1941, the large and relatively sudden shift in mass policy preferences in 1940–42 is consistent with the trends on individual policy questions as well as with the large Republican gains in the 1942 congressional elections.¹⁷ The ability to detect such absolute shifts in public opinion is one of the foremost advantages of our model.

¹⁶One work on this period that does measure absolute ideological changes over time is Ellis and Stimson (2009), who examine trends in symbolic ideology beginning in 1937. They too find a fairly abrupt shift followed by a few years of ideological stability, but a bit earlier than we do, between 1938 and 1939. As these authors emphasize, however, the symbolic ideology has surprisingly little connection to the content of Americans’ policy views, which may explain the temporal disjuncture between our findings. The fact that Ellis and Stimson use the unweighted survey marginals may also be a factor.

¹⁷As an example, consider the question “Do you think the Social Security Program should be changed to include... Farmers?”, which discriminates well between conservatives and liberals. This question was asked in 1941 and 1944, thus spanning most of the conservative shift evident in the attached figure. In 1941, 93% of respondents (unweighted) supported including farmers, the liberal position. In 1944, the figure was 72%, a drop of 21 percentage points (−0.9 on the probit scale). By way of comparison, the largest cross-sectional difference between regions on this question is 6 percentage points. The fitted difference between an urban non-professional without a phone and a rural professional with a phone—two politically disparate demographic groups—is around 20 percentage points. In other words, this is a big drop in support relative to cross-sectional differences. Given this question’s estimated cutpoint ($\kappa_j = -2.7$) and dispersion ($\sigma_j = 1.1$) and the estimated within-group standard deviation of ideal points ($\sigma_\theta = 0.95$), we can use Equation 4 to calculate the predicted change in national support for this question. If we plug the national mean of the estimated ideal points in 1941 and 1944, we get predicted national support levels of 91% for 1941 and 70% in 1944—very close to the observed values of 93% and 72% in the raw data.

4.3 State-Level Trust in the Supreme Court, 1965–2010

Public opinion on the Supreme Court plays a key role in many theories of judicial politics. A large number of scholars have found that the public's confidence in the Supreme Court affects the interaction between the Court and other branches. The Court is sensitive to how it is perceived by the public (Baum, 2009). As a result, it is more likely to issue unpopular decisions or strike down acts of Congress when it is relatively popular (Caldeira, 1987; Carrubba, 2009; Clark, 2011; Hausseger and Baum, 1999). Congress is also sensitive to how the Court is perceived by the public. Members of Congress are more likely to support legislation that limits the Court's power when public support for the Court is low (Clark, 2009, 2011). In addition, scholars have examined the factors that explains changes in the public's confidence in the Court overtime. Mondak and Smithey (1997) find that the Court's support erodes when its decisions diverge from the ideological preferences of the American public.

However, previous work on the role of public opinion in judicial politics has been limited by the difficulty in measuring confidence in the Court either overtime or across states. Clark (2009) writes that "public opinion data about the Court are notoriously sparse" (p. 979). Previous scholars have generally measured support for the Court using aggregated responses to the General Social Survey (GSS) and Harris polls (Caldeira, 1986; Clark, 2009, 2011). But this approach leaves scholars with just a few dozen survey responses in individual states in a given year. Clark (2011) develops better state-level estimates by using a multi-level regression with post-stratification (MRP) model with data from the GSS. But this approach provides no solution to the fact that in some years there is no data at all available from the

GSS or Harris surveys. Moreover, it fails to utilize all of the available data from Gallup and other survey firms on judicial approval or confidence.

Another approach is to aggregate data across a wider range of survey firms, such as Gallup, Pew Research Center for the People and the Press, CBS News, and the GSS (Mondak and Smithey, 1997). But it has been difficult to aggregate data across firms. First, survey questions differ dramatically across surveys. For instance, the GSS asks respondents how much “confidence” they have in “the people running the Supreme Court.” In contrast, recent Pew surveys has asked respondents whether their “overall opinion of the Supreme Court is very favorable, mostly favorable, mostly unfavorable, or very unfavorable?” Second, even after aggregating across all available polls, survey data is still sparse in many individual years and states.

Our model builds upon previous approaches by pooling across survey questions and polling firms to estimate latent trust in the Supreme Court at the state-level. Our dynamic model enables us to estimate trust in the Supreme Court even in years with little or no available survey data. Moreover, our multi-level model partially-pools surveys across states to improve our estimates of states with sparse survey data.

This new measure could enable scholars to re-examine whether Senators are more likely to support legislation that limits the Court’s power when public support for the Court is low. It also enables scholars to expand our analysis of the interaction between the Court and political officials to new arenas. For instance, scholars could examine whether state-level officials are more likely to challenge the Court when the Court is unpopular in their state.

4.3.1 Data

We use data from 65 polls between 1963 and 2010 with approximately 87,000 total respondents. We use the following four question series as indicators of latent trust in the Court:

- Do you approve or disapprove of the way the Supreme Court is handling its job?
- In general, what kind of rating would you give the Supreme Court?
- Would you tell me how much respect and confidence you have in the Supreme Court?
- Is your overall opinion of the Supreme Court very favorable, mostly favorable, mostly unfavorable, or very unfavorable?

Some of these questions have multiple ordinal response categories (e.g., “very favorable,” etc.). To maximize the range of cutpoints with respect to the underlying latent variable, we convert each ordinal variable into a set of dichotomous variables that indicate whether the response was above a given threshold. We model the sum of each of these dichotomous variables, sampling one variable from each respondent so as to avoid having multiple responses from a given individual.

4.3.2 Model

We model opinion in groups defined by states and three demographic categories (race, education, and gender).¹⁸ In order to smooth sampling error for small states, we model the state effects as a function of the *Proportion Evangelical* in each state. As in our other models, the inclusion of aggregate characteristics partially pools information across similar geographical units, improving the efficiency of our estimates of the latent opinion in each state (e.g., Park, Gelman and Bafumi, 2004, 2006). Overall, we have 1,224 groups observed over 48 years. We then weight the annual group estimates to match their proportion in the state populations based on the IPUMS “5-Percent Public Use Microdata Sample” from the Census (Ruggles et al., 2010).

¹⁸ We use three racial categories: white, black, and other. For education, we use five categories: no high school degree, high school degree, some college, college graduate, and graduate school degree. For gender, we use male and female.

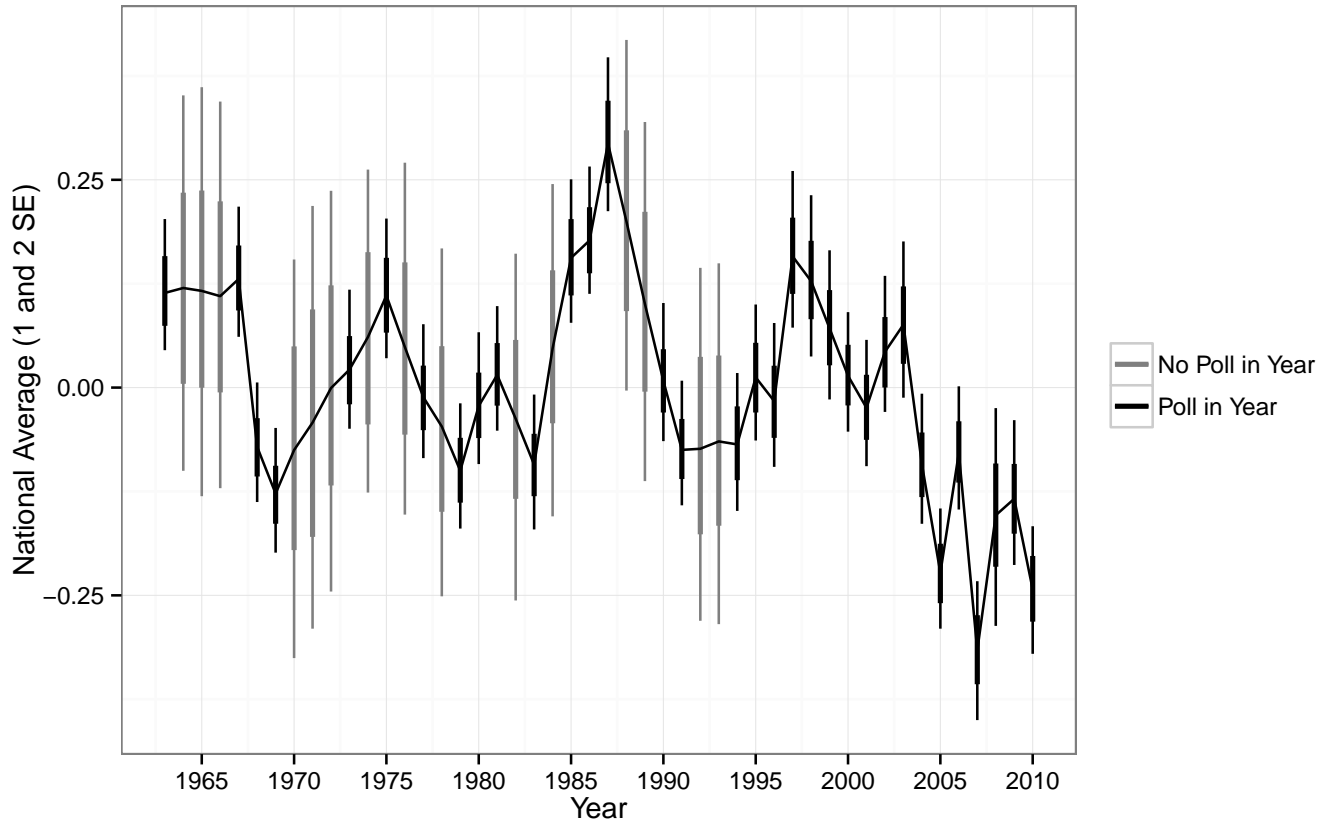


Figure 4: Public Opinion on the Supreme Court - This figure shows how latent confidence in the Supreme Court is changing overtime at the national level.

4.3.3 Estimates

Figure 4 shows our estimates of national-level trust in the Supreme Court. The graph shows how our model smooths the estimates of latent opinion for years where we lack data on public opinion regarding the Supreme Court. However, the confidence intervals on our estimates increase substantially in years where we lack survey data. For instance, the graph shows that the standard errors substantially increase in the mid-1960s when we lack data, but decrease in the late 1960s when there is more survey data available.

The graph show several sensible patterns. First, there is a general drop in support for the Court during the late Warren Court, which may reflect the unpopularity of the due process revolution (e.g., *Miranda v. Arizona*, 1966). Second, there is a large drop in trust

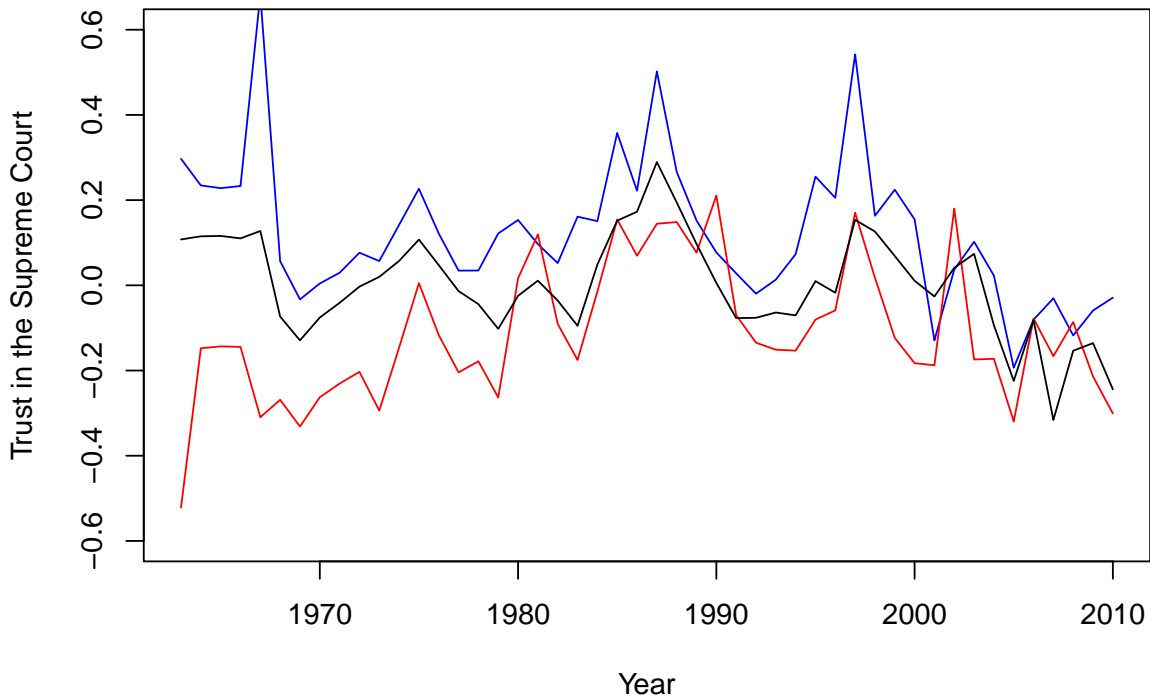


Figure 5: Public Opinion on the Supreme Court - This figure shows how latent confidence in the Supreme Court is changing overtime in Massachusetts (blue) and South Carolina (red) The national level of confidence in the Court is in black.

in the Court in the late 1970s and after the Iraq War, which may reflect general low points in political trust. Third, the graph shows that there is a small net drop in 2001, which may reflect the impact of *Bush v. Gore* on the public’s trust in the Court.

Figure 5 compares state-level support for the Court in Massachusetts and South Carolina. The graph shows that the evolution of trust in the Court in these states reflected changes in the general ideological orientation of the Court. In the early part of the period, there is generally lower support for the Court in South Carolina, which probably reflects southern states’ dissatisfaction with the Court liberal decisions on school de-segregation and criminal justice. In contrast, there is very strong support for the Court in Massachusetts during the 1960s and early 1970s. Overtime, however, support for the Court drops in Massachusetts

Table 4: Correlation of Our Judicial Trust Measure with Clark (2011)’s Measure of Supreme Court Disapproval

AllYears	1980	1984	1988	1992	1996	2000	2004
-.26	-.25	-.90	.17	-.87	-.57	-.59	-.52

and rises in South Carolina. These changes likely reflect the general shift in the Court’s orientation to the ideological right. Finally, notice that support for the Court sharply increases in South Carolina in 2001 in the wake of *Bush v. Gore*, while it drops significantly in Massachusetts. This pattern of differential response to Court rulings between liberal and conservative states is evident at several points, highlighting the value of allowing each state to follow its own temporal trajectory.

4.3.4 Validation

It is difficult to validate our estimates of state-level trust in the Supreme Court since there is no comparable measure that utilizes data from a wide variety of sources. The best existing estimates of state-level trust in the Supreme Court are Clark (2011)’s estimates of the level of explicit *lack of confidence* in the Court in each state from the 94th through the 109th Congresses (1975-2006). These estimates improve upon simple disaggregation by using a multi-level regression with post-stratification (MRP) model with data from the General Social Survey (GSS). Table 4 shows that our measure is relatively quite highly correlated with Clark’s measure of lack of confidence in the Court in each state within each year. However, the overtime correlation is fairly low, which may reflect the noisiness in the GSS’s measure of judicial trust due to its generally small sample sizes.

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 limita-

tions and problems of 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 paper, we develop a new group-level hierarchical IRT model to estimate dynamic measures of public opinion at the sub-national 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. They equip us to re-examine the extent of constituency influence in Congress (Miller and Stokes, 1963). They also equip us to expand our study of the impact of public opinion on policy outcomes (Erikson, Wright and McIver, 1994; Lax and Phillips, 2011).

More generally, our approach could be used for a wide variety of applications in comparative politics, where survey data is generally quite sparse. Our approach enables scholars to construct sensible measures of public opinion at the national or sub-national 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 sub-national levels. It could also be used to measure preferences regarding specific issues or institutions. In this paper, we have shown that our approach can easily be applied to measure the public’s trust in the Supreme Court, which could be used to facilitate

new areas of research in judicial politics. Likewise, our approach could be used to measure the public's approval in Congress, the President, or the media at the state and national levels.

References

- Ansolabehere, Stephen, James M. Snyder, Jr. and Charles Stewart, III. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45(1):136–159.
- 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):215–232.
- 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):519–542.
- Bailey, Michael. 2001. "Ideal Point Estimation with a Small Number of Votes: A Random-Effects Approach." *Political Analysis* 9(3):192–210.
- Baum, Lawrence. 2009. *Judges and Their Audiences: A Perspective on Judicial Behavior*. Princeton, NJ: Princeton University Press.
- Beck, Nathaniel. 1990. "Estimating Dynamic Models Using Kalman Filtering." *Political Analysis* 1:121–156.
- Berinsky, Adam J., Eleanor Neff Powell, Eric Schickler and Ian Brett Yohai. 2011. "Revisiting Public Opinion in the 1930s and 1940s." *PS: Political Science & Politics* 44(3):515–520.
- 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* pp. 327–348.
- Berry, William D, Evan J Ringquist, Richard C Fording and Russell L Hanson. 2007. "The measurement and stability of state citizen ideology." *State Politics & Policy Quarterly* 7(2):111–132.
- Bethlehem, Jalke G. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little. New York: Wiley chapter 18, pp. 275–287.
- Brady, Henry E. 1990. "Traits versus Issues: Factor versus Ideal-Point Analysis of Candidate Thermometer Ratings." *Political Analysis* 2(1):97–129.
- Brinkley, Alan. 1995. *The End of Reform: New Deal Liberalism in Recession and War*. New York: Vintage Books.
- Caldeira, Gregory. 1987. "Public Opinion and the U.S. Supreme Court: FDR's Court-Packing Plan." *American Political Science Review* 81:1139–1153.
- Caldeira, Gregory A. 1986. "Neither the Purse Nor the Sword: Dynamics of Public Confidence in the Supreme Court." *American Political Science Review* 80(4):1209–26.

- Canes-Wrone, Brandice, David W. Brady and John F. Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members' Voting." *American Political Science Review* 96(1):127–140.
- Carrubba, Clifford James. 2009. "A Model of the Endogenous Development of Judicial Institutions in Federal and International Systems." *Journal of Politics* 71(1):55–69.
- Clark, Thomas S. 2011. *The Limits of Judicial Independence*. New York: Cambridge University Press.
- Clark, Tom S. 2009. "The Separation of Powers, Court Curbing, and Judicial Legitimacy." *American Journal of Political Science* 53(4):971–89.
- Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.
- Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David E. Apter. Free Press pp. 206–261.
- DasGupta, Anirban. 2011. *Fundamentals of Probability: A First Course*. Springer (PDF ebook).
- Delli Carpini, Michael X and Scott Keeter. 1993. "Measuring Political Knowledge: Putting First Things First." *American Journal of Political Science* 37(4):1179–1206.
- Deville, Jean-Claude and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–382.
- Ellis, Christopher and James A. Stimson. 2009. "Symbolic Ideology in the American Electorate." *Electoral Studies* 28(3):388–402.
- Ellis, Christopher and James A. Stimson. 2012. *Ideology in America*. New York: Cambridge UP.
- Enns, Peter K and Julianna Koch. In Press. "Public Opinion in the U.S. States: 1956 to 2010." *State Politics and Policy Quarterly*.
- Erikson, Robert S. 1978. "Constituency Opinion and Congressional Behavior: A Reexamination of the Miller-Stokes Representation Data." *American Journal of Political Science* 22(3):511–535.
- Erikson, Robert S, Gerald C Wright and John P McIver. 1994. *Statehouse democracy: Public opinion and policy in the American states*. Cambridge University Press.
- Erikson, Robert S., Gerald C. Wright and John P. McIver. 2006. Public Opinion in the States: A Quarter Century of Change and Stability. In *Public Opinion in State Politics*, ed. Jeffrey E. Cohen. Palo Alto, CA <http://site.ebrary.com/lib/berkeley/Doc?id=10156556&ppg=248>: Stanford University Press chapter 12, pp. 229–253.

- Erikson, Robert S, Gerald C Wright and John P McIver. 2007. "Measuring the public's ideological preferences in the 50 states: Survey responses versus roll call data." *State Politics & Policy Quarterly* pp. 141–151.
- Erikson, Robert S., Michael B. MacKuen and James A. Stimson. 2002. *The Macro Polity*. New York: Cambridge University Press.
- Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications*. 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):271–288.
- Free, Lloyd A. and Hadley Cantril. 1967. *The Political Beliefs of Americans: A Study of Public Opinion*. New Brunswick, NJ: Rutgers UP.
- Garson, Robert A. 1974. *The Democratic Party and the Politics of Sectionalism, 1941–1948*. Baton Rouge: Louisiana State University Press.
- Gelman, Andrew. 2007. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1(3):515–533.
- Gelman, Andrew, Maria Grazia Pittau and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2(4):1360–1383.
- Hausseger, Lori and Lawrence Baum. 1999. "Inviting Congressional Action: A Study of Supreme Court Motivations in Statutory Interpretation." *American Journal of Political Science* 43(1):162–85.
- Hoffman, Matthew D. and Andrew Gelman. In Press. "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.
- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Wiley (PDF ebook).
- Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103(1):59–81.
- Kaplan, David. 2004. *The Sage handbook of quantitative methodology for the social sciences*. Sage.
- Katznelson, Ira, Kim Geiger and Daniel Kryder. 1993. "Limiting Liberalism: The Southern Veto in Congress, 1933–1950." *Political Science Quarterly* 108(2):283–306.

- Kellstedt, Paul, Gregory E. McAvoy and James A. Stimson. 1996. "Dynamic Analysis with Latent Constructs." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 5 (1993):113–150.
- Lax, Jeffrey R. and Justin H. Phillips. 2009a. "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103(3):367–386.
- Lax, Jeffrey R. and Justin H. Phillips. 2009b. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53(1):107–121.
- Lax, Jeffrey R. and Justin H. Phillips. 2011. "The democratic deficit in the states." *American Journal of Political Science* 56(1):148–166.
- 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):736–753.
- Lewis, Jeffrey B. 2001. "Estimating Voter Preference Distributions from Individual-Level Voting Data." *Political Analysis* 9(3):275–297.
- Linzer, Drew A. 2013. "Dynamic Bayesian Forecasting of Presidential Elections in the States." *Journal of the American Statistical Association* 108(501):124–134.
- Little, Roderick J. and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2):161–168.
- Lord, Frederic M. and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- 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):134–153.
- Miller, Warren E. and Donald E. Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57(1):45–56.
- Mislevy, Robert J. 1983. "Item Response Models for Grouped Data." *Journal of Educational Statistics* 8(4):271–288.
- Mondak, Jeffrey J. and Shannon Ishiyama Smithey. 1997. "The Dynamics of Public Support for the Supreme Court." *Journal of Politics* 49(4):1114–42.
- Page, Benjamin I. and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago.
- 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.

- Park, David K., Andrew Gelman and Joseph Bafumi. 2006. State Level Opinions from National Surveys: Poststratification Using Multilevel Logistic Regression. In *Public Opinion in State Politics*, ed. Jeffrey E. Cohen. Stanford, CA: Stanford University Press pp. 209–228.
- Patterson, James T. 1967. *Congressional Conservatism and the New Deal: The Growth of the Conservative Coalition in Congress, 1933–1939*. Lexington: University of Kentucky Press.
- Quinn, Kevin M. 2004. “Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses.” *Political Analysis* 12(4):338–353.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, url<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.
- Schickler, Eric and Devin Caughey. 2011. “Public Opinion, Organized Labor, and the Limits of New Deal Liberalism, 1936–1945.” *Studies in American Political Development* 25(October):1–28.
- Shor, Boris and Nolan McCarty. 2011. “The Ideological Mapping of American Legislatures.” *American Political Science Review* 105(3):530–51.
- Stan Development Team. 2013. “Stan: A C++ Library for Probability and Sampling, Version 1.3.” <http://mc-stan.org/>.
- Stiglitz, Edward H. 2009. “Liberal Relative to Whom? Ideology and Policy Preferences in the American Public.” Unpublished working paper.
- Stimson, James A. 1991. *Public Opinion in America: Moods, Cycles, and Swings*. Boulder: Westview.
- Stimson, James A. 2002. The Micro Foundations of Mood. In *Thinking About Political Psychology*, ed. James H Kuklinski. New York: Cambridge University Press pp. 253–280.
- Tausanovitch, Chris and Christopher Warshaw. 2013. “Measuring Constituent Policy Preferences in Congress, State Legislatures and Cities.” *Journal of Politics* 75(2):330–342.
- Treier, Shawn and D Sunshine Hillygus. 2009. “The Nature of Political Ideology in the Contemporary Electorate.” *Public Opinion Quarterly* 73(4):679–703.
- Warshaw, Christopher and Jonathan Rodden. 2012. “How Should We Measure District-Level Public Opinion on Individual Issues?” *Journal of Politics* 74(1):203–219.

A Derivation of Group Normal Ogive IRT Model

This appendix derives the in group-level model in Equation 4. The same result is shown by Mislevy (1983), but our derivation is different.

The model depends on the following assumptions:

1. The responses to question j are independent conditional on θ_{ig} , κ_j , and σ_j .
2. Within each group, the θ_{ig} are normally distributed with group-specific means and common variance: $\theta_{ig} \sim \mathcal{N}(\mu_g, \sigma_\theta^2)$. (Note that the common variance implies homoskedasticity of the group ability distributions.)
3. The n_{gj} subjects in group g who answer question j were randomly sampled from that group, independently from the $n_{gj'}$ who answer question $j' \neq j$. (This assumption would be violated if each respondent answered more than one question.)

Equation 3 implies that respondent i in group g answers item j correctly if and only if:

$$(\theta_{ig} - \kappa_j)/\sigma_j + \epsilon_{ij} > 0 \quad (11)$$

Multiplying by σ_j , the inequality in Equation 11 becomes:

$$\theta_{ig} - \kappa_j + \epsilon_{ij}\sigma_j > 0 \quad (12)$$

Letting $z_{igj} = \theta_{ig} - \kappa_j + \epsilon_{ij}\sigma_j$, the probability that a randomly sampled member of group g correctly answers question j is:

$$\Pr[y_{igj} = 1] = \Pr[z_{igj} > 0] \quad (13)$$

By Assumption 3, the individual abilities θ_{ig} are distributed $\mathcal{N}(\mu_g, \sigma_\theta^2)$. Since ϵ_{ij} has a standard normal distribution, the term $\epsilon_{ij}\sigma_j$ is distributed $\mathcal{N}(0, \sigma_j^2)$. The sum of two independent normal variables has a normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$ (DasGupta, 2011, 326), so:

$$z_{igj} \sim \mathcal{N}(\mu_g^{(\theta)} - \kappa_j, \sigma_\theta^2 + \sigma_j^2) \quad (14)$$

Since the CDF of a normal variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is $\Phi(\frac{x-\mu}{\sigma})$, the CDF of z_{igj} is:

$$\Pr[z_{igj} \leq x] = \Phi\left[\frac{x - (\mu_g^{(\theta)} - \kappa_j)}{\sqrt{\sigma_\theta^2 + \sigma_j^2}}\right] \quad (15)$$

which implies:

$$\begin{aligned}
\Pr[z_{igj} > 0] &= 1 - \Phi\left[\frac{0 - (\mu_g^{(\theta)} - \kappa_j)}{\sqrt{\sigma_\theta^2 + \sigma_j^2}}\right] \\
&= 1 - \Phi[-(\mu_g^{(\theta)} - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \\
&= \Phi[(\mu_g^{(\theta)} - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \\
&= p_{gj}
\end{aligned} \tag{16}$$

“In other words,” writes Mislevy (1983, 278), “if $[\kappa_j]$ and σ_j are the item threshold and dispersion parameters in the subject-level model, then $[\kappa_j]$ and $\sqrt{\sigma_\theta^2 + \sigma_j^2}$ are the item threshold and dispersion parameters in the group-level model.” The response to each question being a Bernoulli draw with constant probability p_{gj} , the sum of correct answers in group g is distributed $s_{gj} \sim \text{Binomial}(n_{gj}, p_{gj})$, where n_{gj} is the number of valid responses to question j in group g .