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

Citation	Caughey, D., and C. Warshaw. "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." Political Analysis 23.2 (2015): 197–211.
As Published	http://dx.doi.org/10.1093/pan/mpu021
Publisher	Oxford University Press
Version	Original manuscript
Citable link	http://hdl.handle.net/1721.1/106583
Terms of Use	Creative Commons Attribution-Noncommercial-Share Alike
Detailed Terms	http://creativecommons.org/licenses/by-nc-sa/4.0/

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

Devin Caughey[*]
Department of Political Science
MIT

Christopher Warshaw[†]
Department of Political Science
MIT

2014/08/03

ABSTRACT

Since the advent of nationally representative public-opinion polls in the 1930s, millions of Americans have been queried about their political opinions. Until recently, however, it was very rare for respondents to be asked more than a few policy questions, rendering dimension-reduction techniques such as item-response models inapplicable at the individual level. Instead, scholars have been largely constrained to analyzing individual question series, which have uneven continuity and availability over time. In this paper, we develop a dynamic group-level item-response model that overcomes these limitations. Rather than estimating opinion at the individual level, our Bayesian approach 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 enables scholars to measure latent variables across geographic space and over time in a unified framework. As a result, it opens up vast new areas of research on public opinion and representation using the full array of public opinion data from the past 80 years.

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.

[*] Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, caughey@mit.edu

[†] Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, cwarshaw@mit.edu

Contents

1. INTRODUCTION

Since the advent of commercial public-opinion polling in the 1930s, millions of Americans have been surveyed on their political opinions, yielding a wealth of information on the political attitudes of the American public over the past eight decades. Political scientists' 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 dimension-reduction techniques such as factor analysis. Second, few questions have been asked in comparable fashion across many polls, making it difficult to evaluate opinion change over time. This sparseness of the survey data has forced 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, applied studies of state politics have typically relied on measures that crudely proxy for their quantity of interest and that often 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 the weighted average of the scaled roll call votes of elected officials (Berry et al., 1998, 2010) or on time-invariant measures of respondents' ideological self-identification (Erikson, Wright and McIver, 1993). 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 the 20th century.

Our goal in this paper 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; Jessee, 2009; Bafumi and Herron, 2010). We depart from the conventional IRT framework by estimating (latent) opinion not at the individual level, but rather at the level of groups defined by demographic and geographic characteristics (Mislevy, 1983; Lewis, 2001; McGann, 2014). In addition to yielding major computational savings, moving to the group level allows us to use survey data from respondents asked only a single question, vastly increasing the amount of usable data.

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; Tausanovitch and Warshaw, 2013). 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; Jackman, 2005; Linzer, 2013). 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 (Gelman and Little, 1997; Park, Gelman and Bafumi, 2004). Estimating the model with the Bayesian simulation program Stan permits easy calculation of posterior point estimates and their associated uncertainty (Stan Development

Team, 2013).

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 (Ansolabehere, Rodden and Snyder, 2008), but a group-level approach is much more computationally efficient and can incorporate vastly more data because a given respondent need only answer a single question. Our model's dynamic structure naturally accommodates opinion change over time, a central focus of public opinion research (e.g., Stimson, 1991; Page and Shapiro, 1992). 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; see also McGann, 2014).

This modeling framework is quite general and can be applied to a wide variety of dynamic 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 5, using it to estimate the average policy liberalism of U.S. state publics in each year between 1972 and 2012. 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 1965–2010 are provided in Appendices B and C. 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 (Bailey, 2001; Lewis, 2001).

Our paper proceeds as follows. First, we discuss the problem of data sparseness in public opinion surveys in more detail. We then review existing approaches to modeling public opinion and the strengths and weakness of each. Next, we derive and explain our dynamic group-level IRT model of latent opinion. We then present our application of the model to estimating policy liberalism in U.S. states between 1972 and 2012. We conclude with potential extensions and applications of our approach.

2. THE PROBLEM OF SPARSE OPINION DATA IN SURVEYS

An enormous amount of survey data is available to public opinion researchers. Millions of Americans have taken public opinion surveys since the 1930s. Moreover, the available survey data is not limited to the United States. Public opinion surveys have been conducted for decades in most developed countries.

Most of these surveys, however, contain only a handful of political questions, especially those pertaining to policy preferences. The sparseness of policy questions in most surveys largely precludes the use of respondent-level dimension-reduction techniques on the vast majority of available public opinion data. This problem is particularly problematic for studies of representation because most academic surveys prior to the 2000s had very small sample sizes. Thus even when enough questions are available to apply dimension-reduction techniques, small samples and uneven geographic coverage make it difficult to generate accurate estimates of subnational (e.g., state) opinion.

As an illustration of the problem of data sparseness, consider Figure 1. Each cell in this figure represents the intersection of a public-opinion poll and a unique question series related to attitudes towards New Deal liberalism (only questions asked in the same form across at least two polls are included). The cells' shading indicates whether questions were asked in a

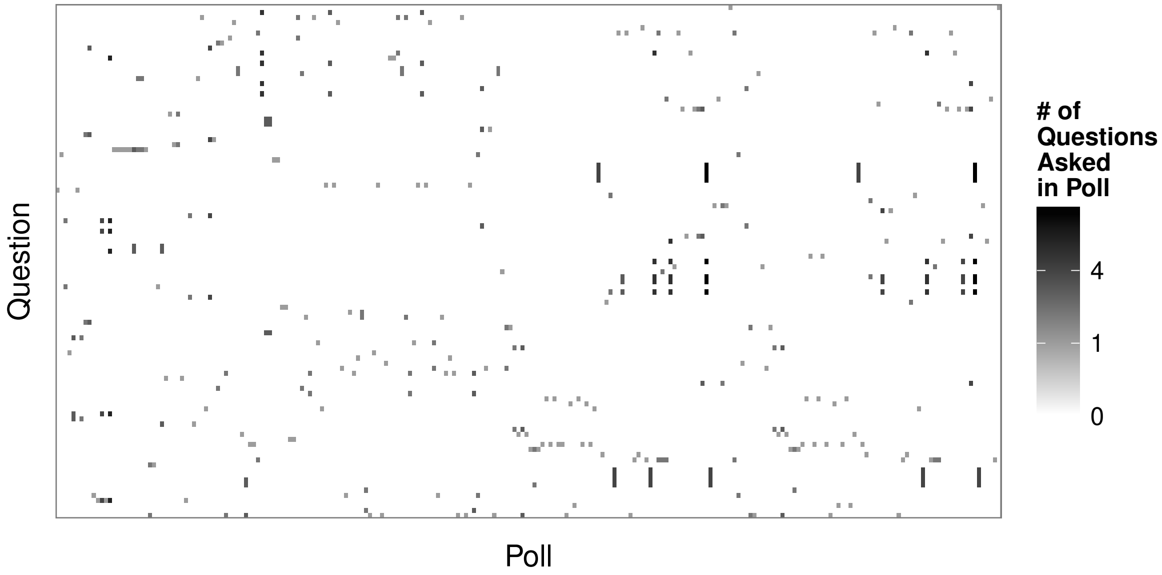

Figure 1: Availability of question series by poll, 1936–1952

particular poll and, if so, how many total questions the poll included. Not only was it rare for a poll to ask more than a couple of policy-relevant questions (the maximum number in this Figure 1 is 8), but the same question was typically asked in at most two polls, often closely spaced in time.

Unfortunately, the problem of data sparseness is not restricted to mid-20th century survey data. Most modern surveys continue to ask only a handful of policy questions. Figure 2 shows the number of policy questions related to about 50 salient economic and social policy issues across over 450 surveys between 1976 and 2012. Academic surveys such as the National Election Survey and Cooperative Congressional Election Study (CCES) have 15 or 16 policy questions. But most surveys from media organizations such as Gallup and the New York Times have just a few questions on public policy issues. Moreover, they only have one or two questions related to other latent quantities, such as confidence in the Supreme Court or

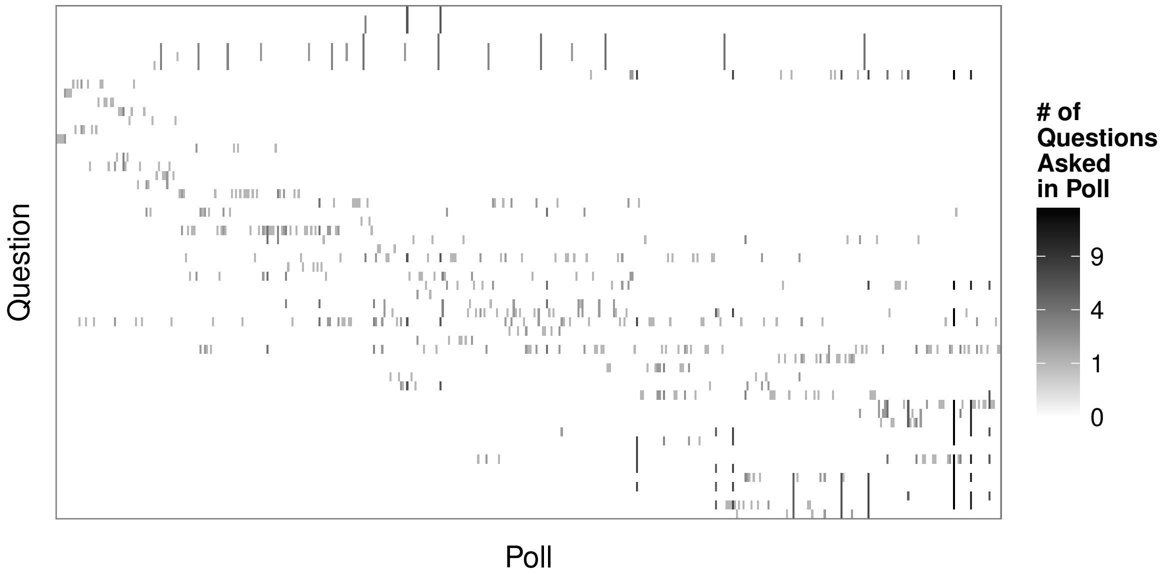

Figure 2: Availability of question series by poll, 1976–2012

political knowledge. The sparseness of the data has made it difficult for political scientists to take full advantage of the wealth of historical public-opinion data that is available.

3. EXISTING APPROACHES TO MODELING PUBLIC OPINION

The measurement model we expound in this paper to overcome these challenges draws upon three important approaches to modeling public opinion: item response theory, multilevel regression 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 lack of constraint in mass issue attitudes noted by Converse (1964) and others, 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 construct (Spearman, 1904; 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 manifest variables.[2] The growing accessibility of Bayesian simulation methods has further increased the range of IRT models, allowing easy characterization of the uncertainty around any parameter estimates or functions thereof (Jackman, 2000).[3]

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

[2] If multivariate normality is not a reasonable approximation, as in the case of binary response variables, "conventional factor analysis can produce biased preference estimates" (Treier and Hillygus, 2009, 684). For a comparison of the utility models underlying factor analysis and ideal-point estimation, see Brady (1990).

[3] By 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.

The second methodological approach we draw upon in this paper is multilevel regression and poststratification (Gelman and Little, 1997; Park, Gelman and Bafumi, 2004). 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 ones based on alternative methods, such as aggregation across polls (Lax and Phillips, 2009*b*; Warshaw and Rodden, 2012; but see Buttice and Highton, 2013 for a cautionary note).

MRP was originally developed to estimate average opinion on particular questions (e.g., support for gay marriage, as in 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, can only be applied under quite specific conditions. First, it requires that each individual survey respondent be asked a large number of issue questions. This means that it would not be applicable to earlier eras when each survey tended to include at most a handful of policy questions. Second, it requires substantial computational resources to estimate the latent ideology of hundreds of thousands of individuals. Finally, Tausanovitch and Warshaw's approach does not directly model changing public opinion over time. The method proposed in this paper 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 upon is the literature 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 (1989) 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). Examples include Martin and Quinn's (2002) dynamic estimation of the ideal points of Supreme Court justices, Jackman's (2005) dynamic model of vote intention over the course of a campaign, and Linzer's (2013) state-level presidential forecasting model. Rather than interpreting DLMs as structural models whose parameters are of interest in themselves, these Bayesian applications tend to use them as convenient means of smoothing estimates over time.

Our work is also closely related to the literature on "public policy mood" that originated with Stimson (1991). Works in this tradition use Stimson's Dyad Ratios algorithm to estimate changes in public preferences for government activity (i.e., left-liberalism). Mood is an inherently aggregate concept, and most studies of mood are concerned only with change over time, not cross-sectional differences (e.g., Bartle, Dellepiane-Avellaneda and Stimson, 2011, 267).[4] Recently, however, Enns and Koch (2013) have combined the Dyad Ratios algorithm with MRP to generate state-level estimates of policy mood.

As McGann (2014) observes, however, the Dyad Ratios algorithm has several unappeal-

[4] See 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 micro-level and aggregate measures of mood are very highly correlated across time.

ing features, most notably its ideological asymmetry and its lack of a 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 4.[5] 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. By contrast, our dynamic group-level IRT model, derived in the following section, accommodates cross-sectional and over-time variation within a common framework.

4. A DYNAMIC HIERARCHICAL GROUP-LEVEL IRT MODEL

In this section, we describe our dynamic measurement model. 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. To understand the logic of the model, it is helpful to derive it step by step, beginning with the group-level IRT model.

4.1. *Group-Level IRT Model*

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

[5] We derived our model independently of McGann (2014), building instead on Mislevy's (1983) earlier work on IRT models for grouped data.

j, and an error term (e_{ij}), where

$$y_{ij} = \begin{cases} 1, & \text{if } \beta_j\theta_i - \alpha_j + \epsilon_{ij} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

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:

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

where Φ is the standard normal CDF (Jackman, 2009, 455; 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 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 impossible. We rely particularly on the work of 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 (for a recent application in political science, see McGann, 2014). The essential idea is to model the θ_i in group g as distributed normally around the group mean $\bar{\theta}_g$ and then marginalize over the distribution of abilities.[6]

[6] For evidence that voter preferences are distributed normally (though not necessarily homoskedastically) within states and congressional districts, see Kernell (2009). Since our real interest is estimating the *average* opinion in different demographic groups, the individual abilities are mere nuisance parameters for us. 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 least a half-dozen responses per individual, whereas

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

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

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.[7] 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

$$p_{gj} = \Phi[(\bar{\theta}_g - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}], \tag{4}$$

where $\bar{\theta}_g$ 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). See Appendix A for a formal derivation of Equation 4.

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 question j out of the n_{gj} responses of individuals in group g (e.g., Ghitza and Gelman, 2013). 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. In Section 4.4, we relax the assumption

respondents are often asked only a single question in the survey data we use.

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

that each respondent answers only one question.

4.2. *Hierarchical Model for Group Means*

As stated in Equation 4, the group-level IRT model generates estimates of the average ability in each group. The number of groups whose opinion can be estimated using this model, however, is very limited due to the sparseness of the 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.[8] Letting $\bar{\boldsymbol{\theta}}$ indicate the vector of group means, ξ an intercept common to all groups, $\mathbf{X}$ the matrix of observed group characteristics, $\boldsymbol{\gamma}$ the vector of hierarchical coefficients, and $\sigma_{\bar{\theta}}$ the standard deviation of group means, the hierarchical linear model for the group means can be written as:

$$\bar{\boldsymbol{\theta}} \sim \mathcal{N}(\xi + \mathbf{X}\boldsymbol{\gamma},\ \sigma^2_{\bar{\theta}}). \qquad (5)$$

If we let it vary by time period, the intercept ξ_t captures opinion dynamics that are common

[8] See Gelman and Little (1997) on multilevel modeling as a solution to sparse opinion data, and Bailey (2001) and Fox and Glas (2001) on hierarchical IRT models.

to all units. We parameterize ξ_t separately to emphasize the tendency for public opinion in different groups to move in parallel according to the national “mood” (Stimson, 1991; Page and Shapiro, 1992).

The matrix $\mathbf{X}$ 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 each racial and gender category. To the extent that there are many members of group g in the data, the estimate of $\bar{\theta}_g$ will be dominated by the likelihood. In the opposite case of an empty cell, $\bar{\theta}_g$ will be shrunk all the way to the linear predictor. In this sense, the hierarchical model functions as an imputation model for individuals 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.

4.3. *Dynamic Model for Hierarchical Parameters*

To estimate opinion change across time, we could simply estimate Equation 5 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 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 5 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). A natural way to do this is via a dynamic linear model (DLM), the Bayesian analogue to the frequentist Kalman filter (Harvey, 1989; West and Harrison, 1997; Jackman, 2009, 471–2).[9] 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, 303–04).

DLMs can be very complex, but a simple model generally suffices for the purpose of pooling information across time. The simplest DLM is a *local-level* model, which the parameter estimate for a given period serves as the prior expected value for the parameter in the following period (a so-called "random walk" prior). We use the following local-level model for the intercept ξ_t:

$$\xi_t \sim \mathcal{N}(\xi_{t-1},\ \sigma^2_\gamma). \tag{6}$$

The innovation variance σ^2_γ determines the weight of the data in period t relative to $t-1$.[10] If there are no new data in period t, then the transition model in Equation 6 acts as a predictive

[9] Political scientists have applied Bayesian DLMs to such problems as estimating the ideal points of Supreme Court justices (Martin and Quinn, 2002), vote intention over the course of a campaign (Jackman, 2005; Linzer, 2013), presidential approval (Beck, Jackman and Rosenthal, 2006), and opinion on individual issues (Voeten and Brewer, 2006; Shirley and Gelman, 2014). For examples of political science applications of the Kalman filter, see Beck (1989), Green, Gerber and De Boef (1999), Green, Palmquist and Schickler (2004), and Baum and Kernell (2001).

[10] 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.

model, imputing an 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:

$$\boldsymbol{\gamma}_t^{\text{demo}} \sim \mathcal{N}(\boldsymbol{\gamma}_{t-1}^{\text{demo}},\ \sigma_\gamma^2). \tag{7}$$

For coefficients corresponding to *geographic* predictors such as state, we write the model more generally so as to permit the optional inclusion of geographic-level covariates (for an exposition of this model, see Jackman, 2009, 471–72). Including geographic-level covariates pools information cross-sectionally among demographically similar geographic units, which can improve the efficiency of geographic effect estimates (Park, Gelman and Bafumi, 2004). Let $\boldsymbol{\gamma}_t^{\text{geo}}$ denote a vector of S geographic effects (e.g., state-specific intercepts), and let $\mathbf{Z}_t$ be an $S \times H$ matrix of corresponding to H geographic-level covariates (e.g., *State Proportion Evangelical/Mormon*). The transition equation for $\boldsymbol{\gamma}_t^{\text{geo}}$ is

$$\boldsymbol{\gamma}_t^{\text{geo}} \sim \mathcal{N}(\boldsymbol{\gamma}_{t-1}^{\text{geo}}\delta_t + \mathbf{Z}_t\boldsymbol{\eta}_t,\ \sigma_\gamma^2), \tag{8}$$

where δ_t is a scalar and $\boldsymbol{\eta}_t$ an H-vector of coefficients. Thus the $\boldsymbol{\gamma}_t^{\text{geo}}$ are modeled as a weighted combination of their value in the previous period ($\boldsymbol{\gamma}_{t-1}^{\text{geo}}$) and the attributes contained in $\mathbf{Z}_t$, with weights δ_t and $\boldsymbol{\eta}_t$, respectively. To the extent that geographic effects are stable and well-estimated, 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 8 (δ_t and $\boldsymbol{\eta}_t$), the standard deviation of group means ($\sigma_{\bar{\theta},t}$), and the standard deviation of abilities within groups ($\sigma_{\theta,t}$). We model the standard deviations on the log scale, as in:

$$\sigma_{\theta,t} \sim \ \ln\mathcal{N}(\ln(\sigma_{\theta,t-1}),\ \sigma_{\sigma}^2), \tag{9}$$

where $\ln\mathcal{N}$ indicates the lognormal distribution and the innovation variance σ_{σ}^2 is a parameter to be estimated.

4.4. *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 post-stratification. 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. First, we estimate a "design effect" d_{gt} for each group g and period t:

$$d_{gt} = 1 + (\frac{\text{sd}_{gt}(w_{i[gt]})}{\text{ave}_{gt}(w_{i[gt]})})^2, \tag{10}$$

where $w_{i[gt]}$ is the sampling weight of individual i, and the average and standard deviation are taken across respondents in group g in period t (see Ghitza and Gelman, 2013 for a similar design effect calculation). Then, we calculate adjusted sample sizes $n^*_{gjt} \leq n_{gjt}$, using the formula

$$n^*_{gjt} = \lceil \sum_{i=1}^{n_{gjt}} \frac{1}{r_{i[gt]} d_{gt}} \rceil, \tag{11}$$

where $r_{i[gt]}$ is the number of questions respondent i answered and $\lceil \cdot \rceil$ is the ceiling function.[11] If each individual in group g answers one question ($r_{i[gt]} = 1, \forall i$) and there is no within-group variation in weights ($d_{gt} = 1$), then $n^*_{gjt} = \sum_{i=1}^{n_{gjt}} 1 = n_{gjt}$. However, the adjusted sample size decreases to the extent that there are multiple responses per individual or variation in weights. Next, we take the weighted mean of each group's responses to item j:

$$\bar{y}^*_{gjt} = \sum_{i=1}^{n_{gjt}} \frac{w_{i[gt]} y_{i[g]jt}}{r_{i[gt]}} / \sum_{i=1}^{n_{gjt}} \frac{w_{i[gt]}}{r_{i[gt]}}. \tag{12}$$

Finally, we replace the raw sums s_{gjt} with the weighted sums $s^*_{gjt} = [n^*_{gjt} \bar{y}^*_{gjt}]$, where $[\cdot]$ is the nearest integer function.[12]

As can be seen in Equation 12, each response $y_{i[g]jt}$ is weighted by i's sampling 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 $\bar{\theta}_{gt}$ is determined by their sampling 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

[11] 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.

[12] Here our approach departs from the formula reported by Ghitza and Gelman (2013, 765), who instead write $s^*_{gjt} = n_{gjt} \bar{y}^*_{gjt}$. In personal correspondence with us, Ghitza and Gelman confirmed that the $n_{gjt} \bar{y}^*_{gjt}$ in the paper is a typo and should be $n^*_{gjt} \bar{y}^*_{gjt}$ instead (Gelman, 2013).

the group.

4.5. *The Full Model*

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

$$s^*_{gjt} = \text{Binomial}(n^*_{gjt},\ p_{gjt}), \tag{13}$$

where

$$p_{gjt} = \Phi[(\bar{\theta}_{gt} - \kappa_j)/\sqrt{\sigma^2_{\theta,t} + \sigma^2_j}]. \tag{14}$$

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

$$\bar{\boldsymbol{\theta}}_t \sim \mathcal{N}(\xi_t + \mathbf{X_t}\boldsymbol{\gamma}_t,\ \sigma^2_{\bar{\theta},t}). \tag{15}$$

Note that 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 has the benefit of bridging the model across time, allowing latent opinion estimates in different periods to be compared on a common metric. In many contexts, however, in may make sense to relax this constraint, a possibility discussed in Section 6.1.

4.6. *Identification, Priors, and Estimation*

IRT models must be identified using 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 code all question responses to have the same polarity (e.g., higher values as 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: $\tilde{\alpha}_j^{(m)} = \alpha_j^{(m)} - J^{-1}\sum_{j=1}^{J}\alpha_j^{(m)}$. 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.[13] We estimated the model using the program Stan, as called from R (Stan Development Team, 2013; R Core Team, 2013).[14]

[13] The first-period priors for all standard deviation parameters are 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 $\delta_{t=1}$ and $\boldsymbol{\eta}_{t=1}$, which are modeled more informatively as $\mathcal{N}(0.5,\ 1)$ and $\mathcal{N}(0,\ 10)$ 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.

[14] 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, 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. Computation time depends on the number of groups, items, and time periods; run times for the models reported in this paper ranged between a day and several weeks.

4.7. *Weighting Group Means to Estimate Geographic Opinion*

The estimates of the yearly group means $\bar{\theta}_{gt}$ 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 population targets can substantially improve estimates of average opinion in states, districts, and other geographic units. Poststratification weighting may be used if the joint population distribution of the variables that define groups is known, but other methods such as raking may be applied if for some variables only the marginal distributions are available (Lumley, 2010, 135–54).

A major advantage of simulation-based estimation is that it facilitates proper accounting for uncertainty in 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, standard deviations, and so on of the posterior distribution. We adopt this approach in presenting the results of the model in the application that follows.

5. 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 more 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?).

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 (Free and Cantril, 1967; 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 47 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 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.[15] 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). In order 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 (e.g., Park, Gelman and Bafumi, 2004, 2006). 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 3 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 Mas-

[15] 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.

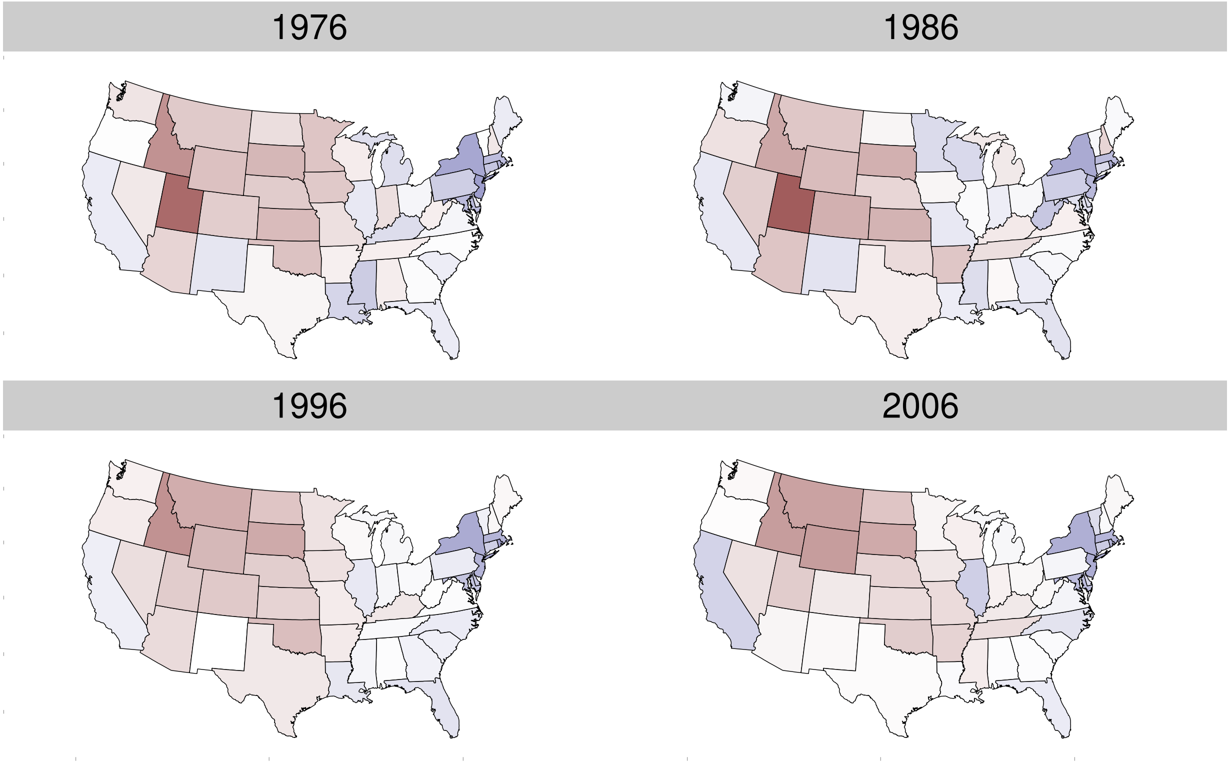

Figure 3: Average state policy liberalism, 1976–2006. The estimates have been re-centered and standardized in each year to accentuate the color contrasts.

sachusetts are always among the most liberal states. Moreover, Figure 3 confirms that the states have remained generally stable in their relative liberalism, consistent with Erikson, Wright and McIver's (2006; 2007) finding that state publics have been stable in terms of ideological identification. According to our estimates, only a few states' policy liberalism has 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.

5.1. *Cross-Validation*

The use of multilevel modeling to smooth subnational opinion estimates across cross-sectional units has been well validated (Lax and Phillips, 2009*b*; 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 $\bar{\theta}_{gt}$ 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).[16] 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 created a training dataset.[17] We used the training data to estimate both the pooled model and the separated model. Based on

[16] 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.

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

the parameter estimates from each model, we calculated the predicted proportion of liberal responses to each item in each group-year:

$$\hat{p}_{gjt} = \Phi[(\hat{\bar{\theta}}_{gt} - \hat{\kappa}_j)/\sqrt{\hat{\sigma}^2_{\theta,t} + \hat{\sigma}^2_j}]. \tag{16}$$

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:

$$\hat{e}_{gjt} = \frac{s^*_{gjt}}{n^*_{gjt}} - \hat{p}_{gjt}. \tag{17}$$

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 10 times, thus producing 10 out-of-sample estimates of bias, MAE, and RMSE for each model.

As Table 5.1 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

	Separated Model			Pooled Model			Diff. 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

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

estimates in the separated model are shrunk as well, but towards the cross-sectional mean rather than towards 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 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.

5.2. *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 only partially speaks 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 (Ansolabehere, Snyder and Stewart, 2001; Canes-Wrone, Brady and Cogan, 2002). 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 4 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.[18] 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).

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

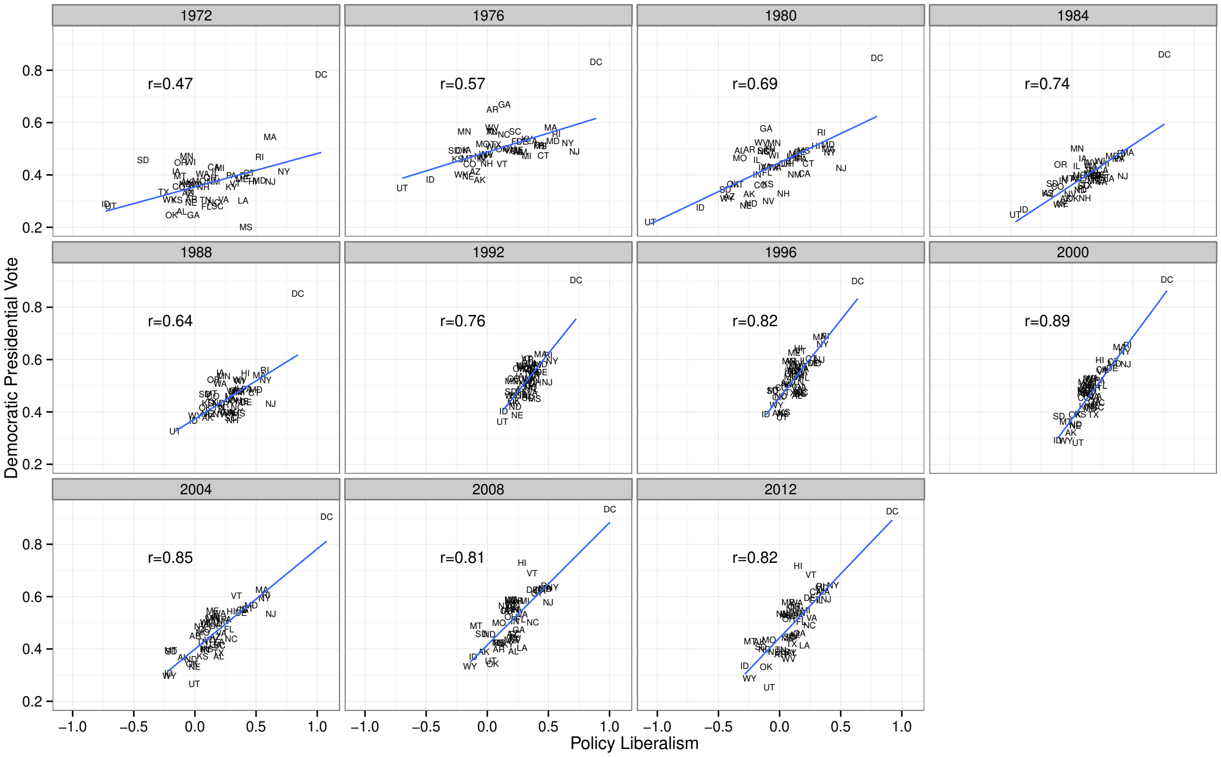

Figure 4: Relationship between policy liberalism and Democratic presidential vote share, 1972–2012.

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 could 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 turn to Stimson's "public policy mood", which is explicitly designed to measure changes in the mass public's policy preferences over time (Stimson, 1991).[19] Of course, we should not expect a perfect correlation between policy liberalism and mood since they are measuring different concepts.

[19] Enns and Koch (2013) use a similar validation strategy for their measures of state-level mood.

Mood is focused on whether the government should be doing “more” or “less” *than it currently is* (e.g., Stimson, 2012, 31). In contrast, our measure of policy liberalism is an absolute measure of the public’s preferences for government spending or activity that is not explicitly tied to the status quo.

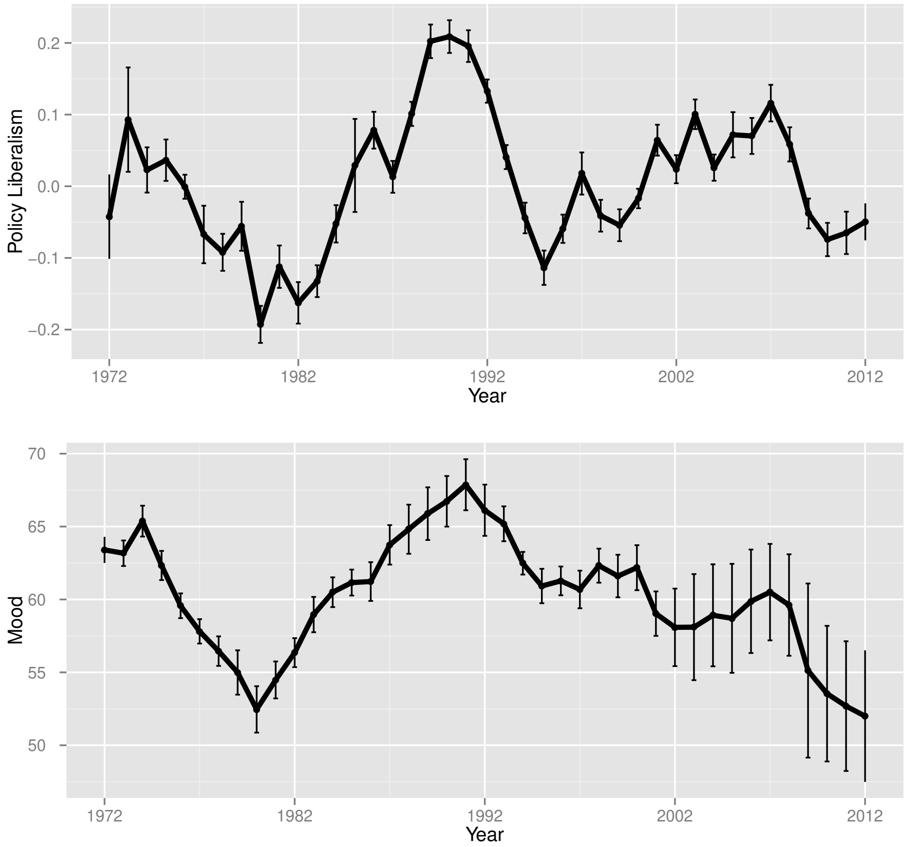

Figure 5: Relationship between national policy liberalism and policy mood, 1972–2012.

Despite these theoretical differences between policy liberalism and Stimson’s 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 2008. The only major divergence between the two scales is in the early 2000s. However, note that Stimson’s mood estimates are quite inefficiently 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.

6. POTENTIAL EXTENSIONS TO THE MODEL

An advantage of our framework is that our model can (and should) be modified to suit particular analytic purposes. Here, we consider three such possibilities: time-varying item parameters, heterogeneous within-group variances, and a multidimensional latent space. We sketch ways of implementing these extensions to our model and describe applications where they might be useful.

6.1. *Time-Varying Item Parameters*

One possible extension to the model would be to allow the item parameters for each question to evolve over time. The assumption of a constant mapping between the latent θ space and the response probabilities is very useful because it justifies the comparability of estimates over time.[20] For certain items, however, it is clearly implausible, especially with regard to the difficulty parameter α_j.[21]

[20] Other dynamic IRT models, notably Martin and Quinn (2002), achieve comparability over time via a different route. In their Supreme Court application, no cases are repeated and so are not available to bridge across time. Rather, their estimates are comparable across time under the random-walk assumption for the innovation of ideal points and the prior distributions for the item parameters. Other approaches, such as Poole and Rosenthal's (2007) DW-NOMINATE, bridge by constraining ideal-point change to be a polynomial function in time. For an example of cross-period bridging using repeated items, see Asmussen and Jo (2011).

[21] Questions gauging support for gay marriage are an obvious example. While these questions may discriminate well between liberals and conservatives at any point in time, the long-term liberal trend on these

One possibility is to specify a local-level transition model for $\alpha_{j,t}$:

$$\alpha_{j,t} \sim \mathcal{N}(\alpha_{j,t-1},\ \sigma^2_\alpha). \tag{18}$$

Based on our experimentation with this model, we have found that it helps to identify the model if the difficulty innovation variance σ^2_α is defined in terms of σ^2_γ, as in $\sigma_\alpha = {}^{\sigma_\gamma}/_{10}$. Substantively, the ratio of σ_α to σ_γ encodes prior beliefs about the magnitude of item-specific change across periods relative to aggregate change in $\bar{\theta}_{gt}$. The downside of allowing $\alpha_{j,t}$ to vary by period is a substantial increase in the number of parameters and as well as in the computational burden of the model. In addition, by altering the mapping between manifest responses and latent opinion, the evolution of item difficulties also complicates the interpretation of opinion estimates from different periods. Whether these additional complexities are worthwhile depends on the application.

6.2. *Heteroskedasticity Across Groups*

As defined in this paper, our model allows $\bar{\theta}_{gt}$ to vary across groups and time but constrains the distributions of $\theta_{i[gt]}$ within groups to be homoskedastic within each period ($\sigma_{\theta,gt} = \sigma_{\theta,t}, \forall g$). This may be misleading if some demographic groups are more heterogeneous than others. For example, in many states African Americans may have more homogenous political preferences than other racial groups, particularly if whites and Hispanics

questions is not shared by other policy questions. Rather, this issue appears to be governed by idiosyncratic long-term dynamics. Ideally, one would want to account for such issue-specific trends while still allowing the issue to inform cross-sectional differences as well as short-term fluctuations in liberalism.

are categorized together, as they often are. It is also possible that, over time, demographic groups may become more or less internally diverse. Heterskedasticity of either form may be accommodated by allowing $\sigma_{\theta,t}$ to vary across groups as well as time.[22]

The simplest heteroskedastic specification of the model would simply estimate group-specific values of $\sigma_{\theta,t}$. However, for the same reasons that it makes sense to model $\bar{\theta}_{gt}$ as a function of group covariates, it may also be advantageous to model the $\sigma_{\theta,gt}$. One possible approach is a variance-function regression, which models the variance of the error term as a function of covariates, possibly the same ones as used to model the mean (Park, 1966; see Western and Bloome, 2009 for a Bayesian implementation). One common specification is a log-normal regression. So, for example, the vector of within-group variances of θ_i could be modeled as

$$\boldsymbol{\sigma}_{\theta}^{2} \sim \ln\mathcal{N}(\mathbf{X}\boldsymbol{\lambda},\ \sigma_{\sigma_\theta}^{2}), \tag{19}$$

where $\mathbf{X}$ is a matrix of group characteristics (including an intercept), $\boldsymbol{\lambda}$ is a vector of coefficients, and $\sigma_{\sigma_\theta}^{2}$ is the prior variance of $\boldsymbol{\sigma}_{\theta}^{2}$ on the log scale. In addition to potentially providing a better fit to the data, the group variance vector $\boldsymbol{\sigma}_{\theta}^{2}$ might be of substantive interest for its own sake.

[22] On a side note, it would also be possible to allow the variance of the response-level error term e_{ij} (Equation 1), which currently has a standard normal distribution, to vary across groups. This would be equivalent to the approach of Jessee (2010) and Lauderdale (2010), who use such heteroskedasticity to allow for some individuals to behave more "spatially" than others. While the specification would be slightly different, both their approach and the one outlined above would have a similar effect of inflating the denominator of the group-level IRT model (Equation 14) with an additional variance component.

6.3. *Multidimensionality*

A third natural extension to the model would be to allow for multiple latent dimensions. The question of whether the issue attitudes of the mass public are best modeled with one or multiple dimensions, or possibly none, is an old one and not easily resolved. Between the extremes of unidimensionality (e.g., Jessee, 2009; Tausanovitch and Warshaw, 2013) and little structure at all (e.g., Converse, 1964) lie studies that identify two or three latent dimensions (e.g., Poole, 1998; Peress, 2013). One issue with these multidimensional findings is that secondary dimensions often lack substantive interpretation and do not always correspond to the typical classification of questions into economic, social, and other issue domains (Ellis and Stimson, 2012; cf. Miller and Stokes, 1963; Ansolabehere, Rodden and Snyder, 2008; Treier and Hillygus, 2009).

Adding a second or even a third dimension to the group-level IRT model might shed new light on this long-standing debate. However, it is also likely to exacerbate the computational complexity of the IRT model by greatly increasing the number of parameters (which is approximately proportional to the number of dimensions) as well as the difficulty of identifying the model and mixing through the posterior distributions. As a result, successful estimation of a multidimensional model might require that the model be simplified in other ways (e.g., with groups defined only by state).

7. 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 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 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 is generally quite sparse. Our approach enables scholar to contract 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. 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.

REFERENCES

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

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.

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.

Asmussen, Nicole and Jinhee Jo. 2011. "Anchors Away: A New Approach for Estimating Ideal Points Comparable Across Time and Chambers." Unpublished manuscript. Available for download at http://my.vanderbilt.edu/nicoleasmussen/files/2011/08/Anchors-Away-updated-March-28-2011.pdf.

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.

Bartle, John, Sebastian Dellepiane-Avellaneda and James Stimson. 2011. "The Moving Centre: Preferences for Government Activity in Britain, 1950–2005." *British Journal of Political Science* 41(2):259–285.

Baum, Lawrence. 2009. *Judges and Their Audiences: A Perspective on Judicial Behavior.* Princeton, NJ: Princeton University Press.

Baum, Matthew A. and Samuel Kernell. 2001. "Economic Class and Popular Support for Franklin Roosevelt in War and Peace." *Public Opinion Quarterly* 65(2):198–229.

Beck, Nathaniel. 1989. "Estimating Dynamic Models using Kalman Filtering." *Political Analysis* 1(1):121–156.

Beck, Nathaniel, Simon Jackman and Howard Rosenthal. 2006. "Presidential Approval: The Case of George W. Bush." Paper presented at the Summer Meeting of the Society for Political Methodology, University of California, Davis, July 19, 2006.

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70(4):499–529.

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* 42(1):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.

Berry, William D., Richard C. Fording, Evan J. Ringquist, Russell L. Hanson and Carl E. Klarner. 2010. "Measuring Citizen and Government Ideology in the US states: A Reappraisal." *State Politics & Policy Quarterly* 10(2):117–135.

Brady, Henry E. 1990. "Traits versus Issues: Factor versus Ideal-Point Analysis of Candidate Thermometer Ratings." *Political Analysis* 2(1):97–129.

Buttice, Matthew K. and Benjamin Highton. 2013. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21(4):449–467.

Caldeira, Gregory. 1987. "Public Opinion and the U.S. Supreme Court: FDR's Court-Packing Plan." *American Political Science Review* 81(4):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.

Caughey, Devin. 2012. "Congress, Public Opinion, and Representation in the One-Party South, 1930s–1960s." PhD dissertation, UC Berkeley.

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, Jean M. 1987. *Survey Research in the United States: Roots and Emergence.* Berkeley: University of California Press.

Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David E. Apter. London: 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.

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

Enns, Peter K. and Julianna Koch. 2013. "Public Opinion in the U.S. States: 1956 to 2010." *State Politics and Policy Quarterly* 13(3):349–372.

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.

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: Stanford University Press 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* 7(2):141–151.

Fiorina, Morris P. and Samuel J. Abrams. 2008. "Political Polarization in the American Public." *Annual Review of Political Science* 11(1):563–588.

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.

Gelman, Andrew. 2007. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1(3):515–533.

Gelman, Andrew. 2013. "Typo in Ghitza and Gelman MRP paper." *Statistical Modeling,* *Causal Inference, and Social Science* (blog), November 9, 2013, http://andrewgelman.com/2013/11/09/typo-ghitza-gelman-mrp-paper.

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.

Gelman, Andrew and Thomas C. Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23(2):127–135.

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):762–776.

Green, Donald P., Alan S. Gerber and Suzanna De Boef. 1999. "Tracking Opinion over Time: A Method for Reducing Sampling Error." *Public Opinion Quarterly* 63:178–192.

Green, Donald P., Bradley Palmquist and Eric Schickler. 2004. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven, CT: Yale University Press.

Harvey, Andrew C. 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge University Press.

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. Forthcoming. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *Journal of Machine Learning Research* .

Jackman, Simon. 2000. “Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation.” *Political Analysis* 8(4):307–332.

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.* Hoboken, NJ: Wiley.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Springer (PDF ebook).

Jessee, Stephen A. 2009. “Spatial Voting in the 2004 Presidential Election.” *American Political Science Review* 103(1):59–81.

Jessee, Stephen A. 2010. “Partisan Bias, Political Information and Spatial Voting in the 2008 Presidential Election.” *Journal of Politics* 72(2):327.

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.

Kernell, Georgia. 2009. “Giving Order to Districts: Estimating Voter Distributions with National Election Returns.” *Political Analysis* 17(3):215–235.

Key, Jr., V. O. 1984[1949]. *Southern Politics in State and Nation.* Knoxville: University of Tennessee Press.

Ladd, Everett Carll and Charles D. Hadley. 1975. *Transformations of the American party system: Political Coalitions from the New Deal to the 1970s.* New York: Norton.

Lauderdale, Benjamin E. 2010. "Unpredictable Voters in Ideal Point Estimation." *Political Analysis* 18:151–171.

Lax, Jeffrey R. and Justin H. Phillips. 2009*a*. "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. 2009*b*. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53(1):107–121.

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.

Lord, Frederic M. and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lumley, Thomas S. 2010. *Complex Surveys: A Guide to Analysis Using R.* Hoboken, NJ: Wiley (PDF ebook).

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.

McGann, Anthony J. 2014. "Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm." *Political Analysis* 22(1):115–129.

Mickey, Robert W. 2014. *Paths Out of Dixie: The Democratization of Authoritarian Enclaves in America's Deep South.* Princeton, NJ: Princeton UP.

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.

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):1040–1054.

Park, R. E. 1966. "Estimation with Heteroscedastic Error Terms." *Econometrica* 34(4):888.

Peress, Michael. 2013. "Candidate Positioning and Responsiveness to Constituent Opinion in the U.S. House of Representatives." *Public Choice* 156(1-2):77–94.

Persily, Nathaniel, Jack Citrin and Patrick J. Egan, eds. 2008. *Public Opinion and Constitutional Controversy*. New York: Oxford University Press.

Poole, Keith T. 1998. "Recovering a Basic Space From a Set of Issue Scales." *American Journal of Political Science* 42(3).

Poole, Keith T. and Howard Rosenthal. 2007. *Ideology & Congress*. New Brunswick, NJ: Transaction Publishers.

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, 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.

Shirley, Kenneth E. and Andrew Gelman. 2014. "Hierarchical Models for Estimating State and Demographic Trends in US Death Penalty Public Opinion." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* .

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3):530–51.

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15(1):72–101.

Stan Development Team. 2013. "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.

Stimson, James A. 1999. *Public Opinion in America: Moods, Cycles, and Swings.* 2nd ed. Boulder, CO: 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.

Stimson, James A. 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):330–342.

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

Voeten, Erik and Paul R Brewer. 2006. "Public Opinion, the War in Iraq, and Presidential Accountability." *Journal of Conflict Resolution* 50(6):809–830.

Warshaw, Christopher and Jonathan Rodden. 2012. "How Should We Measure District-Level Public Opinion on Individual Issues?" *Journal of Politics* 74(1):203–219.

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):526–546.

West, Mike and Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models.* New York: Springer.

Western, Bruce and Deirdre Bloome. 2009. “Variance Function Regressions for Studying Inequality.” *Sociological Methodology* 39(1):293–326.

A. DERIVATION OF GROUP-LEVEL NORMAL OGIVE IRT MODEL

This appendix derives the 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: $\theta_{ig} \sim \mathcal{N}(\bar{\theta}_g,\ \sigma_\theta^2)$. 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:

$$(\theta_{ig} - \kappa_j)/\sigma_j + \epsilon_{ij} > 0 \tag{20}$$

Multiplying by σ_j, the inequality in Equation 20 becomes:

$$\theta_{ig} - \kappa_j + \epsilon_{ij}\sigma_j > 0 \tag{21}$$

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:

$$\Pr[y_{igj} = 1] = \Pr[z_{igj} > 0] \tag{22}$$

By Assumption 3, the individual abilities θ_{ig} are distributed $\mathcal{N}(\bar{\theta}_g,\ \sigma_\theta^2)$. 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:

$$z_{igj} \sim \mathcal{N}(\bar{\theta}_g - \kappa_j,\ \sigma_\theta^2 + \sigma_j^2) \tag{23}$$

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:

$$\Pr[z_{igj} \leq x] = \Phi[\frac{x - (\bar{\theta}_g - \kappa_j)}{\sqrt{\sigma_\theta^2 + \sigma_j^2}}] \tag{24}$$

which implies:

$$\begin{aligned}
\Pr[z_{igj} > 0] &= 1 - \Phi[\frac{0 - (\bar{\theta}_g - \kappa_j)}{\sqrt{\sigma_\theta^2 + \sigma_j^2}}] \\
&= 1 - \Phi[-(\bar{\theta}_g - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \\
&= \Phi[(\bar{\theta}_g - \kappa_j)/\sqrt{\sigma_\theta^2 + \sigma_j^2}] \\
&= p_{gj} \qquad (25)
\end{aligned}$$

“In other words,” writes Mislevy (1983, 278), “if [κ_j] and σ_j are the item threshold and disperson 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.

B. MASS SUPPORT FOR THE NEW DEAL, 1936–1952

In the mid-1930s, just as Franklin Roosevelt's New Deal program of liberal reform was reaching its peak, commercial survey firms began fielding the first national opinion polls. The advent of systematic polling was thus well-timed to document the shifts in mass opinion that occurred in the wake of this political watershed. By 1952, when the first American National Election Study was fielded, George Gallup and others had conducted hundreds of commercial opinion polls, querying a total of over one million Americans for their opinions on a multitude of political attitudes and topics (Converse, 1987). Recently, a team led by Adam Berinsky and Eric Schickler has cleaned and standardized the data from these early polls, making them much more accessible to political scientists (Berinsky et al., 2011).

Aside from data issues, a major problem with these early polls is that they were collected with quota-sampling techniques that rendered them unrepresentative of the U.S. population. It is therefore desirable to weight the polls to match known population benchmarks, such as the racial and occupational make-up of each state (Berinsky, 2006). Another difficulty is that a given respondents was rarely asked more than a couple of political questions, and few questions were asked in consistent fashion over many polls. These limitations present a substantial challenge to summarizing the enormous amount of information contained in these polls, either at the individual level (in the form of dimension-reduction techniques) or over time (by, say, tracking consistent question series). These difficulties are what first motivated us to develop the dynamic group-level IRT model described in this paper.

The data for this analysis were derived from quota-sampled Gallup polls fielded between November 1936 and December 1952. These polls contain 453 unique question series asked in identical form across time. Three-quarters of the questions were asked in only a single year; just 18 were asked in more than three different years. Only questions related to such New Deal issues as labor unions, taxation, regulation of the economy, and social welfare were included. A total of 644,370 unique respondents are represented in the data. We coded their responses as either favoring or opposing the New Deal, dichotomizing ordinal responses at an appropriate midpoint.

Respondents were grouped into categories defined by *State* and race-by-region variable (*White South* × *Black*) with three levels: black, Southern white, and non-Southern white.[23] Within each group, respondents were poststratified to match the joint distribution of *Female* and *Professional* in the population, and the group totals were weighted accrodingly. Including these variables in the model ameliorates the biases introduced by the severe gender, occupational, racial, and regional discrepancies between the poll samples and the population. Mean support for the New Deal in each group was modeled hierarchically as an linear combination of *State* and *White South* × *Black*. Except in the first year, when the state intercepts were modeled as a function of four-category region, no state-level characteristics were included in the model.

One of the virtues of estimating opinion by group is that the group estimates can be weighted to match whatever the population of interest happens to be. In this case, it is useful to focus not on the U.S. adult population as a whole, but rather on the population minus Southern blacks. We do this for two reasons. First, in this period Southern blacks were almost entirely disfranchised, so they were not part of the potential electorate (Key,

[23] Following Gallup's regional categorization scheme, the South was defined as the eleven states of the former Confederacy plus Kentucky and Oklahoma.

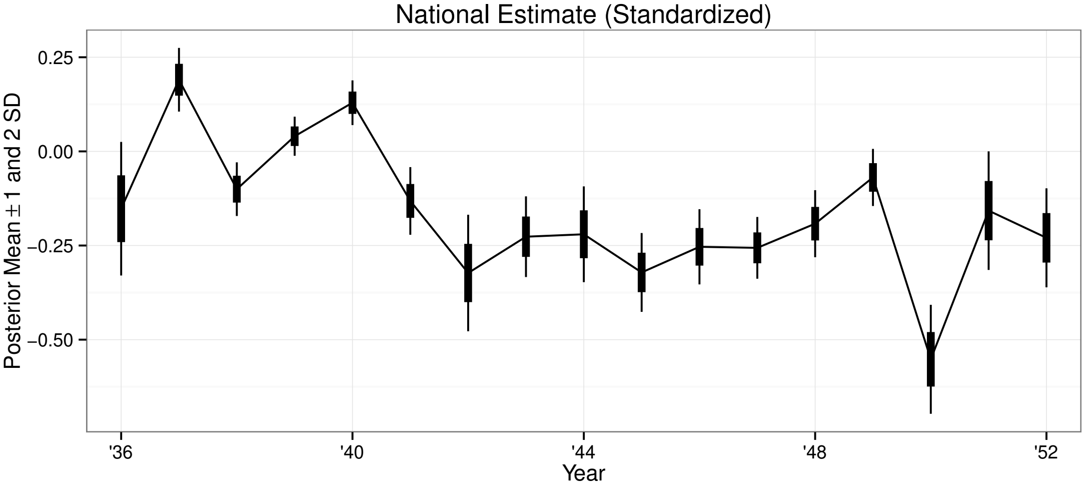

Figure 6: Dynamic group IRT estimates of mean support for the New Deal in the United States, 1936–52. Error bars represent 1 and 2 standard deviations around the mean of the posterior distribution. The estimates have been standardized by the cross-sectional standard deviation of New Deal support in the median year.

1984[1949]). Second, for this reason, blacks were severely undersampled in Southern states, so our estimates for Southern blacks would be extrapolating heavily (via the multilevel model) from the opinions of Northern blacks. Thus, though black respondents from the South were included in the data used to estimate the model, we poststratify the estimated group means to match the population minus Southern blacks, implicitly given them zero weight in our estimates for Southern states. All estimates below are based on this definition of the U.S. population.

Figure 6 plots estimated mean support for the New Deal in the United States between (the last two months of) 1936 and 1952. The figure displays a large and sharp turn against the New Deal that coincided with U.S. mobilization for the Second World War (1941–42). Since the estimates have been scaled by the standard deviation across individuals in a typical year, the figure implies that that the American public moved almost half a standard deviation to the right between 1940 and 1942. Aside from an anomalous deviation in 1950—which

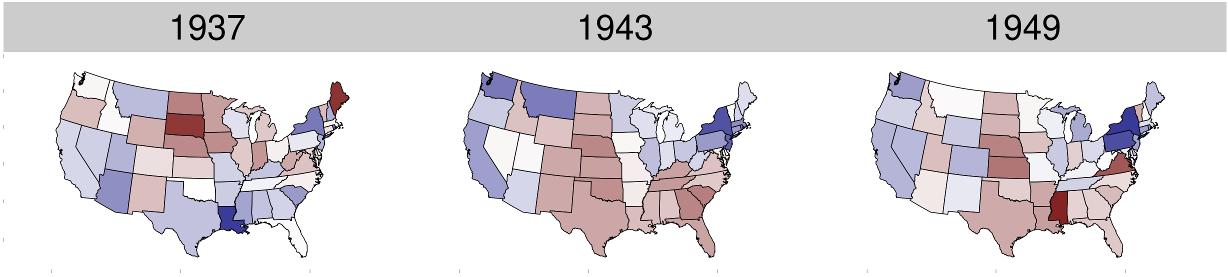

Figure 7: State support for New Deal liberalism in 1937, 1943, and 1949. Estimates have been centered and standardized in each year to accentuate the color contrasts.

may reflect the small number of questions (seven) in that year—national support for the New Deal was relative stable after 1942.

Now consider the cross-sectional state comparisons presented in 7, which require less-stringent assumptions than do the over-time comparisons. These maps reveal a striking realignment of state opinion between 1937 and 1943, as the South transformed from the region most supportive of the New Deal to the most conservative region. While the South's turn against liberalism has been noted by scholars (e.g., Ladd and Hadley, 1975), this is the first time its extent and timing has been documented with any precision.

The state-level opinion estimates also permit examination of the relationship between mass liberalism and the voting records of their representatives in Congress. Figure 8 plots the average first-dimension DW-NOMINATE score of state Senate delegations against mean support for New Deal liberalism in the state publics. Two noteworthy patterns emerge from this graph. First, consistent with Figure 7, the Southern white public began the period more liberal than the rest of the nation but quickly become more conservative than average. This sharp regional shift is an exception to the normal pattern of state ideological stability, at least in survey-based measures (for a debate on this point, see Berry et al., 2007 and Erikson,

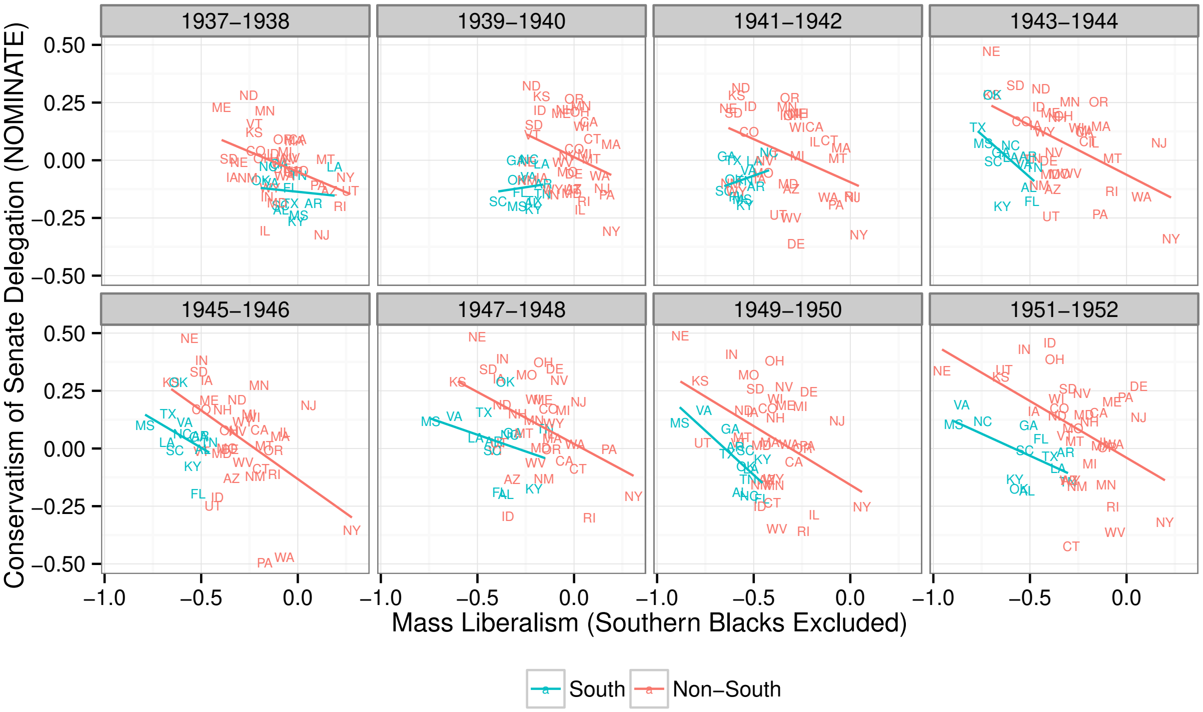

Figure 8: Relationship between mean liberalism of state publics and mean conservatism of state Senate delegations (as measured by first-dimension DW-NOMINATE scores). State opinion estimates exclude Southern blacks, who were disenfranchised at this time. Estimates are pooled within two-year periods corresponding to congressional sessions.

Wright and McIver, 2007).

Second, within each region, mass liberalism and Senate conservatism are quite negatively correlated, especially after 1942. The empirical correspondence between these theoretically related measures provides additional construct validation for our model. The strength of the relationship in the (white) South is somewhat surprising, however, given the absence of partisan competition in the one-party region and the fact that many whites were disfranchised along with blacks (Key, 1984[1949]; Mickey, 2014). But it is consistent with other recent evidence of representatives' responsiveness to the preferences of the potential electorate in the one-party South (Schickler and Caughey, 2011; Caughey, 2012).

C. STATE CONFIDENCE IN THE SUPREME COURT, 1965–2010

Public opinion on the Supreme Court plays a key role in many theories of judicial politics (for an overview, see Persily, Citrin and Egan, 2008). A major theme in this literature is the effect of public confidence in the Supreme Court on the interaction between the Court and other branches. Because the Court is sensitive to how it is perceived by the public (Baum, 2009), 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 explain changes in the public's confidence in the Court over time. Mondak and Smithey (1997) find that the Court's support erodes when its decisions diverge from the ideological preferences of the American public.

Previous empirical work on the role of public opinion in judicial politics has been hampered by the difficulty in measuring confidence in the Court either over time or across states. Clark (2009) writes that "public opinion data about the Court are notoriously sparse" (p. 979). 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.[24] Our model builds upon previous approaches by pooling across survey ques-

[24] Clark (2011) develops better state-level estimates by using a multi-level regression with poststratification (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.

tions and polling firms to estimate latent trust in the Supreme Court at the state-level. Our dynamic model enables us to estimate latent confidence in the Supreme Court even in years with little or no available 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.

We use data from 72 polls between 1963 and 2010 with approximately 166,000 total respondents. We use four question series as indicators of confidence in the Court.[25] Some of these questions have multiple ordinal response categories (e.g., "very favorable", "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.

Figure 9 compares state-level support for the Court across the past five decades. In the early part of the period, there is generally lower support for the Court in the South, which probably reflects Southern whites' dissatisfaction with the Court liberal decisions on school de-segregation and criminal justice. In contrast, there is very strong support for the Court in liberal, northern states during the 1960s and early 1970s. Over time, however, support for

[25] We use the items: 1) Do you approve or disapprove of the way the Supreme Court is handling its job? 2) In general, what kind of rating would you give the Supreme Court? 3) Would you tell me how much respect and confidence you have in the Supreme Court? 4) Is your overall opinion of the Supreme Court very favorable, mostly favorable, mostly unfavorable, or very unfavorable?

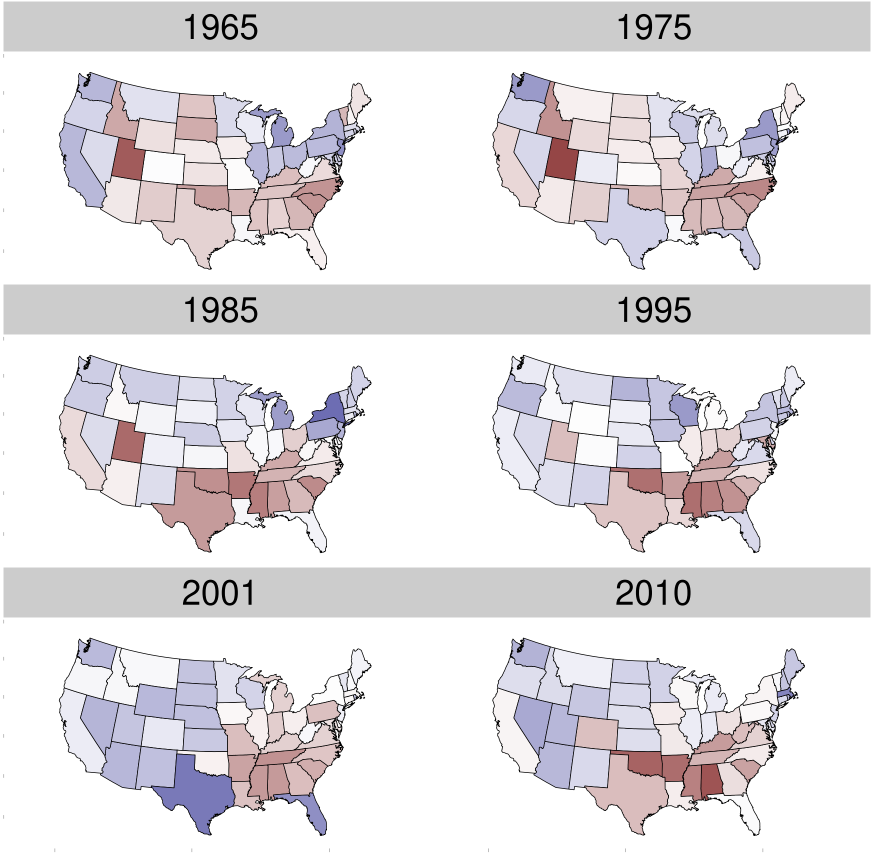

Figure 9: Average state confidence in the Supreme Court, 1965–2010. Blue indicates greater confidence. State estimates have been normalized in each year to highlight cross-sectional differences.

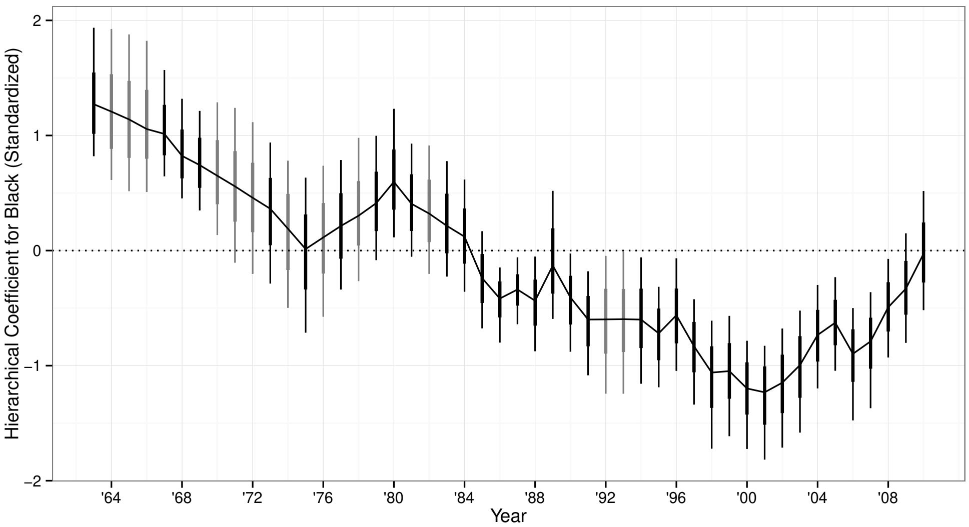

Figure 10: Year-specific estimates of the hierarchical coefficent for the demographic predictor *Black*. The estimates have been standardized by the cross-sectional standard deviation of latent judicial confidence in a typical year. Gray bars indicate years for which no poll data are available.

the Court drops in northern states and rises in southern states. These changes likely reflect the general shift in the Court's orientation to the ideological right.

A different angle on these same phenomenon is provided by Figure 10, which plots the yearly estimated coefficients for *Black* in the hierarchical model. The estimates have been standardized by the cross-sectional standard deviation of latent judicial confidence in a typical year. In 1963, blacks were predicted to be over a standard deviation more confident in the Supreme Court than non-blacks, conditional on their other demographic and geographic characteristics. Black support dropped as the Court became less closely associated with civil rights and more conservative generally. After bottoming out around 2000, blacks' judicial confidence rebounded, especially after the election of Barack Obama in 2008. These shifts in blacks' relative confidence in the Court highlight the importance of allowing the hierarchical model to evolve over time.

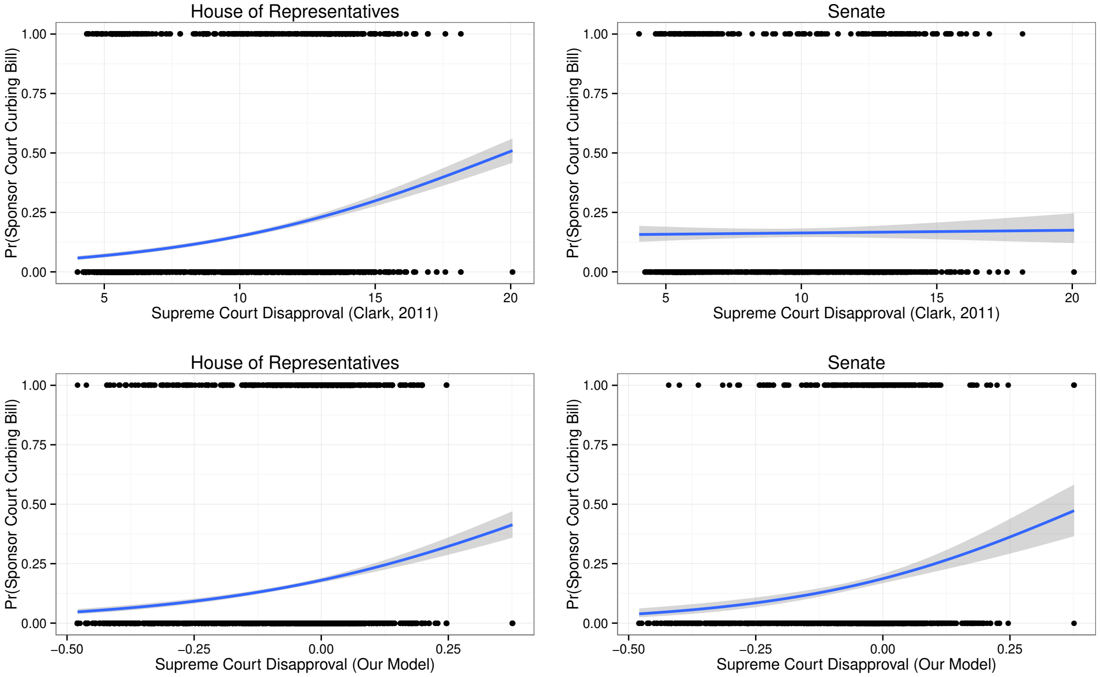

Figure 11: Relationship between probability of sponsoring a court-curbing bill and state public disapproval of the Supreme Court. Top row uses estimates from Clark (2011); bottom row uses our group-level IRT estimates of confidence in the Supreme Court (reverse-coded).

We validate our estimates by using them to predict co-sponsorship of court-curbing bills in Congress. Clark (2011) argues that legislators should be more likely to sponsor court-curbing bills when there is substantial disapproval of the Court in the legislators' constituency. Clark (2011) shows that members of the U.S. House are more likely to sponsor court curbing bills when there is substantial disapproval of the Court in their home state. However, Clark's theoretical logic is actually stronger for the Senate than the House since there should be a closer fit there between state-level estimates of judicial confidence and senators' constituencies.

In the top-row of figure 11, we replicate Clark's results for members of the U.S. House using his MRP-based measure of judicial confidence and co-sponsorships of court curbing bills by representatives. As the top-right panel shows, however, Clark's state-level estimates of Court disapproval are uncorrelated with senatorial support for court-curbing. The bottom row of Figure 11 conducts the same analysis using our measure of state-level confidence in the Court (reverse-coded). Our estimates not only predict House court-curbing as well as Clark's, but they also predict it in the Senate. These results both reinforce the validity of our measure of confidence in the Supreme Court and demonstrate our model's empirical usefulness for studying constructs other than policy preferences.

```
data {
    int<lower=1> G; // number of covariate groups
    int<lower=1> Q; // number of items/questions
    int<lower=1> T; // number of years
    int<lower=1> N; // number of observed cells
    int<lower=1> S; // number of geographic units (e.g., states)
    int<lower=1> P; // number of hierarchical parameters, including geographic
    int<lower=1> H; // number of predictors for geographic unit effects
    int<lower=1> H_prior; // number of predictors for geographic unit effects (t=1)
    int<lower=1> D; // number of difficulty parameters per question
    int<lower=0,upper=1> constant_item; // indicator for constant item parameters
    int<lower=0,upper=1> separate_years; // indicator for no over-time smoothing
    int s_vec[N]; // long vector of responses
    int n_vec[N]; // long vector of counts
    int<lower=0> MMM[T, Q, G]; // missingness array
    matrix<lower=0, upper=1>[G, P] XX; // indicator matrix for hierarchical vars.
    row_vector[H] ZZ[T, S]; // data for geographic model
    row_vector[H_prior] ZZ_prior[1, S]; // data for geographic model
}
transformed data {
}
parameters {
    vector[Q] diff_raw[D]; // raw difficulty
    vector<lower=0>[Q] disc_raw; // discrimination
    vector[T] xi; // national mean (common intercept)
    vector[P] gamma[T]; // hierarchical parameters
    vector[T] delta_lag; // weight placed on geo. effects from prev. period
    vector[H] delta_pred[T]; // weight on geographic predictors
    vector[H_prior] delta_pred_prior; // weight on geographic predictors (t=1)
    vector[G] theta_bar[T]; // group mean ability
    vector<lower=0>[T] sd_theta_bar; // sd of group ability means (by period)
    vector<lower=0>[T] sd_theta; // sd of abilities (by period)
    real<lower=0> sd_geo; // prior sd of geographic effects
    real<lower=0> sd_geo_prior; // prior sd of geographic effects (t=1)
    real<lower=0> sd_demo; // sd of demographic effecs
    real<lower=0> sd_innov_delta; // innovation sd of delta_pred and delta_lag
    real<lower=0> sd_innov_logsd; // innovation sd of sd_theta
    real<lower=0> sd_innov_gamma; // innovation sd of gamma, xi, and (opt.) diff
}
transformed parameters {
    vector[Q] diff[D]; // adjusted difficulty
    vector[Q] kappa[D]; // threshold
    vector<lower=0>[Q] disc; // normalized discrimination
    vector<lower=0>[Q] sd_item; // item standard deviation
    vector<lower=0>[Q] var_item; // item variance
    vector<lower=0>[T] var_theta; // variance of abilities
    vector[G] xb_theta_bar[T]; // linear predictor for group means
    vector[G] z[T, Q]; // array of vectors of group deviates
    real prob[T, Q, G]; // array of probabilities
    // Identify model by rescaling item parameters (Fox 2010, pp. 88-89)
    // scale (product = 1)
    disc <- disc_raw * pow(exp(sum(log(disc_raw))), (-inv(Q)));
    for (q in 1:Q) {
        sd_item[q] <- inv(disc[q]); // item standard deviations
    }
    for (d in 1:D) {
        // location (mean in first year = 0)
        diff[d] <- diff_raw[d] - mean(diff_raw[1]);
        kappa[d] <- diff[d] ./ disc; // item thresholds
    }
    var_item <- sd_item .* sd_item; // item variances
    // Abilities
    var_theta <- sd_theta .* sd_theta; // within-group variances of abilities
    for (t in 1:T) { // loop over years
        xb_theta_bar[t] <- xi[t] + XX * gamma[t]; // Gx1 = GxP * Px1
        for (q in 1:Q) { // loop over questions
            real var_tq; //
            var_tq <- sqrt(var_theta[t] + var_item[q]);
            // Group-level IRT model
            if (constant_item == 0) {
                z[t, q] <- (theta_bar[t] - kappa[t][q]) / var_tq;
            }
            if (constant_item == 1) {
                z[t, q] <- (theta_bar[t] - kappa[1][q]) / var_tq;
            }
            for (g in 1:G) { // loop over groups
                prob[t, q, g] <- Phi_approx(z[t, q, g]); // fast approx. of normal CDF
            } // end group loop
        } // end question loop
    } // end year loop
    // Convert counts and probabilities from array to vector
}
model {
    // TEMPORARY VARIABLES
    real prob_vec[N]; // long vector of probabilities (empty cells omitted)
    int pos;
    pos <- 0;
    // PRIORS
    if (constant_item == 1) {
        diff_raw[1] ~ normal(0, 1); // item difficulty (constant)
    }
    disc_raw ~ lognormal(0, 1); // item discrimination
    sd_geo ~ cauchy(0, 2.5); // sd of geographic effects
    sd_geo_prior ~ cauchy(0, 2.5); // prior sd of geographic effects
    sd_demo ~ cauchy(0, 2.5); // prior sd of demographic parameters
    sd_innov_delta ~ cauchy(0, 2.5); // innovation sd of delta_pred/delta_lag
    sd_innov_gamma ~ cauchy(0, 2.5); // innovation sd. of gamma, xi, and diff
    sd_innov_logsd ~ cauchy(0, 2.5); // innovation sd of theta_sd
    for (t in 1:T) { // loop over years
        if (separate_years == 1) { // Estimate model anew each period
            xi[t] ~ normal(0, 10); // intercept
            for (p in 1:P) { // Loop over individual predictors (gammas)
                if (p <= S) gamma[t][p] ~ normal(ZZ[t][p]*delta_pred[t], sd_geo);
                if (p > S) gamma[t][p] ~ normal(0, sd_demo);
            }
        }
        if (t == 1) {
            if (constant_item == 0) {
                diff_raw[t] ~ normal(0, 1); // item difficulty
            }
            // Priors for first period
            sd_theta_bar[t] ~ cauchy(0, 2.5);
            sd_theta[t] ~ cauchy(0, 2.5);
            delta_lag[t] ~ normal(0.5, 1);
            delta_pred[t] ~ normal(0, 10);
            delta_pred_prior ~ normal(0, 10);
            if (separate_years == 0) {
                xi[t] ~ normal(0, 10); // intercept
                for (p in 1:P) { // Loop over individual predictors (gammas)
                    if (p <= S) {
                        gamma[t][p] ~ normal(ZZ_prior[1][p]*delta_pred_prior,
                                             sd_geo_prior);
                    }
                    if (p > S) gamma[t][p] ~ normal(0, sd_demo);
                }
            }
        }
        if (t > 1) {
            // TRANSITION MODEL
            // Difficulty parameters (if not constant)
            if (constant_item == 0) {
                diff_raw[t] ~ normal(diff_raw[t - 1], sd_innov_gamma);
            }
            // predictors in geographic models (random walk)
            delta_lag[t] ~ normal(delta_lag[t - 1], sd_innov_delta);
            delta_pred[t] ~ normal(delta_pred[t - 1], sd_innov_delta);
            sd_theta_bar[t] ~ lognormal(log(sd_theta_bar[t - 1]), sd_innov_logsd);
            sd_theta[t] ~ lognormal(log(sd_theta[t - 1]), sd_innov_logsd);
            if (separate_years == 0) {
                // Dynamic linear model for hierarchical parameters
                xi[t] ~ normal(xi[t - 1], sd_innov_gamma); // intercept
                for (p in 1:P) { // Loop over individual predictors (gammas)
                    if (p <= S) {
                        gamma[t][p] ~ normal(delta_lag[t]*gamma[t - 1][p] +
                                             ZZ[t][p]*delta_pred[t], sd_innov_gamma);
                    }
                    if (p > S) gamma[t][p] ~ normal(gamma[t - 1][p], sd_innov_gamma);
                }
            }
        }
        // RESPONSE MODEL
        // Model for group means
        // (See ‘transformed parameters’ for definition of xb_theta_bar)
        theta_bar[t] ~ normal(xb_theta_bar[t], sd_theta_bar[t]); // group means
        for (q in 1:Q) { // loop over questions
            for (g in 1:G) { // loop over groups
                if (MMM[t, q, g] == 0) { // Use only if not missing
```

```
                    pos <- pos + 1;
                    prob_vec[pos] <- prob[t, q, g];
                }
            } // end group loop
        } // end question loop
    } // end time loop
    // Model for group responses
    s_vec ~ binomial(n_vec, prob_vec); // fully vectorized
}
generated quantities {
    vector<lower=0>[T] sd_total;
    for (t in 1:T) {
        sd_total[t] <- sqrt(variance(theta_bar[t]) + square(sd_theta[t]));
    }
}
```