

Estimating the Political Center from Aggregate Data: An Item Response Theory
Alternative to the Stimson Dyad Ratios Algorithm

Anthony J. McGann

Accepted by Political Analysis
December 4, 2013

Abstract

This paper provides an algorithm to produce a time series estimate of the political center (or median voter) from aggregate survey data, even when the same questions are not asked in most years. This is compared to the existing Stimson dyad ratios approach, which has been applied to various questions in political science. Unlike the dyad ratios approach, the model developed here is derived from an explicit model of individual behavior – the widely used item response theory model. I compare the results of both techniques using the data on public opinion from the United Kingdom from 1947-2005 from Bartle, Dellepiane-Avellaneda and Stimson (2011a). Measures of overall model fit are provided, as well as techniques for testing model's assumptions and the fit of individual items. Full code is provided for estimation with free software WinBUGS and JAGS.

1. Introduction

There are many contexts where we wish to estimate the aggregate level of public opinion – the position of the political center, the location of the median voter, the prevailing policy mood, or the overall level of support for a particular policy. Often we do not have individual level panel data for the entire time period we are interested in. This is not an insurmountable problem as we are interested in aggregate, rather than individual, level change. Indeed, in order to increase the coverage of the time period, we may use questions for which we only have aggregate response data. However, we face another problem. Even if we have many question items administered in each year, it is often the case that no item is administered in more than a small proportion of the years. However, if the questions that are asked in different years overlap sufficiently, it is still possible to generate a comparable time series measure for the level of public opinion for all years. This paper proposes a model for doing this based on item response theory, and evaluates it against the existing approaches such as the Stimson dyad ratios algorithm.¹

The Stimson dyad ratios algorithm has now been used to address this kind of problem in a considerable number of contexts. It was developed in Stimson's work on American public opinion (Stimson 1991, 1999). It is an important part of the Erikson, McKuen and Stimson's *Macropolity* project (Erikson, MacKuen, and Stimson 2002; Stimson, Mackuen, and Erikson 1995). More recently the approach has been extended to

¹ I would like to thank Professor William Batchelder for introducing me to item response theory. I would also like to thank Professor Simon Jackman. I would like to thank Dr. John Bartle, Professor James Stimson and Dr. Sebastian Dellepiane-Avellaneda, whose work inspired this project, and who were kind enough to share their data with me.

the United Kingdom (Bartle, Dellepiane-Avellaneda, and Stimson 2011a; Bartle, Dellepiane-Avellaneda, and Stimson 2011b) and France (Stimson, Thiébaud, and Tiberj 2009). There have also been a number of unrelated studies that use the dyad ratios algorithm (Cohen 2000; Chanley, Rudolph, and Rahn 2000; Kellstadt 2003; Voeten and Brewer 2006; Baumgartner, De Boef, and Boydston 2008).

This ability to measure change in public opinion over long time period provides the opportunity to address many additional outstanding problems. This includes not just issues of political behavior, but also political economy and comparative political institutions. For example, to operationalize some rational choice theories, it is necessary to have a measure of the position of the median voter. When studying phenomena such as the growth and retrenchment of the welfare state, it would be hugely helpful to have an independent measure of public demand for such programs. There are various studies that attempt to measure the responsiveness to public opinion of various electoral systems or constitutional arrangements (Powell 2000; McDonald and Budge 2005). The problem they face is that they are forced to use extremely indirect measures of public opinion – left-right self-placement from survey data in Powell; the declared position of the median party in parliament in MacDonal and Budge (see also Kim and Fording 2001). The methods discussed here provide a way to estimate this directly using existing data.

The Stimson dyad ratios algorithm is an ingenious approach to these problems that was computationally tractable given the computer resources available in the 1990s. However, it is theoretically ad hoc – there is no individual level model linking individual level response behavior to the aggregate outcomes we observe. For this reason, it is

uncertain exactly what is being measured. Given that computer resources are now far less of a constraint, other approaches are now possible.

As an alternative, I propose an approach that explicitly estimates the central tendency of the distribution of public opinion based on an individual level model of behavior. This adapts an established item response theory model from psychometrics – a model that has also been widely applied to individual level data in political science in recent years. This approach has the added advantage of estimating not only the central tendency, but also the variance of the distribution. I provide code for implementing the model using freely available software (BUGS and JAGS). I also compare the results of the item response theory approach with those using the dyad ratios algorithm and other approaches, using the data on British domestic public opinion from Bartle, Dellepiane-Avellaneda and Stimson (2011a).² In addition to comparing overall measures of model fit, I provide tools for testing the assumptions of the models and evaluating item selection.

2. The Dyad ratios Algorithm and existing IRT approaches

We can compare the model proposed in this paper to the existing models for estimating policy mood from aggregate data (Stimson 1991; Voeten and Brewer 2006; Jackman 2005), and also to existing IRT approaches. The existing policy mood models lack an individual level model of response, but instead simply assume the existence of aggregate policy mood. Item response theory provides such an individual level model.

² Replication data is available from Anthony McGann "Replication data for: Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm", <http://dx.doi.org/10.7910/DVN/22861> IQSS Dataverse Network [Distributor] V1 [Version]

There have been many applications of item response theory to political attitudes and ideology. These existing IRT models, however, use individual level data. The approach pursued in this paper is essentially to extend these models to the case where we only have aggregate data.

The Stimson dyad ratios algorithm was the first attempt (to my knowledge) to estimate the central tendency of the policy mood for each year from aggregate data.³ The key assumption of the algorithm is that the ratio of left responses between two years in which a given question is asked represents an estimate of the relative policy mood of these two years. It is then possible to piece together the various estimates of the relative magnitudes of policy mood to produce a series of estimates. Thus, each administered survey question is recoded so that answers are classified as either left-wing or not left-wing, and the proportion of left-wing responses ($\text{left} / (\text{left} + \text{not left})$) is recorded for each question. If a question is asked in years $t+i$ and $t+j$, then R_{ij} is the ratio of the proportion of left-wing responses to the question administered in year $t+i$ to the proportion of left-wing responses in year $t+j$. Thus the key assumption of the model (Bartle, Dellepiane-Avellaneda, and Stimson 2011a, 268) is:

$$\boxed{\phantom{R_{ij} = \frac{x_{t+i}}{x_{t+j}}}} \quad (1)$$

where x_{t+i} and x_{t+j} are positions of the political center in years $t+i$ and $t+j$. As is made clear in the text, what this means is that R_{ij} (for which we have observable data) can be

³ In laying out this algorithm, I use the terminology from Bartle, Dellepiane-Avellaneda and Stimson (2011a), as this is most recent exposition and the notation is most elegant. However, the fullest explication of the algorithm is in Stimson (1999, 133-7).

used as an estimate of the relative magnitude of x_{t+i} and x_{t+j} . If we have assumed or imputed a value for x_{t+j} , we can use R_{ij} to estimate x_{t+i} . Of course there are usually multiple items that are asked in any pair of years, so there are also multiple estimates of the policy mood in any year. The algorithm essentially averages together these estimates of policy mood. It then uses these estimates to calculate the communality of each question – how well it predicts policy mood. The policy mood estimates are then recalculated weighted by these communalities. This process is repeated iteratively until the policy mood and communality estimates converge.

As the Stimson dyad ratios algorithm lacks an explicit individual level model of response, it is not clear exactly what it is measuring. However, the central assumption of the model – that the ratio of the percentage of left responses to the same question asked in different years provides a measure of the relative policy mood of the two years – has some worrying implication that cast doubt on its accuracy as a measure of policy mood. A change from 10% answering a question in a left-wing manner to 20% represents the same dyad ratio as a change from 20% to 40%, or from 40% to 80%, and thus by the assumption of the algorithm represents the same change in policy mood. Apart from seeming arbitrary, this produces a startling asymmetry between left and right. A movement from 20% to 40% gives a ratio of 1:2, but a movement from 80% to 60% only gives a ratio of 4:3. However, if we counted the probability of right-wing answers instead of left-wing, these ratios would be reversed. This suggests that we may get different results if we calculate from the percentage of right-wing answers, even though this is providing essentially the same information. This phenomenon is especially problematic when we consider questions that are extremely hard or extremely easy to answer in a left-

wing manner. If 90% of the population answers a question in a left-wing manner in years when the political center is around its long-run median, it is not possible for this question to reflect a sharp move to the left. Even if we have 100% give the left-wing answer, we only get a ratio of 10:9.⁴

Another, more recent, approach to this problem is that of Voeten and Brewer (2006). This uses Bayesian estimation of a linear function to estimate policy mood from aggregate data. The policy mood variable they consider is support for the US war in Iraq, and the data is the percentage of respondents answering in a pro-war manner to various questions. Questions are asked repeatedly, but no question is asked in every period. Thus the model is:

$$Y_{jt} = a_j + b_j \theta_t + \epsilon_{jt} \quad (2)$$

where Y_{jt} is the percentage answering yes to question j in period t , a_j is the bias of the question i , b_j is the loading of question i and θ_t is the policy mood in period t . The policy mood θ_t is assumed to be subject to a random walk adjustment process. This model is closely related to the model that Jackman (2005) proposes to pool opinion polls from different survey houses – they are both linear models of aggregate public opinion with dynamic adjustment processes.

⁴ This problem may be mitigated somewhat by the fact that Stimson's algorithm iteratively reweights items based on communalities. If extremely easy or hard items exhibit the odd response behavior we predict, we would expect them to correlate poorly with the estimate of policy mood, and be assigned a low weight. Essentially there would be item selection in favor of items with average difficulty. The cost of this is that the information contained in very easy or difficult questions is not used.

As with the Stimson algorithm, these models provide no individual level model of response – it is simply assumed that the aggregate responses react to changes in policy mood in a linear manner. These linear models do not produce the asymmetry in response between left and right that we see with the dyad ratios algorithm. In fact, a given change in policy mood is assumed to produce the same change in the percentage of left answers, regardless of how many people were giving left answers to start with. This assumption is itself problematic. If we have a policy mood that is already quite left-wing, we would expect even a large movement to the left to have very little effect (virtually everyone is answering left-wing already). If, however, the policy mood is quite centrist, we would expect a change in policy mood to have a much greater effect.

Item response theory models allow us to model response in a more plausible way, so that a given move in policy mood can have a different effect depending on how likely a left response was anyway. It usually does this by using a logistic or cumulative normal link function, in a manner similar to logistic or probit regression.

Item response theory models have long been used in psychology, but have also recently been applied to many problems in political science. They have been applied to ideal point estimation in legislatures using roll call data (Jackman 2001; Clinton, Jackman, and Rivers 2004). Indeed Jackman (2000) points out that the Poole and Rosenthal (1997) NOMINATE algorithm is closely related to the item response model. Other scholars have used item response models to estimate the ideal points of Supreme Court justices (Martin and Quinn 2002; Bafumi et al. 2005; Peress 2009). There have also

been some applications of item response theory to mass opinion using individual level survey data (Jessee 2009; Bafumi and Herron 2010; Levendusky and Pope 2010).⁵

The model set out in this paper differs from these applications in two ways. Firstly, although the model is built from an individual level model of survey response, the data is at the aggregate level. As with the Stimson dyad ratios and the Voeten and Brewer approaches, we do not have data on individual responses, but only the percentage of respondents who responded each way to a question. Secondly we do have questions that are repeated at different points over time. This is not typically the case when estimating the ideal points of legislators or Justices – bills and court cases only happen once. Most existing item response models either estimate ideal points for a single legislature or Court, assume that legislators or Justices retain fixed positions over their careers, or use a dynamic updating model (for example, Martin and Quinn 2002), whereby it is assumed that ideal points change gradually over time.

3. An item response theory model of policy mood

I adapt the item response theory model to measure policy mood using aggregate survey response data. While existing methods, like the dyad ratios approach, *may* track the political center, we have no theoretical basis for the claim that we are measuring the central tendency of the distribution of public opinion. With an IRT approach we can explicitly model the central tendency and dispersion of public opinion and draw inferences from it.

⁵ These articles all use the Cooperative Congressional Election Study, which asks respondents questions referring to specific Congressional Bills, thus allowing voters and legislators to be placed in the same space.

The standard item response theory model uses responses to test questions to simultaneously estimate the ability of the respondents and the characteristics of the questions (Nunnally and Bernstein 1994, 393-409). We assume that each individual has an ability level that is unidimensional, with x_i representing the ability level of individual i . Each question q can have various parameters, but we assume it has two, difficulty and discrimination, represented as λ_q and α_q . The most commonly used functional forms are the lognormal and (as here) the normal cumulative distribution function:

(3)

This is, of course, closely related to probit regression, which is commonplace in political science. Figure 1 graphs this response function for three questions with different values of λ and α .

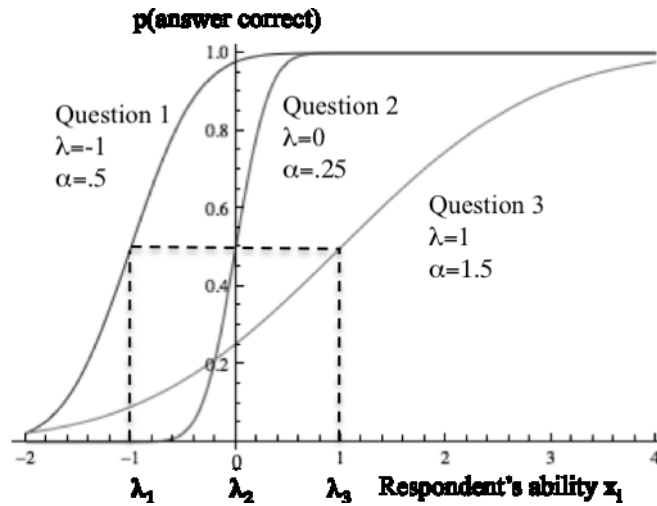


Figure 1 – Probability of correct response to three questions

If the position of the respondent x_i is equal to λ_q , then the probability of giving the correct response is .5. If x_i is greater (less) than λ_q , then the probability of a correct response is greater (less than) 0.5. Thus question 1, with its low value of λ , is easy to answer correctly (a very low level of ability is required to have a high probability answering incorrectly), while question 3 has a high value of λ and is very difficult to answer correctly. How quickly the probability of a correct response increases or decreases as the respondent's ability changes depends on the discrimination parameter α_q . Question 1 has a low value of α , so if the respondent has ability much greater than λ_1 , the probability of a correct response rapidly approaches 1. However question 3 has a high α , so as the respondent's ability moves lower than λ_3 , the probability of a correct response only falls slowly.

Looking at Figure 1, we can see why the IRT model provides a more plausible model of response than the models implied by the Stimson dyad ratios approach or the Voeten and Brewer (2006) linear model. In contrast to what is implied by the dyad ratios algorithm, the response function is symmetric, so we get equivalent results whether we use left or right-wing responses. Unlike the Voeten and Brewer model, the response is not linear, so a given change in ability (or preference) can make more difference when a respondent's ability gives them a fifty-fifty chance answering a question correctly (or left-wing) than for a respondent who is almost certain to answer the question correctly anyway.

If we had individual level data, this model could be adapted to our problem in a straightforward way. The probability of answering in a left-wing manner would simply

replace the probability of answering correctly; and “ability” would be reinterpreted as propensity to answer in a left wing manner, or simply as how “left-wing” a respondent was (with higher scores for more left-wing respondents). The difficulty and discrimination parameters of the questions would work exactly as before. The only change required would be that the same respondent would be allowed to have a different policy position for each year they participated.

We, of course, do not have sufficient individual panel data with which to estimate this model, but instead have to rely on aggregate data. Therefore we cannot estimate the position of each individual, but only the average position of the population in any given year. This, however, is exactly what we are interested in – the central tendency of the distribution of public opinion. Indeed, if we assume that the population is normally distributed, not only can we estimate the population mean, but also the standard deviation, which gives us a measure of how polarized public opinion is.

Let us lay out the model formally. Let us denote the year by y with the range $\text{startyear} \dots \text{endyear}$. Let us denote the number of the question or item being asked as q , which ranges from $1 \dots Q$. (Each question will be asked in at least two different years.) We have data on the proportion of respondents who gave a “left-wing” answer to each question that was asked. This we denote as $\text{leftr}_{y q}$, the proportion that gave a left wing answer to question q in year y . For each year, we estimate the mean policy position of population, which we will call μ_y . This is simply a real number with the convention that low scores stand for more right-wing positions and high scores for more left-wing positions. We will also estimate the standard deviation of the distribution of the respondents’ positions in each year, σ_y . In addition we need to estimate parameters that

characterize each question. As before, the parameter λ_q stands for the “position” of question q , with high scores for question that are hard to answer in a left-wing manner, while parameter α_q represents how effectively the question discriminates between left and right-wing respondents.

Although we only have aggregate data, we need to start with a model of individual choice on responses. We use the same cumulative normal model as above, with mean λ_q and standard deviation α_q . If $e_{i y q}$ is the probability of respondent i in year y giving a left wing response to question q , and $x_{i y}$ is the policy position of respondent i in year y , and Φ is the cumulative normal distribution function,⁶ then:

$$\boxed{\phantom{e_{i y q} = \Phi\left(\frac{x_{i y} - \lambda_q}{\alpha_q}\right)}} \quad (4)$$

We assume that the respondents in year y are normally distributed with mean μ_y and standard deviation σ_y . Essentially we are assuming that all respondents in a given year are drawn from the same distribution of policy positions, and that each question behaves on average in the same way, whatever year it is asked in. The response function $e_{y q}$ (Equation 5) gives us the probability of a respondent with position x giving the left-wing response. We assume that the probability of a respondent having policy position x in year y is given by the normal probability density function with mean μ_y and standard deviation σ_y . Given this, the probability of a randomly selected agent giving the left response to question q in year y (which we will call $m_{y q}$) is the product of $e_{y q}$ and the

normal density function integrated over all possible values of x . Thus if ϕ is the normal probability density function:

$$\boxed{\phantom{\int_{-\infty}^{\infty} \phi(x) dx}} \quad (5)$$

Integrating, we get:

$$\boxed{\phantom{\int_{-\infty}^{\infty} \phi(x) dx}} \quad (6)$$

We may note the symmetry in the effect of the parameters α_q and σ_y . The standard deviation of the normal distribution function that gives the probability of a left response is the geometric mean of the two parameters. Intuitively we would expect the two parameters to have a similar effect. A question that does not discriminate well leads to a significant number of right wing responses, even given a left-wing population. A population with a high variance will also lead to a significant number of right-wing responses, even if the mean of the population is quite left-wing.

Given the function $m_{y,q}$ for the expected probability of a random respondent giving a left-wing answer, we can model the total number of left-wing responses to a question using the beta-binomial distribution. The reason for using the beta-binomial is to allow some stochastic variation to account for the fact that the same question may be applied or understood in slightly different ways in different years. Thus the final probability of a left-wing response to question q in year y is distributed according to the beta distribution with expectation $m_{y,q}$. Given that the expectation of the beta distribution

with parameters α and β is $\alpha/(\alpha+\beta)$, we can reparameterize the beta distribution in terms of its expectation. Thus the probability of a random respondent giving a left-wing response, which we call p_{yq} , is distributed thus:

$$p_{yq} \sim \text{beta} \left(\frac{m_{yq}}{1 - m_{yq}} \beta, \beta \right) \quad (8)$$

The number of left-wing responses expected for question q in year y , where n_{yq} is the number of responses to that question, is then distributed binomially:

$$\text{leftr}_{yq} \sim \text{binomial}(p_{yq}, n_{yq}) \quad (9)$$

Assigning non-informative uniform priors to the parameter vectors β, σ, λ and α , as well as to the beta parameter β , we can estimate the model using Bayesian inference software such as BUGS or JAGS. Figure 2 gives a graphical summary of the model. Appendix 1 gives complete code for JAGS and BUGS.

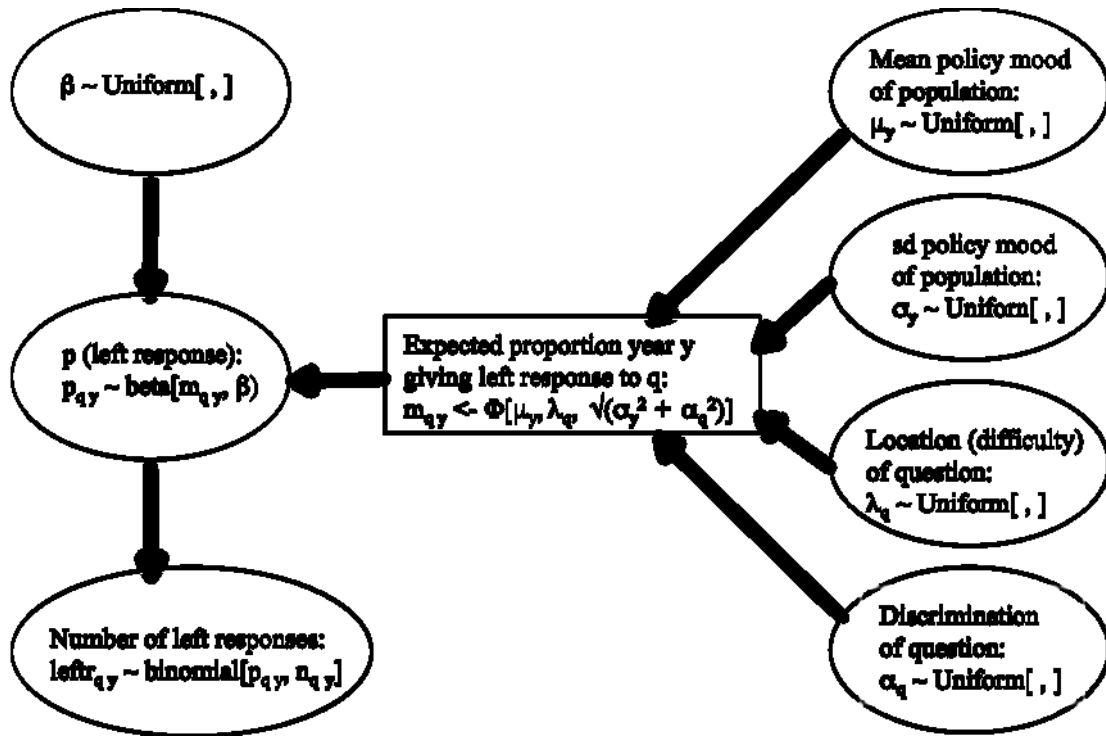


Figure 2 – Graph of the model

4. Results and Model Fit

The IRT algorithm was applied to data from Bartle, Dellepiane-Avellaneda and Stimson (2011a). The authors collected a database of the aggregate results of survey questions about political matters in the UK between 1947 and 2005. The results were recoded to give the proportion of respondents giving the “left-wing” answer to each question. Following the authors, I have only used the question dealing with domestic policy issues. There were 364 different question items (two of which were dropped for reasons for reasons of differential item functioning⁷). Each of these was asked in at least

⁷ Testing for differential item functioning – the violation of the assumption that items behave the same way whatever year they are asked in – is dealt with in section 5. However, cursory analysis of the data revealed that two items (dealing with spending on education and the National Health Service) strongly violated this assumption.

two different years, resulting in a total of 2377 question administrations.⁸ Each line of the input dataset corresponds to one question administration, with variables for the proportion of left responses ($\text{left} / (\text{left} + \text{not left})$), the year the question was asked, the number of the question item, and the number of respondents.

The model was estimated using JAGS. I used 3 chains, each with 30,000 iterations, and a 15,000 iteration burn-in. Convergence was checked for by inspecting the traceplots and by calculated Geweke and Raftery / Lewis diagnostics. These diagnostics are provided in the web/reviewer appendix. The full CODA file is available on request. The model estimates policy position and standard deviation parameters for each of the 59 years from 1947 to 2005, position and discrimination parameters for each of the 362 questions and one coefficient for the beta distribution, which captures the degree of stochastic variation between question administrations. Given we are more interested in the policy positions of the population than the characteristics of the individual questions, the parameters relating to the question items (λ_q and α_q) are given in the web/reviewer appendix, while the parameters for the policy position of the population (μ_y and σ_y) are given in the main appendix and graphed in Figures 3 and 4

Figure 3 graphs the estimate of policy mood by year from the IRT method outlined in the last section, and compares this to the results from Bartle, Dellepiane-Avellaneda and Stimson (2011a, 271) computed using the Stimson dyad ratios method. The IRT results represent the probability of a respondent answering in a left-wing manner to a typical question in a given year ($\lambda_q = -.357$, $\sigma_y = 6.16$). The Stimson dyad ratios

⁸ The average question was asked in between six and seven different years. The two most often asked questions were asked in 40 and 20 out of the 59 years.

method does not directly allow the estimation of response to individual items, but the raw policy mood scores have been standardized using the “validity-weighted means and standard deviations of the input items” (Bartle, Dellepiane-Avellaneda, and Stimson 2011a, 269).



Figure 3 -- Estimated policy mood – IRT method and dyad ratios results from Bartle, Dellepiane-Avellaneda and Stimson (2011).

It is apparent that the results from the two methods are quite similar. The correlation coefficient between the Bartle, Dellepiane-Avellaneda and Stimson dyad ratios and the IRT estimates from 1950 to 2005 is 0.71. However, this coefficient is depressed by the period 1947-62 for which there is very little data. From 1963 to 2005 the correlation is .91, and for the years 1970-2005 it is .93. We can observe the same patterns in both series of estimates. In general public opinion moves against whoever is in

office. The British electorate leans to the left in the early postwar period, but through the fifties and sixties it seems to gradually move towards the center (although data is sparse until the mid-1960s, so we have limited confidence in our measurements). There is instability in the early seventies, followed by a sharp move to the right following the re-election of a Labour government under Harold Wilson in 1974. However, the electorate moves steadily leftward under the Conservative governments of Thatcher (1979-90) and Major (1990-7). The victory of Tony Blair and the Labour Party in the 1997 general election leads to a gradual movement back towards the right, although this is temporarily halted in the election years of 2001 and 2005.

There are, however, some substantively interesting differences. According to Bartle, Dellepiane-Avellaneda and Stimson (2011a) estimates, policy mood in the UK moves to the right until the election of Margaret Thatcher in 1979. However, the IRT estimates indicate that policy mood reached its rightmost point in 1976, the year when the Sterling crisis forced the UK to seek IMF loans, and subsequently moved somewhat back to the left under the Callaghan administration (1976-79). The movement of policy mood to the right under Blair (1997-2005) is rather less pronounced according to the IRT estimates than with the dyad ratios estimates. While the dyad ratios estimates are smoothed as part of the algorithm, the IRT results are not. This allows us to observe some interesting patterns on a year by year basis. For example, in the election years of 2001 and 2005, the rightward movement of policy mood under Blair was temporarily reversed. In 1992, we observe a similar effect – the leftward movement of public opinion is temporarily halted, and John Major won the general election.

The IRT algorithm allows us to estimate not just the central tendency of the population's policy mood, but also the dispersion of the population. Figure 4 illustrates this. The black line charts the probability of a randomly selected respondent giving a left wing answer to a typical question in a given year. The grey lines chart that probability of a person drawn from the population to the left or right of the median giving a left wing response. These are calculated using the following formulae:

$$\boxed{\phantom{\text{Empty box for formula (10)}}}$$

(10)

$$\boxed{\phantom{\text{Empty box for formula (11)}}}$$

(11)

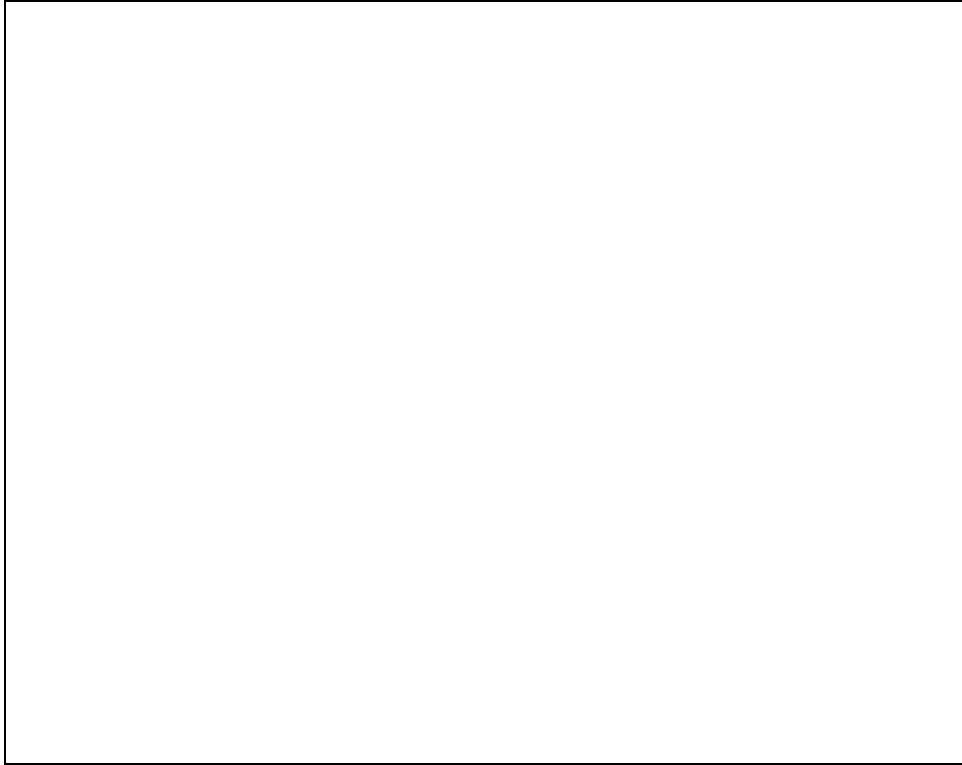


Figure 4 – Predicted percentage giving left response to a typical question for whole population and left-wing and right-wing halves of population

We can see that the polarization of the population varies considerably across time. During the 1970s there is considerable polarization. In the late 1970s and early 1980s, this polarization decreases as the left and right wing parts of the population converge. After this polarization increases again. These changes in the dispersion or polarization are interesting in themselves. Estimating the dispersion of the population also allows us to get better estimates of its central tendency. It is notable that the period where the IRT estimates of policy mood diverge most from the dyad ratios estimate are when there is a big change in polarization, for example in the mid-1970s. This can be explained by the

fact that the dyad ratios algorithm will interpret an increase or reduction in polarization as a change in the central tendency of policy mood.⁹

Model Fit and Alternative Models

I have argued that the IRT model outlined here provides a more plausible model of response behavior than the previously existing models. We can compare the IRT model with other methods in terms of how well it fits the data. In this context model fit means how well we can reconstruct the original data from the model parameters we have estimated. That is to say, given our estimates of policy mood and the parameters describing the behavior of each survey item, how well can we predict the response to each item administration?

A simple measure of fit is the root mean square residual. This is defined as:

$$\sqrt{\frac{1}{A} \sum_{a=1}^A \text{leftp}_a - \hat{\text{leftp}}_a^2}$$

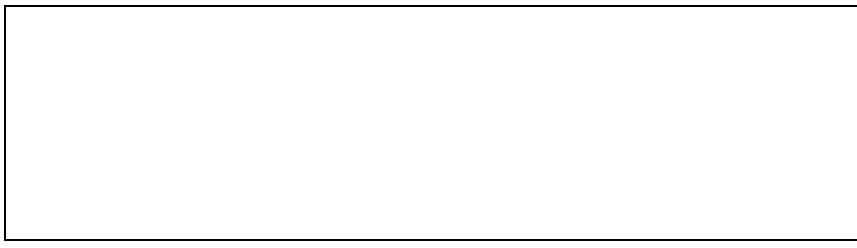
(11)

where leftp_a is the percentage giving a left wing answer to question a , a is the index of the question administrations $1 \dots A$, and the hat sign means predicted value. Intuitively, this is how many percentage points we are off on average in estimating the aggregate response to a question.

Of course, it is possible that we could predict responses quite accurately, but that this is simply a result of the fact that some questions always draw few left-wing answers

⁹ If the central tendency is to the right and dispersion decreases, then fewer respondents will give left wing answers to questions that are hard to answer in a left wing manner. This will be interpreted by the dyad ratios algorithm as shift to the right in policy mood.

while other questions are uniformly easy to answer in a left-wing manner, so that we can explain most of the variance with just the item means. To account for this possibility, we can calculate a by item R^2 measure. This is identical to a normal R^2 , except that instead of calculating the denominator using the mean of all question administrations, we use the mean of the relevant item. The by item R^2 is defined:



(12)

Roughly speaking, this measures the proportion of the variance that the model explains, over an above what is explained by the item means.

The models that we are comparing have differing numbers of parameters. For this reason I calculate an adjusted by item R^2 measure. This is defined:



(13)

where df_m is the number of degrees of freedom of the predictive model, while df_n is the number of degrees of freedom of the estimate of the variance. Given that the mean is a vector of means for each item, df_n is the number of question administrations minus the number of items.

We can compare the fit of the IRT model to various other models. Firstly, there is the linear model of the type used by Voeten and Brewer (2006) (see Equation 2). Secondly there is a simple probit model. This is specified as:

$$\boxed{\phantom{y_{ij} = \alpha_j + \lambda_j x_{ij}}} \quad (14)$$

where x_{ij} is policy mood, λ_j is the difficulty parameter, α_j is the discrimination parameter and m_{y_j} is the expected proportion of left-wing responses. This is then input into the same beta-binomial distribution as with the IRT model to estimate response. Finally there is the Stimson dyad ratios algorithm. This model does not directly predict the responses to individual items. However, we can take the estimates of policy mood from Bartle, Dellepiane-Avellaneda and Stimson (2011a) and use either linear or probit regression to estimate the responses to the different items that were asked, giving each item a difficulty and a discrimination parameter.

Table 1 compares the fit of the various models – that is to say, how well he can recreate the original responses from our estimates of policy mood and question difficulty. First we consider the root mean squared residuals. Roughly speaking this is the average error we get when we estimate the responses to all the questions from our parameters. With all our models this is between 5 and 6 percentage points. By way of comparison using just the mean response levels of the various questions items, we get a residual of 7.24 percentage points. The residual for the IRT model is 5.38%, which is somewhat lower than the other models.

	Root mean squared residual	By item R^2	Adjusted by item R^2
IRT model	5.35	0.481	0.319
Linear model	5.68	0.416	0.290
Probit model	5.58	0.435	0.313
Dyad ratios / linear predictions	5.78	0.397	0.238
Dyad ratios / probit predictions	5.96	0.359	0.191
Item means only	7.24	0 (by construction)	0 (by construction)

Table 1 – Fit of the IRT model and various other models

When we consider the by item R^2 , once again the IRT model does somewhat better than the other models. The IRT model has a by item R^2 of 0.48, which roughly means that it explains about half the variance in responses over and above what is explained by the item means. As with the root means squared residuals, the probit model does slightly worse than the IRT model (by item $R^2 = 0.435$), followed by the linear model (0.416). The dyad ratios model does worst ($R^2 = 0.397$ or 0.359 depending on whether we use linear or probit predictions). When we consider the by item R^2 adjusted by the number of parameters, the IRT model once again does best, but its adjusted R^2 is almost identical to that of the simple probit model. That is to say, the additional variance explained just about justifies the extra parameters used.

Thus the IRT model fits this data somewhat better than the probit and linear models, and considerably better than the dyad ratios scores. The improvement in fit over the simple probit model is marginal – the closer fit barely justifies the additional parameters. This is not surprising as the IRT model and the probit model are mathematically extremely similar. However, the IRT model has the added advantages of being better justified in terms of individual-level behavior and the additional parameters

(the polarization of the population in terms of policy mood) are substantively interesting. Nevertheless, the estimates of all the models are quite similar. As stated above, even the dyad ratios scores correlate very well with the estimates of the IRT model. The fact that different estimation methods produce very similar results provide some reassurance about the validity of the substantive results produced by scholars using the dyad ratios approach.

5. Item Analysis and Differential Item Functioning

In addition to considering the overall fit of the model, we can evaluate the individual item questions. This is important for three reasons. Firstly there is the question of item selection – which item questions really contribute to the measurement of the underlying construct, and which should be dropped. Secondly, there is the related question of interpreting the scale – what is it really measuring. Finally, there is the question of whether the assumptions made by the model are empirically valid. The IRT approach provides us with tools to deal with these three problems.

We could just consider the residual of each question administration or the average residual over all administrations of an item. The problem here is deciding how large a residual has to be before we reject an item. A solution to this is to ask whether the residuals are greater than we would expect if the model were true. We can test this by generating a replication data set drawn from the posterior predictive distribution of the proportion of left answers for question q in year y , using the model parameters we have estimated (see Gelman et al. 2004, 167-174). We can then calculate residuals for the

replication data set and ask how often the actual residuals exceed this. Thus the sampled test statistic is:

$$\frac{\sum_{i=1}^n T_i(y, q)}{n} \quad (15)$$

$T_i(y, q)$ has a value of true if the inequality holds and false otherwise. We can use this to test the null hypothesis that the observed residual is less or equal to the residual predicted by the model for a given administration. If $T_i(y, q)$ is true for 95% of replications, then we can reject the null hypothesis at the 5% level.

The model assumes that a given item functions in the same way no matter what year it is asked. Thus it is assumed that the item parameters λ_q and α_q are fixed for each item. This is equivalent to the assumption in item response theory that there is not differential item functioning. In item response theory it is usually assumed that a given test question has the same probability of being answered correctly by any respondent with a given ability. It is violated if a question is more difficult for one group than another (again conditional on ability). The assumption that group membership does not affect the probability of answering correctly is usually tested using either cross tables or logistic regression (see Marascuilo and Slaughter 1981; Swaminathan and Rogers 1990). In our model we are assuming that that the probability of getting a left-wing answer is the same in any year, given the policy mood of that year. We can test this assumption if we assume that our estimate of policy mood for each year is unbiased (a very non-trivial assumption). We can do this by re-estimating the model allowing the difficulty parameter

(λ_{α}) to vary for each question administration, but setting all other parameters (μ, σ, α) to values sampled from the posterior distribution derived from the original estimation. We can then test the null hypothesis that the difficulty parameter for a given question administration (λ_{α}) is equal to the single difficulty parameter estimated for the item this administration uses (λ_q), as estimated in the original model. The null hypothesis is rejected if the estimated value of the difficulty of the item lies outside the confidence limits of the posterior of the difficulty parameter of the administration.

However, we need to be careful in interpreting the results of this test. We do not know the true value of policy mood for each year, but only have estimates based on item responses. If differential item functioning is present, then the estimates of policy mood may be biased. If one particular item fails the test for differential item functioning, it may be because this item exhibits differential item functioning. However, it may also be the case that this item has failed the test due to bias in the estimate of policy mood due to differential item functioning in other items. For this reason, the presence of a significant number of items that fail the test for differential item functioning only show that differential item functioning is present in some items; it does not indicate which items violate uniform item functioning.

A final method that is useful for item analysis is simply graphing the responses over time. Consider Figure 5, which graphs the percentage of left-wing (permissive) responses to two items on abortion over time, together with the response predicted by the IRT model to a typical question. The two questions are very similar – ABORT2 asks respondent whether they think abortion should be illegal if a couple want no more

children, while ABORT12 asks respondents if they approve of abortion in this circumstance. However ABORT2 fails both the tests outlined above, while ABORT12 is flagged by neither of them. By inspecting Figure 5 we can see why this is the case. Attitudes towards abortion become more liberal over time. However ABORT12 is only asked between 1981 and 1999, a period in which policy mood happened to be moving to the left. ABORT2, however was asked until 2004, by which time policy mood had starting moving back to the right, resulting in a divergence with attitudes on abortion, which continued to move to the left. Inspecting the graphs reveals that the correlation between policy mood and ABORT12 is probably an artifact of the years in which ABORT12 was asked, something that other diagnostics would not have revealed.

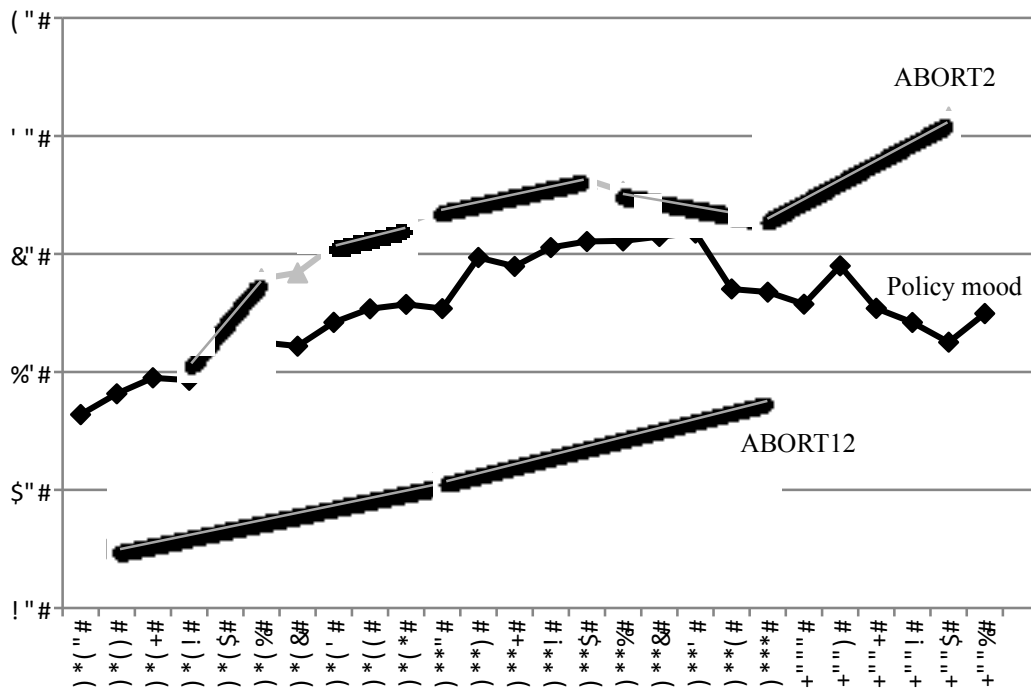


Figure 5 – Policy Mood and Two Items measuring Attitudes towards Abortion

A full discussion of item selection and scale construction for policy mood in the United Kingdom is beyond the scope of this paper. However, based on the tests outlined

here, there is reason to believe that the policy mood measure is really measuring attitudes towards economic policy rather than a general preference for left-wing policy on both social and economic issues. There are some kinds of social issues (such as environmental policy) for which all of the items fail the residuals test.¹⁰ There are others (such as abortion and capital punishment) where some items fail the residuals test, but other, very similarly worded items do not. As explained above, I suspect that this is an artifact of what years the items were asked in. For this reason I have constructed a scale that only uses items on economic issues. (I have also excluded economic items that were flagged by either test.) This scale has 128 items and is graphed in Figure 6.

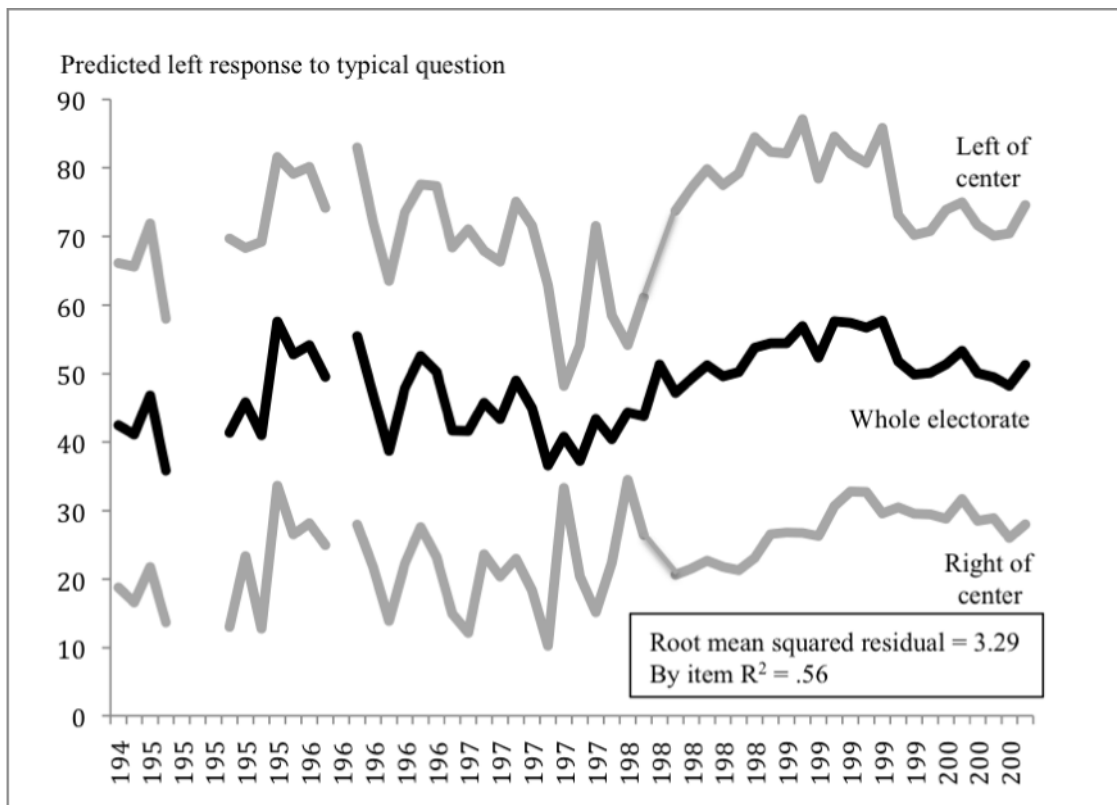


Figure 6 – Scale Measuring Role of Government in the Economy

¹⁰ A total of 121 out of 362 items failed the residuals test. 80 items failed the DIF test. However all but 11 of them also failed the residuals test.

The fit of the scale is somewhat better than that of the scale with all items in terms of root mean squared residual and by item R^2 . It is notable that the scale follows a very similar pattern to the IRT scale with all items graphed in Figure 4. This is consistent with the proposition that the all-item scale is really measuring policy mood towards economic issues. However, the new scale still exhibits evidence of differential item functioning. Of the 128 items, 27 fail the DIF test for at least one administration.

6. Conclusion

I present a model for estimating the central tendency and dispersion of public opinion over time from aggregate data. This technique is appropriate where we do not have individual level data, but only the proportion of respondent answering in a particular manner. It is not necessary for the same questions to be asked in every time period, providing there is some overlap. There are a considerable number of research questions in political behavior, political economy and comparative political institutions where we face this problem.

The main advantage of this technique over existing approaches such as the Stimson dyad ratios algorithm and the Voeten and Brewer linear model is that it is based on a plausible model of individual behavior. The other models are based on ad hoc assumptions about the relationship between aggregate responses and the central tendency of public opinion. Furthermore, the response behaviors implied by these models are implausible. The dyad ratios model treats left-wing and right-wing response in a startlingly asymmetric manner. The linear model assumes that a given change in preferences always produces the same shift in the response percentage, even in cases

where the respondents are almost certain to give left-wing responses anyway. The model presented here, on the other hand, is derived from a well established item response theory model of individual choice, which is widely used in psychometrics, political science and other fields. For that reason we can be far more confident of what we are actually measuring. The item response theory model has the added advantage of being able to estimate the dispersion of public opinion as well as its central tendency.

I provide measures of fit, and compare the performance of the IRT model to other approaches using the British public opinion data from Bartle, Dellepiane-Avellaneda and Stimson (2011a). The IRT model fits the data somewhat better than the other models (dyad ratios, linear, probit). However, all the models produce rather similar results. Nevertheless, even if the model fit of the IRT model were identical to the other models, it would still be preferable on theoretical grounds, being based on a viable model of individual choice. Indeed, the fact that the dyad ratios model produces similar results to the more theoretically justified IRT model provides some assurance that the dyad ratios model is in fact measuring what it is claimed to be measuring.

The code for the item response theory model is provided below, and can be run in the free software WinBUGS and JAGS.

Appendix A: Code

```
##Constants to be set in data or in code: startyear, endyear,  
##nquest (number of questions), len (number of administrations of questions)  
  
##variables in data: leftp (proportion answering left), year,  
##q (the number of the question asked), n (number of respondents)  
  
##Data statement is for JAGS. Omit next 5 lines if using BUGS
```

```

data{
for (i in 1:len){
leftr[i]<-round(leftp[i]*n[i]/100)
}
}

model{

##loop over the data
for (i in 1:len){
## leftr[i]<-round(leftp[i]*n[i]/100)  **Uncomment this line if using BUGS**
p[i]~dbeta(a[i], b)
a[i]<--b*m[i]/(m[i]-1)
m[i]<-phi(x[i])
x[i]<-(mu[year[i]] - lambda[q[i]]) / sqrt((alpha[q[i]])^2+(sigma[year[i]])^2)
}

##priors for this model – make sure uniform priors are wide enough not to constrain
b~dunif(0, 100)
lambda[1]<-0
alpha[1]<-0.25

for (i in 2:nquest){
lambda[i]~dunif(-10,10)
alpha[i]~dunif(0,10)
}

for (i in startyear:(endyear)){
mu[i]~dunif(-10,10)
sigma[i]~dunif(0,10)
}
}

```

Appendix B: Results

	Mean	SD	Naive SE	Time-series SE
b	2.20E+01	8.39E-01	1.53E-02	2.03E-02
mu [1947]	-1.29E-01	2.77E+00	5.06E-02	5.94E-02
mu [1948]	1.14E+00	3.84E+00	7.01E-02	1.24E-01
mu [1949]	8.79E-01	3.38E+00	6.17E-02	1.19E-01
mu [1950]	1.71E+00	4.01E+00	7.32E-02	1.21E-01
mu [1951]	-2.53E-01	3.83E+00	6.99E-02	1.28E-01
mu [1952]	5.35E+00	2.03E+00	3.70E-02	4.00E-02
mu [1953]	-2.67E+00	2.36E+00	4.30E-02	4.61E-02
mu [1954]	-1.25E-01	5.79E+00	1.06E-01	1.10E-01
mu [1955]	3.05E-01	9.86E-01	1.80E-02	1.92E-02
mu [1956]	1.86E+00	1.98E+00	3.61E-02	3.79E-02

mu [1957]	-3.91E-01	7.18E-01	1.31E-02	1.52E-02
mu [1958]	-5.93E-01	7.98E-01	1.46E-02	1.63E-02
mu [1959]	1.10E+00	8.11E-01	1.48E-02	1.57E-02
mu [1960]	9.32E-01	1.58E+00	2.89E-02	3.11E-02
mu [1961]	-2.63E-01	4.73E-01	8.64E-03	1.61E-02
mu [1962]	-2.40E+00	2.61E+00	4.77E-02	5.30E-02
mu [1963]	5.75E-01	6.04E-01	1.10E-02	1.26E-02
mu [1964]	7.94E-03	4.54E-01	8.30E-03	9.69E-03
mu [1965]	7.16E-02	6.26E-01	1.14E-02	1.28E-02
mu [1966]	3.40E-01	3.56E-01	6.50E-03	7.60E-03
mu [1967]	-7.33E-02	4.69E-01	8.56E-03	1.08E-02
mu [1968]	-4.42E-01	4.41E-01	8.06E-03	1.12E-02
mu [1969]	3.63E-01	7.00E-01	1.28E-02	1.31E-02
mu [1970]	-1.16E+00	7.02E-01	1.28E-02	1.36E-02
mu [1971]	5.14E-01	9.90E-01	1.81E-02	2.12E-02
mu [1972]	-1.27E+00	5.50E-01	1.00E-02	1.12E-02
mu [1973]	-4.70E-01	4.82E-01	8.79E-03	9.78E-03
mu [1974]	-8.98E-01	2.17E-01	3.96E-03	6.88E-03
mu [1975]	-2.03E+00	2.66E-01	4.85E-03	9.48E-03
mu [1976]	-2.08E+00	2.39E-01	4.37E-03	9.65E-03
mu [1977]	-1.72E+00	2.40E-01	4.38E-03	8.54E-03
mu [1978]	-1.41E+00	2.12E-01	3.87E-03	7.83E-03
mu [1979]	-1.44E+00	1.97E-01	3.59E-03	7.17E-03
mu [1980]	-9.19E-01	1.88E-01	3.44E-03	7.87E-03
mu [1981]	-6.51E-01	1.83E-01	3.35E-03	5.96E-03
mu [1982]	-4.34E-01	1.73E-01	3.16E-03	5.48E-03
mu [1983]	-4.68E-01	1.37E-01	2.50E-03	4.65E-03
mu [1984]	-3.50E-01	1.56E-01	2.84E-03	4.85E-03
mu [1985]	4.11E-02	1.34E-01	2.44E-03	3.98E-03
mu [1986]	-1.14E-02	1.35E-01	2.47E-03	3.90E-03
mu [1987]	3.29E-01	1.52E-01	2.78E-03	4.74E-03
mu [1988]	5.02E-01	2.48E-01	4.53E-03	5.35E-03
mu [1989]	5.90E-01	1.69E-01	3.08E-03	4.87E-03
mu [1990]	5.25E-01	1.59E-01	2.90E-03	5.15E-03
mu [1991]	1.44E+00	2.30E-01	4.20E-03	6.39E-03
mu [1992]	1.15E+00	2.06E-01	3.76E-03	5.68E-03
mu [1993]	1.57E+00	2.13E-01	3.88E-03	5.37E-03
mu [1994]	1.64E+00	1.94E-01	3.54E-03	5.42E-03
mu [1995]	1.68E+00	2.15E-01	3.92E-03	5.56E-03
mu [1996]	1.70E+00	2.08E-01	3.80E-03	5.92E-03
mu [1997]	1.68E+00	2.06E-01	3.76E-03	5.77E-03
mu [1998]	7.82E-01	1.79E-01	3.27E-03	4.72E-03
mu [1999]	7.42E-01	2.15E-01	3.92E-03	5.80E-03
mu [2000]	5.51E-01	1.89E-01	3.46E-03	5.47E-03
mu [2001]	1.17E+00	2.29E-01	4.19E-03	6.80E-03
mu [2002]	5.04E-01	2.18E-01	3.97E-03	6.40E-03
mu [2003]	3.56E-01	2.51E-01	4.59E-03	6.16E-03
mu [2004]	4.27E-02	1.62E-01	2.96E-03	3.92E-03
mu [2005]	4.92E-01	2.05E-01	3.75E-03	5.44E-03
deviance	1.92E+04	6.81E+01	1.24E+00	1.28E+00
sigma [1947]	5.63E+00	2.85E+00	5.20E-02	5.45E-02
sigma [1948]	5.60E+00	2.86E+00	5.22E-02	5.14E-02
sigma [1949]	5.88E+00	2.81E+00	5.13E-02	5.25E-02
sigma [1950]	5.68E+00	2.82E+00	5.14E-02	4.36E-02
sigma [1951]	5.66E+00	2.84E+00	5.18E-02	5.26E-02
sigma [1952]	5.40E+00	2.87E+00	5.24E-02	5.07E-02
sigma [1953]	6.49E+00	2.58E+00	4.71E-02	4.65E-02

sigma[1954]	4.94E+00	2.90E+00	5.29E-02	5.80E-02
sigma[1955]	5.39E+00	2.17E+00	3.96E-02	4.46E-02
sigma[1956]	5.52E+00	2.88E+00	5.26E-02	5.05E-02
sigma[1957]	5.34E+00	2.10E+00	3.84E-02	3.93E-02
sigma[1958]	2.12E+00	1.96E+00	3.58E-02	3.71E-02
sigma[1959]	3.68E+00	2.41E+00	4.40E-02	4.51E-02
sigma[1960]	5.54E+00	2.86E+00	5.22E-02	4.84E-02
sigma[1961]	7.76E-01	5.35E-01	9.77E-03	1.36E-02
sigma[1962]	5.47E+00	2.88E+00	5.27E-02	5.29E-02
sigma[1963]	4.22E+00	1.99E+00	3.63E-02	3.63E-02
sigma[1964]	5.31E+00	1.30E+00	2.38E-02	2.59E-02
sigma[1965]	2.14E+00	9.04E-01	1.65E-02	2.16E-02
sigma[1966]	3.71E+00	1.30E+00	2.38E-02	3.14E-02
sigma[1967]	1.47E+00	7.04E-01	1.29E-02	2.07E-02
sigma[1968]	1.54E+00	7.44E-01	1.36E-02	2.25E-02
sigma[1969]	4.22E+00	1.80E+00	3.28E-02	3.35E-02
sigma[1970]	6.38E+00	1.76E+00	3.22E-02	3.41E-02
sigma[1971]	4.77E+00	2.62E+00	4.78E-02	4.94E-02
sigma[1972]	4.64E+00	1.75E+00	3.19E-02	3.30E-02
sigma[1973]	5.45E+00	1.71E+00	3.12E-02	3.29E-02
sigma[1974]	1.83E+00	5.82E-01	1.06E-02	1.57E-02
sigma[1975]	2.24E+00	4.79E-01	8.75E-03	9.89E-03
sigma[1976]	2.03E+00	3.81E-01	6.95E-03	8.22E-03
sigma[1977]	1.52E+00	4.30E-01	7.85E-03	1.06E-02
sigma[1978]	1.22E+00	4.59E-01	8.38E-03	1.27E-02
sigma[1979]	1.42E+00	4.80E-01	8.77E-03	1.28E-02
sigma[1980]	7.33E-01	4.12E-01	7.52E-03	1.42E-02
sigma[1981]	1.56E+00	5.11E-01	9.33E-03	1.56E-02
sigma[1982]	6.13E-01	4.08E-01	7.45E-03	1.96E-02
sigma[1983]	1.27E+00	3.64E-01	6.65E-03	1.68E-02
sigma[1984]	8.08E-01	4.49E-01	8.19E-03	1.73E-02
sigma[1985]	1.25E+00	3.25E-01	5.93E-03	1.64E-02
sigma[1986]	1.30E+00	3.34E-01	6.10E-03	1.66E-02
sigma[1987]	2.09E+00	3.69E-01	6.73E-03	1.70E-02
sigma[1988]	1.66E+00	3.45E-01	6.30E-03	1.28E-02
sigma[1989]	2.30E+00	4.21E-01	7.69E-03	1.71E-02
sigma[1990]	2.17E+00	2.68E-01	4.90E-03	1.27E-02
sigma[1991]	3.95E+00	3.97E-01	7.25E-03	1.37E-02
sigma[1992]	2.55E+00	3.29E-01	6.01E-03	1.26E-02
sigma[1993]	3.74E+00	3.21E-01	5.86E-03	1.19E-02
sigma[1994]	3.56E+00	3.28E-01	5.99E-03	1.07E-02
sigma[1995]	3.77E+00	3.26E-01	5.96E-03	1.09E-02
sigma[1996]	3.43E+00	3.02E-01	5.51E-03	1.19E-02
sigma[1997]	2.88E+00	3.07E-01	5.61E-03	1.14E-02
sigma[1998]	1.83E+00	3.29E-01	6.00E-03	1.30E-02
sigma[1999]	1.91E+00	4.18E-01	7.63E-03	1.39E-02
sigma[2000]	1.24E+00	4.03E-01	7.36E-03	1.80E-02
sigma[2001]	2.68E+00	5.11E-01	9.34E-03	1.80E-02
sigma[2002]	1.54E+00	3.88E-01	7.07E-03	1.90E-02
sigma[2003]	2.80E+00	7.00E-01	1.28E-02	1.82E-02
sigma[2004]	1.40E+00	4.72E-01	8.61E-03	1.99E-02
sigma[2005]	2.94E+00	5.78E-01	1.06E-02	1.89E-02

Table A1 – Parameter Estimates of IRT model (see text)

Bibliography

- Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan. 2005. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13 (2):171-187.
- Bafumi, J., and M. 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.
- Bartle, John, Sebastian Dellepiane-Avellaneda, and James Stimson. 2011a. The Moving Centre: Preferences for Government Activity in Britain, 1950–2005. *British Journal of Political Science* 41 (2):259-285.
- Bartle, John, Sebastian Dellepiane-Avellaneda, and James A. Stimson. 2011b. The Policy Mood and the Moving Centre. In *Britain at the polls 2010*, edited by N. Allen and J. Bartle. London: SAGE.
- Baumgartner, Frank R., Suzanna L. De Boef, and Amber E. Boydston. 2008. *The Decline of the Death Penalty and the Discovery of Innocence*. Cambridge: Cambridge University Press.
- Chanley, Virginia A., Thomas J. Rudolph, and Wendy M. Rahn. 2000. The Origins and Consequences of Public Trust in Government: A Time Series Analysis. *Public Opinion Quarterly* 64:239-256.
- Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review* 98 (2):355-370.
- Cohen, Jeffrey E. 2000. The Polls: Public Favourability towards the First Lady. *Presidential Studies Quarterly* 30:575-585.
- Erikson, Robert S., Michael MacKuen, and James A. Stimson. 2002. *The Macro Polity, Cambridge studies in political psychology and public opinion*. New York: Cambridge University Press.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian data analysis*. 2nd ed, *Texts in statistical science*. Boca Raton, Fla.: Chapman & Hall/CRC.
- 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. 2001. Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking. *Political Analysis* 9 (3):227-241.
- Jackman, Simon. 2005. Pooling the Polls Over an Election Campaign. *Australian Journal of Political Science* 40 (4):499-517.
- Jessee, S. A. 2009. Spatial Voting in the 2004 Presidential Election. *American Political Science Review* 103 (1):59-81.
- Kellstadt, Paul. 2003. *The Mass Media and the Dynamics of American Racial Attitudes*. Cambridge: Cambridge University Press.
- Kim, Hee Min, and Richard Fording. 2001. Extending Party Estimates to Voters and Governments. In *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*, edited by I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum. Oxford: Oxford University Press.

- Levendusky, M. S., and J. C. Pope. 2010. Measuring Aggregate-Level Ideological Heterogeneity. *Legislative Studies Quarterly* 35 (2):259-282.
- Marascuilo, Leonard, and Robert Slaughter. 1981. Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational Measurement* 18 (4):229-248.
- 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.
- McDonald, Michael, and Ian Budge. 2005. *Elections, Parties, Democracy: Conferring the Median Mandate*. Oxford: Oxford University Press.
- McGann, Anthony. 2013. "Replication data for: Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm", <http://dx.doi.org/10.7910/DVN/22861> IQSS Dataverse Network [Distributor] V1 [Version]
- Nunnally, Jum C., and Ira H. Bernstein. 1994. *Psychometric Theory*. 3 ed. New York.: McGraw-Hill.
- Peress, Michael. 2009. Small Chamber Ideal Point Estimation. *Political Analysis* 17 (3):276-290.
- Poole, Keith, and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford: Oxford University Press.
- Powell, G. Bingham Jr. 2000. *Elections as Instruments of Democracy: Majoritarian and Proportional Visions*. New Haven: Yale University Press.
- Stimson, James A. 1991. *Public opinion in America : moods, cycles, and swings, Transforming American politics*. Boulder: Westview Press.
- Stimson, James A. 1999. *Public opinion in America : moods, cycles, and swings*. 2nd ed, *Transforming American politics*. Boulder: Westview Press.
- Stimson, James A., Michael B. Mackuen, and Robert S. Erikson. 1995. Dynamic Representation. *American Political Science Review* 89 (3):543-565.
- Stimson, James, Cyrille Thiébaud, and Vincent Tiberj. 2009. The Structure of Policy Attitudes in France. In *Annual Meetings of the Midwest Political Science Association*. Chicago.
- Swaminathan, Hariharan, and H. Jane Rogers. 1990. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement* 27 (4):361-370.
- Voeten, Erik, and Paul R. Brewer. 2006. Public Opinion, the War in Iraq, and Presidential Accountability. *Journal of Conflict Resolution* 50:809-830.