

Online Appendix A:

Details of Graded Response Model & Comparison of Aldrich-McKelvey Scaling with Projection Method

Details on the Bayesian Graded Response Model & Convergence Statistics

Under the assumption of local independence of the item-responses, the graded response model can be formulated as

$$\Pr[Y_{ikt} = c | \Xi] = \begin{cases} \Lambda(\eta_{ikt,c}) & \text{if } c = 1, \\ 1 - \Lambda(\eta_{ikt,c-1}) & \text{if } c = C_k, \\ \Lambda(\eta_{ikt,c}) - \Lambda(\eta_{ikt,c-1}) & \text{otherwise} \end{cases} \quad (1)$$

where Y_{ikt} is the response of respondent i on item k in year t , $\Lambda(\cdot)$ is the standard Logistic cumulative distribution function, C_k is the largest response category of item k , and Ξ is the set of parameters in the model. Further

$$\eta_{ikt,c} = \kappa_{k,c} - \boldsymbol{\gamma}_k^\top \boldsymbol{\theta}_{it},$$

with $\boldsymbol{\theta}_{it}$ representing an $D \times 1$ vector of person i 's position in a D -dimensional latent space at time t , $\boldsymbol{\gamma}_k$ a $D \times 1$ vector of discrimination parameters for item k , and $\kappa_{k,c}$ the c th cutpoint of item k , which has a total of $C_k - 1$ cutpoints with $\kappa_{k,c} < \kappa_{k,c+1}$ for $c = \{1, \dots, C_k - 2\}$ if $C_k > 2$.

While the equations look ugly, the GRM might be thought of a ordered logistic regression with unobserved regressors $\boldsymbol{\theta}_{it}$. It is also equivalent to the non-linear factor model (Takane and De Leeuw 1987) or the mixed factor model (Quinn 2004) with only ordinal variables.

Hence, the discrimination parameters are similar to factor loadings, although caution has to be given in the interpretation of the parameters since the number of response categories might differ across items. The cutpoint parameters have no equivalent in linear factor models, but can be thought as having the same interpretation as the cutpoints in ordinal logistic regression except that, in the GRM, the regressors themselves parameters to be estimated.

The model is estimated as follows. Let $\Xi = \{\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Omega}\}$ the set of the parameters in the model. By Bayes rule, the joint posterior distribution of the parameters is proportional to the likelihood times the prior distribution. I assume that the prior distributions are *a priori* independent, except for the latent ideological positions which depend on the year-specific mean and variance of each latent dimension. Thus, abusing notation a little bit,

$$\begin{aligned} p(\Xi | \mathbf{Y}) &\propto p(\mathbf{Y} | \Xi)p(\Xi) \\ &= p(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Omega})p(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Omega})p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\boldsymbol{\Omega})p(\boldsymbol{\gamma})p(\boldsymbol{\kappa}), \end{aligned}$$

where \mathbf{Y} is the matrix of responses, and the other parameters are as above. Given the assumption of local independence, the contribution of a single observation to the likelihood is given in equation (1) as

$$\phi_{ikt}(c) = \Pr[Y_{ikt} = c | \Xi]$$

Hence the likelihood of the data can be expressed as

$$p(\mathbf{Y} | \Xi) = \prod_{i=1}^N \prod_{t=1}^T \prod_{k=1}^K \prod_{c=1}^{C_k} [\phi_{ikt}(c)]^{I_{\{y_{ikt}=c\}}(y_{ikt})},$$

where N and T are, respectively, the number of individuals and number of time points in the data, K is the number of items, $I_{\{y_{ikt}=c\}}(y_{ikt})$ is an indicator function which is 1 if $y_{ikt} = c$ and zero otherwise, and C_k is the number of response categories of item k . I use Normal priors for the latent ideological positions, i.e., $\boldsymbol{\theta}_{it} \sim N_d(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\Sigma}_t = \text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\Omega}_t\text{diag}(\boldsymbol{\sigma}_t)$,

where $\boldsymbol{\mu}_t$ is a $D \times 1$ mean vector, $\text{diag}(\boldsymbol{\sigma}_t)$ is a diagonal $D \times D$ matrix containing the standard deviations, $\{\sigma_{dt}\}$, of each dimension, and $\boldsymbol{\Omega}_t$ is a $D \times D$ correlation matrix. For the other parameters, I use weakly informative priors: namely, $\mu_{dt} \sim N(0, 10^2)$, $\sigma_{dt} \sim U(0, 50)$, $\kappa_{k,c} \sim N(0, 10^2)$, and $\gamma_k \sim N(0, 10^2)$. The off-diagonal elements of $\boldsymbol{\Omega}_t$ —i.e., the correlations between the dimensions—are given non-informative uniform priors over the interval $[-1, 1]$ for all years with the constraint that the resulting matrix has to be symmetric and positive definite. This can be formulated as

$$\begin{aligned}
p(\boldsymbol{\gamma}) &= \prod_{k=1}^K \prod_{d=1}^D p(\gamma_{kd}), & \gamma_{kd} &\sim N(0, 10^2) \quad \forall k, d \\
p(\boldsymbol{\kappa}) &= \prod_{k=1}^K \prod_{c=1}^{C_{k-1}} p(\kappa_{k,c}), & \kappa_{k,c} &\sim N(0, 10^2) \quad \forall k, c \\
p(\boldsymbol{\mu}) &= \prod_{t=1}^T \prod_{d=1}^D p(\mu_{dt}), & \mu_{dt} &\sim N(0, 10^2) \quad \forall d, t \\
p(\boldsymbol{\sigma}) &= \prod_{t=1}^T \prod_{d=1}^D p(\sigma_{dt}), & \sigma_{dt} &\sim U(0, 50) \quad \forall d, t \\
p(\boldsymbol{\Omega}) &= \prod_{t=1}^T p(\boldsymbol{\Omega}_t), & \boldsymbol{\Omega}_t &\sim \text{Ljk}_D(1) \quad \forall t \\
p(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Omega}) &= \prod_{i=1}^N \prod_{t=1}^T p(\theta_{it}), & \boldsymbol{\theta}_{it} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t &\sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad \forall i, t
\end{aligned}$$

where $\boldsymbol{\mu}_t = [\mu_{1t}, \mu_{2t}, \dots, \mu_{Dt}]$, $\boldsymbol{\sigma}_t = [\sigma_{1t}, \sigma_{2t}, \dots, \sigma_{Dt}]$, and $\boldsymbol{\Sigma}_t = \text{diag}(\boldsymbol{\sigma}_t)\boldsymbol{\Omega}_t\text{diag}(\boldsymbol{\sigma}_t)$, and where $\text{Ljk}_D(1)$ denotes the Ljk distribution over a $D \times D$ matrix, with shape parameter equal to 1.¹

The model as such is not identified. For example, subtracting a constant vector \mathbf{a} from all latent positions and simultaneously adding $\boldsymbol{\gamma}_k^\top \mathbf{a}$ to all cut points would not change the likelihood of the model; the likelihood remains also unchanged when all discrimination pa-

¹The Ljk_D distribution with shape parameter one is uniform over all correlation matrices of order D . In other words, it is equivalent to assigning $U(-1, 1)$ priors to off-diagonal elements of $\boldsymbol{\Omega}$ (the correlation parameters) under the constraint that $\boldsymbol{\Omega}$ is symmetric, has diagonal entries of ones, and is positive definite.

rameters are multiplied and the latent positions divided by the same constant. As noted in the main text, I fix the prior mean of each latent dimension in the first year of the analysis to zero—i.e., $\boldsymbol{\mu}_t = \mathbf{0}$ if $t = 1986$ —and fix one discrimination parameter for each latent dimension equal to one in order to identify the latent scale of the ideological positions.² This completes the model.

I use a Bayesian approach to estimate the parameters of the graded response model due to two practical, rather than philosophical, reasons. In order to meaningfully compare the latent ideological positions across individuals and different years, all the responses have to be scaled *jointly* such that they are estimated on the same scale. However, the sparse nature of the data matrix renders this task difficult for the maximum likelihood estimator, especially as the number of latent dimensions increases—a shortcoming which is largely overcome by the Monte Carlo algorithms with which Bayesian models are estimated. Another limitation of the maximum likelihood estimator is that it offers no information of the uncertainty in the estimated latent positions. Asymptotic standard errors are calculated only for the discrimination parameters; the latent positions, in contrast, are “predicted” under the assumption that the discrimination parameters are estimated without error and, hence, have no accompanying uncertainty measures. In the Bayesian approach, on the other hand, *all* parameters—including the latent positions—are objects of statistical inference. And the posterior distributions of these parameters summarize the uncertainty in the estimates.

Direct calculation of the posterior distribution is analytically intractable. Yet, it is possible to construct a Markov Chain that is assured to converge to the posterior distribution under reasonable assumptions. Statistical inference is then done by repeatedly sampling from the stationary distribution to which the Markov Chain has converged (hence the name

²Note that we have ignored in the formulation of the model the parameters that were fixed in order to resolve the scale and rotational indeterminacy of the model. Including them would amount to define an indicator variable, I_E (where E is the set of estimated parameters) which is one for estimated parameters and zero for fixed ones, and using $p(\cdot)^{I_E}$ instead of $p(\cdot)$ in the model formulation.

Markov Chain Monte Carlo or MCMC). The two most often utilized algorithms are the the Gibbs sampler and the Metropolis-Hastings algorithm. However, due to their random walk behavior, both procedures are rather inefficient in generating samples from the posterior distribution. Therefore, I rely on a Hamiltonian Monte Carlo algorithm (the No-U-Turn Sample) implemented in *STAN*. The major difference of Hamiltonian Monte Carlo and the Metropolis-Hastings algorithm is that the former generates proposals for new states (samples) by simulating a trajectory according to Hamiltonian dynamics with the help of “momentum” variables. This results in higher acceptance rates as well as proposals that can be far apart in the parameter space and, thus, is more efficient in exploring the posterior distribution. The relative efficiency of HMC over the Metropolis-Hastings algorithm is generally greater for more complex models, which is the major reason of relying of HMC in this study. Indeed, the number of iterations that were used to estimate the parameters (15,000), which might appear small for readers familiar with other MCMC algorithms, was a quite conservative choice for models fitted in *STAN*. For more detail regarding the HMC, see Gelman et al. (2014).

Convergence Statistics

Although convergence of MCMC chains is usually monitored for each quantity of interest separately, due to the abundance of parameters in the model, I present summary statistics in Table S1 for all parameters of the final model. The table shows clearly that all the chains have reached a stationary distribution and that they have mixed well. The effective sample size is above 500 for all parameters (first column), and largest the potential scale reduction factor amongst all of the parameters in the model is below 1.02 (last column).

	(Min)	(25%)	(50%)	(Mean)	(75%)	(Max)
<i>Effective Sample Sizes</i>						
γ	779.600	987.100	1154.000	1115.000	1240.000	1500.000
κ	294.600	470.200	568.600	634.900	810.100	1371.000
θ	826.700	1427.000	1500.000	1454.000	1500.000	1500.000
$\rho_{econ,civil}$	1269.000	1432.000	1500.000	1448.000	1500.000	1500.000
$\rho_{econ,moral}$	1203.000	1367.000	1445.000	1411.000	1482.000	1500.000
$\rho_{civil,moral}$	1252.000	1414.000	1439.000	1439.000	1500.000	1500.000
μ	503.400	581.400	701.600	741.600	858.300	1082.000
τ	844.800	1009.000	1094.000	1102.000	1213.000	1396.000
<i>Potential Scale Reduction Factor, \hat{R}</i>						
γ	0.998	1.000	1.001	1.001	1.002	1.008
κ	0.998	1.001	1.004	1.005	1.007	1.017
θ	0.996	0.999	1.000	1.000	1.001	1.010
$\rho_{econ,civil}$	0.997	0.998	0.999	1.000	1.000	1.002
$\rho_{econ,moral}$	0.997	0.998	0.999	1.000	1.001	1.002
$\rho_{civil,moral}$	0.997	0.998	1.000	1.000	1.001	1.002
μ	0.999	1.002	1.003	1.004	1.006	1.012
τ	0.998	0.999	1.001	1.001	1.002	1.007

Table S1: Convergence Statistics for Model 7 (Final Model).

Model Fit

As other information criteria, the WAIC is useful in comparing the relative predictive fit of different models, but it does not give us an objective sense of how well the model fits the data at hand. For example, if all competing models fit the data poorly, the final model can have the lowest WAIC but still be a bad model for the data. In order to examine objective fit of the model to the data, I use the posterior distribution of the parameters to simulate 500 data sets that are implied by the model. For each of these datasets, I calculate the means of all items that were analyzed, which are then compared to the means of the actual data. Figure S1 shows the results of the analysis. Out of 43 items, the the mean of 39 items are contained within the 5th and 95th percentile of the simulated datasets. Note also that the most salient issues are all well predicted by the model. Even for items that the model seems

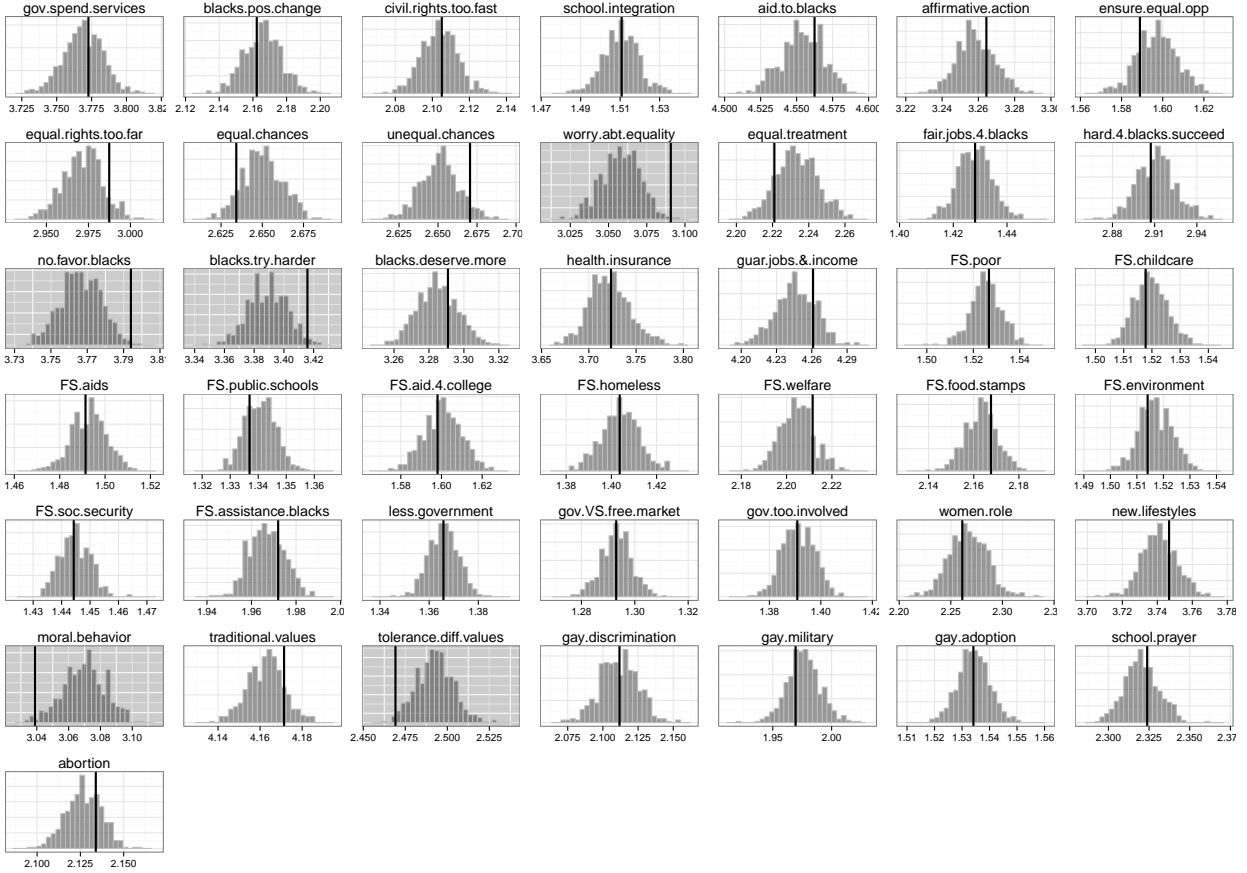


Figure S1: Distribution of 500 simulated datasets based on the posterior distribution of the model parameters. The vertical line shows the actual mean of the data for each item. The histogram shows the distribution of predicted means by the model.

to miss (histograms with gray-shaded background color), the distance between the predicted mean and the actual mean is smaller than .04 in every case on a 5-point scale. Although not perfect, it seems safe to conclude that the model shows a reasonable fit to the data.

Comparison of Aldrich-McKelvey Scaling and the Projection Method

The basic idea behind A-M scaling is to assume that the position of selected political stimuli (such as politicians or parties) are fixed and that each respondent reports a “distorted” version of it when asked to place these stimuli on the liberal-conservative scale. The stimuli

positions are then used as *anchors* (King et al. 2004) to scale the self-placement on a common scale with the stimuli. Formally, the model underlying A-M scaling is

$$\alpha_i + \beta_i X_{ij} = Y_j + \epsilon_{ij}, \quad (2)$$

or, equivalently,

$$X_{ij} = \frac{1}{\beta_i}(Y_j - \alpha_i) + \frac{\epsilon_{ij}}{\beta_i} \quad (3)$$

where X_{ij} is the placement of stimuli j by respondent i on the liberal-conservative scale, Y_j is the “true” position of stimuli j , and ϵ_{ij} is an error term that is assumed to be independently distributed with zero mean and constant variance (Aldrich and McKelvey 1977: 114). Thus, it is assumed that each individual reports a noise-added affine transformation of the true stimuli positions. Since the ANES asks the respondent to place not only themselves but also politicians and the two major parties on the same 7-point liberal-conservative scale, it is possible to estimate all the coefficients by fixing the mean and variance of the Y_j ’s to two arbitrary constants.

The *individual-specific* parameters, (α_i, β_i) , show how much each respondent “distorts” the true stimuli positions. For instance, a positive α_i would indicate that individual i places the stimuli positions too low, because we would have to add α_i to X_{ij} in equation (2) in order to recover Y_j . On the other hand, β_i shows how respondent i stretches or contracts the stimuli positions.³ Once the distortion parameters are estimated from the data, the corrected position of each respondent can be calculated by using the “same transformation that his perceptions were subjected to” (Aldrich and McKelvey 1977: 117):

$$\hat{p}_i = \hat{\alpha}_i + \hat{\beta}_i z_i, \quad (4)$$

³Note that the interpretation of the intercept is different from Hare et al. (2015). This is because Hare et al. use equation (3) in their estimation procedure, while I use the equivalent version of (2). Doing so allows me to avoid the division by a small constant when recovering the ideal points.

where z_i is respondent i 's reported self-placement on the liberal-conservative scale. Note that this transformation corrects for the distortion in that it rescales the raw self-placement of respondent i based on the estimates of how much i is distorting the true stimuli positions. At the same time, the transformation projects the liberal-conservative self-placements into the same space as the stimuli.

However, while A-M scaling offers a satisfactory solution to the incomparability problem within each cross-section, it does not allow the comparison of scaled responses over different years, unless all years are scaled jointly. The estimation procedure of the A-M model shows why. Estimates of the parameters α_i , β_i , and Y_j are obtained by solving the constrained minimization problem

$$\text{Minimize } \sum_{i=1}^n \sum_{j=1}^q \epsilon_{ij}^2 \quad \text{s.t.} \quad \sum_{j=1}^q Y_j = 0 \text{ and } \sum_{j=1}^q Y_j^2 = 1$$

which is simply a least-squares problem with the two constraints that the sum of the stimuli positions is zero and the variance of the stimuli positions is equal to one. The constraints on the mean and the variance of the stimuli positions are necessary to identify the latent scale of the stimuli (and therefore the scale of the corrected ideological positions). But, at the same time, it prevents the direct comparison between scaled responses over different years. Even if the true distribution at time point t has greater variance than the distribution at $t + 1$, the A-M procedure would rescale both distributions to have mean zero and variance one when the two years are scaled separately. So it is impossible, after the scaling is done, to compare the mean and variance of the distributions of time t and $t + 1$: whatever the true difference in the distributions, their mean and variance would be equal.

In order to “bridge” the scaled self-placements over time, two different methods might be used. First, it might be assumed that the stimuli positions do not change over time. While this assumption is obviously violated for the parties which became much more polarized

over the last decades (McCarty et al. 2006, Poole and Rosenthal 2011), it is a plausible assumption for individual politicians whose ideology has been shown to be highly stable (Poole 2007). Indeed, the assumption of stable ideologies is probably the most often evoked assumption to estimate a common scale for actors across different political institutions or over time in the political science literature (Shor et al. 2010, Shor and McCarty 2011, Poole and Rosenthal 2011). Since the A-M procedure provides estimates of the “true” positions of politicians, they might be used as anchors that bridge the scale of different years. In theory, only two politicians that are measured repeatedly over two different time periods are sufficient to identify a unidimensional common scale. The ratio of the the distances between the same two politicians at two different time points would equal the scaling factor that can be used to map one scale into the other. A limitation of this approach is that the number of “anchoring” politicians has to be sufficiently large. When only few “bridges” are used, small measurement errors can lead to uninterpretable results. Unfortunately, the number of politicians that are placed on the liberal-conservative item in the ANES is quite limited. For most consecutive presidential election years, only one or two figures can be used as bridges. On the other hand, both the Republican and Democratic parties have become more polarized over the period under study, so that it would be unrealistic to assume that the position of the parties remained the same. Hence, a different method is called for.

The projection method used in the paper tires to overcome the incomparability problem by substituting the Y_j 's in equation (2) by an external scale that 1) measures approximately the same dimension as the liberal-conservative item in the ANES and 2) remains constant over the entire period of the study. In the analysis, I use the first dimension of the Common Space DW-NOMINATE space (Poole and Rosenthal 2011) for this purpose. While the choice of the NOMINATE scores might look arbitrary, it is the standard ideology measure for politicians and political parties in the political science literature. And, indeed, why relying on the perceptions of survey respondents to estimate the “true” positions of politicians

when behavioral measures are readily available? Now, everything that is changed is that Y_j in equation (2) is replaced by the corresponding NOMINATE scores, Y_j^* , i.e.,

$$\alpha_i + \beta_i X_{ij} = Y_j^* + \nu_i,$$

where Y_j^* is the first-dimension DW-NOMINATE score of stimuli j , X_{ij} is the placement of stimuli j on the liberal-conservative 7-point item by respondent i , and ν_i is an error term. Also, as equation (4) gives us the best linear projection (in the least squares sense) into the space defined by the stimuli Y_j , so is

$$\hat{p}_i^* = \hat{\alpha}_i + \hat{\beta}_i z_i$$

the least-squares projection into the NOMINATE space.

Online Appendix B: Supplementary Figures

Operational Ideology

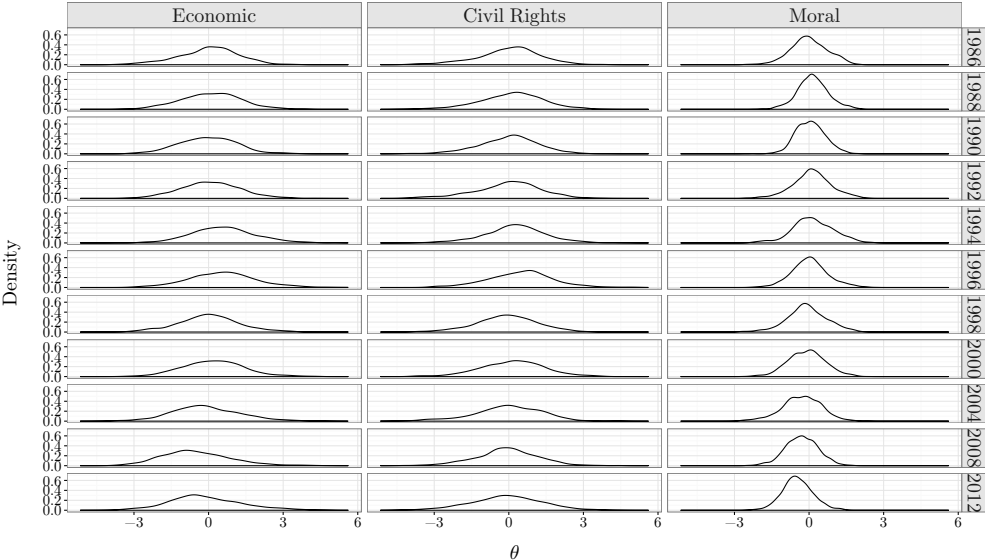


Figure S2: Kernel density estimates of ideological distribution: operational ideology, 1988-2012. Estimated distribution of posterior means for all years.

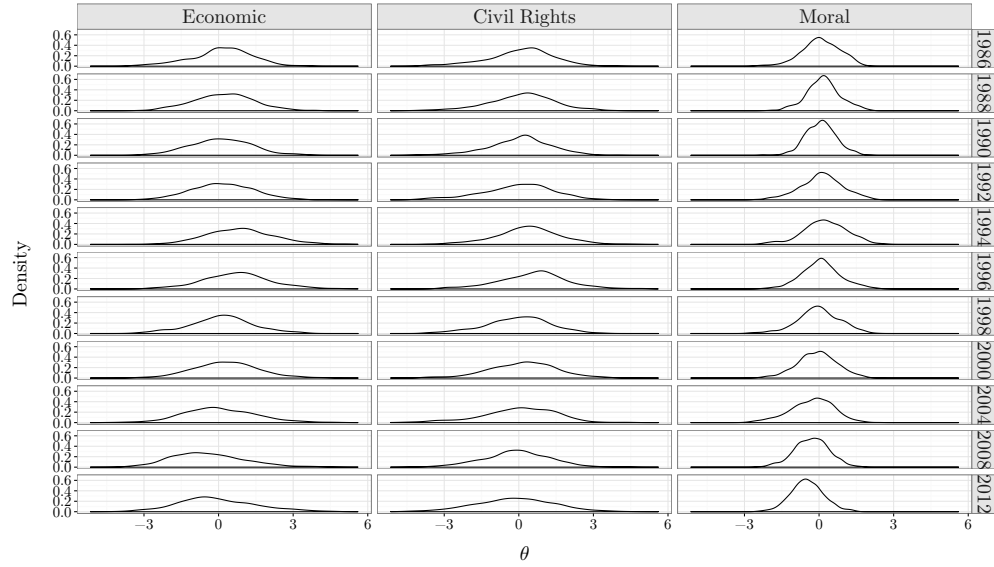


Figure S3: Kernel density estimates of ideological distribution: operational ideology, respondents who reported to have voted in last election, 1988-2012. Estimated distribution of posterior means for all years.

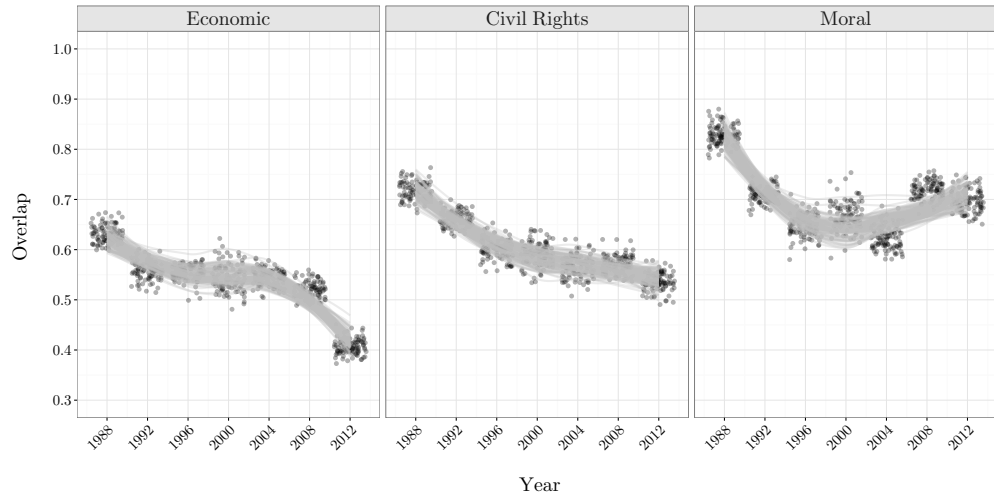


Figure S4: Sorting on the economic, civil rights, and moral domain, respondents that were also used in the analysis of symbolic ideology, 1986-2012.

Symbolic Ideology

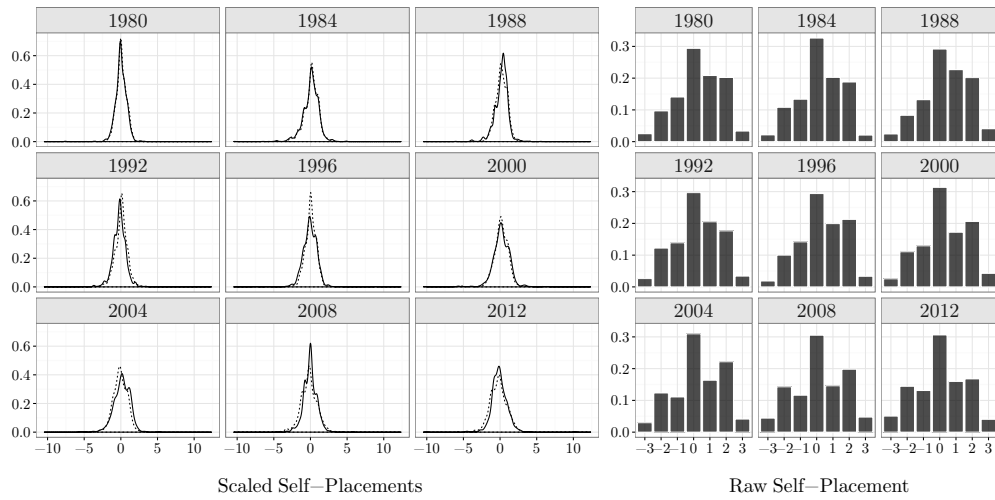


Figure S5: Scaled ideology distributions estimated from projection into the NOMINATE (solid line) and CFscore (dashed line) Space, all respondents, 1980-2012.

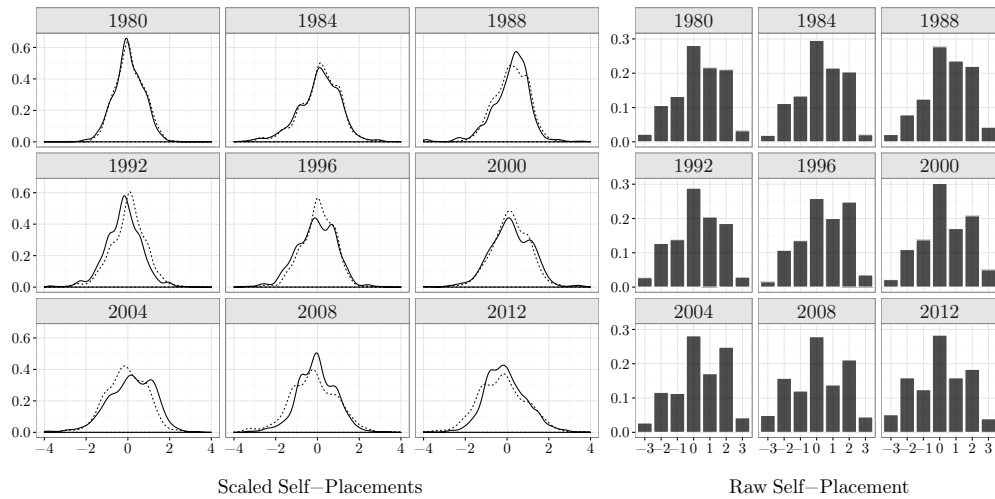


Figure S6: Scaled ideology distributions estimated from projection into the NOMINATE (solid line) and CFscore (dashed line) Space, respondents who reported to have voted in last election, 1980-2012.

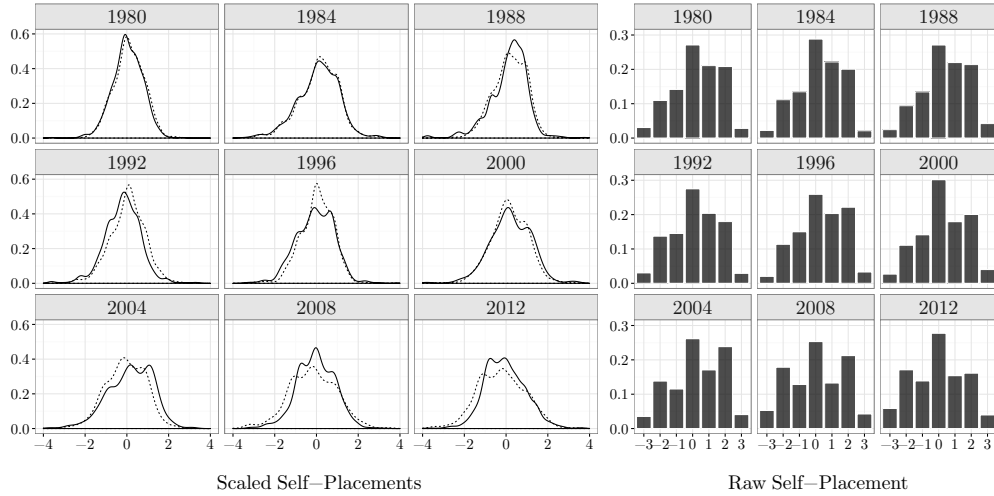


Figure S7: DIF-corrected ideology distribution estimated from projection into the NOMINATE (solid line) and CFscore (dashed line) Space, respondents with positive $\hat{\beta}_i$, 1980-2012.

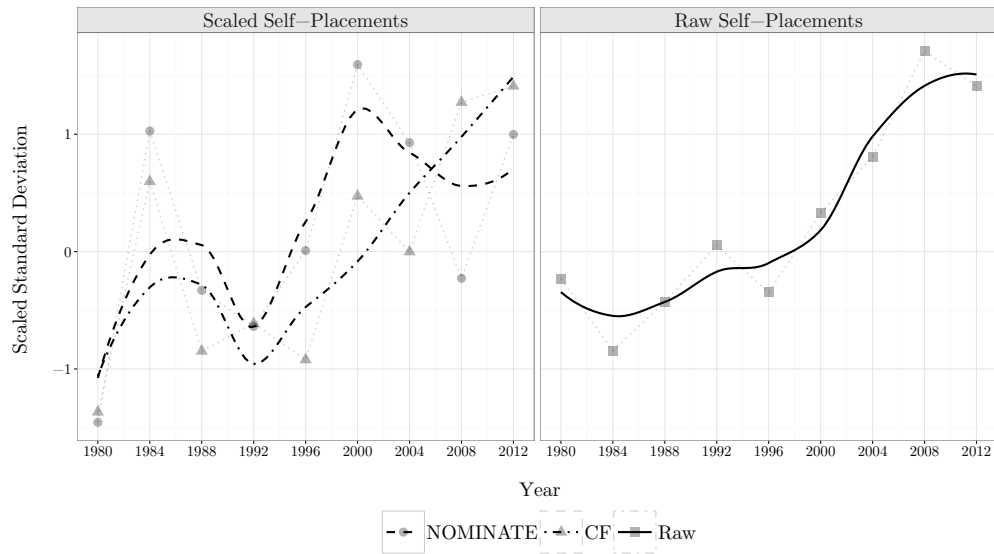


Figure S8: Estimated standard deviation of symbolic ideology, 1980-2012. The standard deviation of each time-series is scaled to have mean zero and variance one in order to facilitate the comparison between the time trends.

References

- Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *The American Political Science Review* 71:111–130.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.
- Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2014. *Bayesian Data Analysis (3rd Ed.)*. Taylor & Francis.
- Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll, and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59:759–774.
- King, Gary, Christopher JL Murray, Joshua A Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98:191–207.
- McCarty, Nolan, Keith T Poole, and Howard Rosenthal. 2006. *Polarized America*. MIT Press.
- Poole, Keith T. 2007. "Changing Minds? Not in Congress!" *Public Choice* 131:435–451.
- Poole, Keith T and Howard L Rosenthal. 2011. *Ideology and Congress*. Transaction Publishers.
- Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12:338–353.

Shor, Boris, Christopher Berry, and Nolan McCarty. 2010. "A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-institutional Common Space." *Legislative Studies Quarterly* 35:417–448.

Shor, Boris and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105:530–551.

Takane, Yoshio and Jan De Leeuw. 1987. "On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables." *Psychometrika* 52:393–408.