When Does Joint Scaling Allow For Direct Comparisons of Preferences?*

Jeffrey B. Lewis
UCLA
Department of Political Science
jblewis@ucla.edu

Chris Tausanovitch
UCLA
Department of Political Science
ctausanovitch@ucla.edu

April 28, 2015

Abstract

Achen's (1978) famous critique of Miller and Stokes (1963) shows that correlations between the policy stances taken by legislators and the policy stances taken by constituents do not establish whether or not these policy stances are proximate to one another. In general, proximity cannot be established when the two measures are not on the same scale. The recent literature on joint scaling proposes a solution to this problem. Indeed, the proposed solution is general enough that it has been applied to a variety of contexts, comparing the political positions of everyone from interest groups (Bonica, 2013) to twitter users (Barberá, 2015). We show that joint scaling works well in contexts where the underlying assumptions are correct. However, it fails in some of its most well-known applications. In particular, we show that methods for jointly scaling between media outlets and legislators, and between legislators and constituents, are problematic.

*Prepared for the 2015 Conference on Ideal Point Models at MIT. We thank participants in the Vanderbilt Center for the Study of Democratic Institution's Conference on Political Representation Fifty Years After Miller and Stokes. We would like to extend our thanks to Joseph Bafumi, Matthew Gentzkow, Tim Groseclose, Michael Herron, Stephen Jessee, Jeffrey Milyo, and Jessee M. Shapiro for sharing the data on which this paper is based. The authors benefited greatly from a draft of a paper by Stephen Jessee that shares some of the insights of the current paper.

> The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

— John Tukey

1 Introduction

Preferences are one of the most important primitives in political science. As a discipline, political science is about understanding preferences and how groups of people can solve the problems that arise when preferences conflict. The measurement of preferences has taken its place as an important and vital subfield of the discipline. Some of the most interesting applications in this young subfield arise when the political preferences of very different political actors are at stake. For example, the study of representation is about how the preferences of citizens affect the actions of representatives. To take another example, a central question in the study of government institutions is who gets their way when different branches of government disagree. All of these questions suffer from a key problem: it is difficult to measure preferences in a comparable fashion across diverse political settings.

A second problem arises from the simple matter of measuring preferences for any one group. Political scientists have long been attuned to the nuances of individual preferences. Choices are made with error and bias. While economics hews closely to the notion of "revealed preferences," political science recognizes that sometimes preferences are revealed in only a trivial sense (Zaller, 1992). When an individual faces the same choice a large number of times, we may learn on average what they prefer. However, if this choice is a narrow one, we can only make use of this measure in a limited number of political situations. Understanding which choice is preferred in a single situation is far less useful than understanding the basis of a wide variety of choices.

The method of joint scaling was developed to solve these twin problems. Creating preference scales by reducing the dimensionality of a large choice space allows scholars to apply these measures with a high degree of generality. The fact that multiple indicators are used reduces the degree of error in the resulting measure (Ansolabehere, Rodden and Snyder, 2008). Finally, if disparate groups face some of the same choices, then these preference spaces can be bound together. For instance, if voters from two different states respond to the same survey questions, then their preferences can plausibly be compared. Joint scaling solves the measurement problem *and* the comparability problem.

Creative and innovative applications abound among top scholars in political science. The method of joint scaling has been used to study members of different chambers of Congress across years (Poole and Rosenthal, 1997), the ideologies of government agencies (Clinton et al., 2012), the ideologies of presidents (Clinton, Jackman and Jackman, 2013; Treier, 2009), the preferences of judges (Bailey, 2007), state legislators (Kousser, Lewis and Masket, 2007; Shor and McCarty, 2011), Twitter users (Barberá, 2015), Facebook users (Bond and Messing, 2015), campaign donors (Bonica, 2013), voters (Jessee, 2009; Bafumi and Herron, 2010), voters across history and across different survey instruments (Hill and Tausanovitch, 2014; Tausanovitch and Warshaw, 2013), the media (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2010), and even legislators and the electorates of disparate countries (Lo, Proksch and Gschwend, 2014).

While applications of joint scaling abound, tests of the underlying assumptions have been rare (but for an exception, see Jessee (2015)). In particular, the assumption that allows for comparability across groups is that these groups face the very same choices. This is rarely true in a literal sense. When a voter expresses a preference on an issue, and a newspaper editorial board writes an editorial expressing a similar preference, this is quite a different act. The voter expresses this view in a low-stakes and most likely low-information environment. The newspaper editorial board carefully researches their view, and understands that expressing their view publicly could have a broad impact. It may or may not be the case, in such circumstances, that the act of expressing this particular view has the same relationship with the underlying preferences or ideology of the voter and the editorial board as a group. Assuming that this is the same choice is a strong assumption. This “constant item” (or “constant behavior”) assumption is the underpinning of joint scaling exercises. The purpose of this paper is to demonstrate how this assumption can be tested using existing tools, and to provide some intuition for the severity of the problem when the assumption fails.

As test cases we use two of the most well-known and important applications. The first case is that of jointly scaling media outlets and legislators or their constituents. We test these assumptions in two highly cited papers that use somewhat different data and methods: Groseclose and Milyo (2005) and Gentzkow and Shapiro (2010). Our second case is that of jointly scaling legislators and citizens. Jessee (2009) jointly scale legislators and citizens in order to understand spatial voting. Bafumi and Herron (2010) jointly scale legislators and citizens in order to examine questions of representation. These are highly cited papers asking fundamental questions. We find that the joint scaling assumptions fall short in all four of these papers, to such a degree that the primary conclusions are in doubt.

Simple hypothesis tests can be too stringent when parsimonious models deviate in small ways from actual data generating processes. In order to show that our results are substantively important, we compare the magnitude of the problems from the joint scaling assumptions to other model constraints. We demonstrate how the severity of violations of the joint scaling assumptions can be evaluated in the context of the comparisons the researcher desires to make.

We start by reviewing the literature on scaling and joint scaling. In the following sections we explain the models and methods used in the media and representation applications. We show that the underlying assumptions fail in the context of jointly scaling legislators and the media as well as legislators and citizens, respectively. Then we confirm that there are contexts where joint scaling works. We conclude with some remarks about the generality of these results and speculation about what can be done about them.

2 Literature

Joint scaling is not really a method at all, but an application of scaling methods to sparse datasets. In common usage, “joint” scaling (Poole, 2007), “bridging” (Shor, Berry and McCarty, 2010) or “gluing” (Poole and Rosenthal, 2001) consists of applying scaling to populations that would not normally encounter the same types of choice situations. Most commonly, the scale itself is based off of explicit statements of preference. When a legislator votes on a bill, she has to state whether she is for or against (or wishes to abstain). However, in some cases the scale is based on a behavior. For instance, it may be assumed that whether and to whom one gives political donations is motivated by political preferences (Bonica, 2013), or that what words a party uses in their manifesto is similarly motivated (Gabel and Huber, 2000). A distinguishing feature of these methods is that the scales are estimated via an explicit model of choice. It is still common for researchers to simply combine measurements using their own judgement (e.g. Abramowitz and Saunders, 2008; Volden, Wiseman and Wittmer, 2013) but in these cases underlying assumptions cannot be tested.

Although a voluminous literature has spawned in recent years, the joint scaling approach is not new. Two famous examples are that of Poole and Rosenthal (1984), who joined multiple years of congressional preferences using legislators that remain in their respective chambers, and Aldrich and McKelvey (1977) , who use the fact that legislators are an object of voter choice to jointly scale voters and candidates. These seminal works differ from much of the current scholarship (including a few of our own papers) in being self-consciously circumspect about the joint scaling assumptions. In more recent work, Poole and Rosenthal (1997) heap caution on comparisons over long stretches of time. As far back as Aldrich and McKelvey (1977), these authors explicitly model the possibility of differential scale use across individuals.

The necessity of joint scaling arose from the study of political representation.Miller and Stokes (1963) is the paper that arguably began the era of large-scale, empirical studies of ideological representation in political science. Miller and Stokes were the first to collect data on the political positions of both legislators and their constituents and to reduce that data into a set of scales that could be compared. Although few doubted the extent of Miller and Stokes' innovation, it was criticized on methodological grounds. Many of these problems, such as the critique that the sample sizes used were too small (Erikson, 1978), have been overcome in more recent work. However, Achen (1977) posed a challenge that has lasted until the present day. Miller and Stokes' analysis rests upon correlations between the positions of legislators and the positions of constituents. Achen (1977) pointed out that these correlations do not imply that legislators are actually taking positions that are proximate to the positions of their constituents, even if the correlations are very high. His 1978 paper proposes an alternative measure which he calls centrism: the squared distance between the legislator's position and the mean position of their constituents. This measure represents the extent to which legislators take positions that are close to the center of the distribution of their constituents' positions.

To see why correlations are poor measures of representation, consider a linear regression of a measure of legislator preferences, x^l on an equivalent measure of mean (or median) voter preferences, x^m, where i indexes districts. Assume that the true relationship between these variables is as follows, where ϵ is a normally distributed error term:

$$x_i^l = \gamma_1 + \gamma_2 x_i^m + \epsilon$$

Now consider a different measure of mean voter preferences, x^{m*}, such that $x^{m*} = \delta_1 + \delta_2 x^m$ and the values of δ_1 and δ_2 are unknown. The relationship between these measures is:

$$x_i^l = \gamma_1^* + \gamma_2^* x_i^{m*} + \epsilon$$

where $\gamma_2^* = \gamma^2/\delta_2$ and $\gamma_1^* = \gamma_1 - \gamma_2\delta_1/\delta_2$. If δ_2 is less than 1, then the slope of this relationship will be increased even though the underlying behavior has not changed. If δ_2 is greater than 1, then the slope will be reduced. Using an estimate of γ_2^* as a measure of representation will conflate the scale of the measurement with the strength of the relationship. In fact, the only meaningful hypothesis that can be tested with such an estimate is the hypothesis that $\gamma_2^* = 0$, in which case it must also be the case that $\gamma_2 = 0$.

Even if the common scale values of x^l and x^m were known it is difficult in practice to distinguish between representation of the median or mean and representation of other quantiles of the distribution of constituents by fitting regressions of the form above. Romer and Rosenthal (1979) show that it is very hard to statistically differentiate between the correlation of legislators positions with the median and the correlation of legislator positions with arbitrary quantiles of the constituent distribution. The cause of this problem is that medians and other quantiles of the distribution of constituent preferences are themselves highly correlated.

If researchers had a common measure for both legislators and constituents, then we could use Achen's proposed "centrism score." Computing the centrism score would be simple: just take the squared of absolute value difference between the two measures of preferences. Indeed, Miller and Stokes attempt to use a common measure insofar as they use similar instruments to measure the preferences of legislators and constituents. If we believed that the questions Miller and Stokes achieve a common scale then we could stop there and use Achen's centrism score to measure representation. However, the public opinion literature has pointed out some problems with using a single measure of constituent opinion. Scholars since Converse (1964) have pointed out that individual preferences are much more prone to error than legislator preferences, and may have a different structure altogether. Ansolabehere, Rodden and Snyder (2008) note that scales of voter ideology based on multiple measures demonstrate much more ideological behavior than measures based on only one measure, such as those used by Miller and Stokes. If even similar instruments do not yield a common measure when applied to different groups, then proximity comparisons using these measures are not valid.

More than 30 years after Achen's papers, joint scaling has become the reigning solution to the problem of comparing legislators to constituents, and a host of other similar problems. The first paper to apply this method to representation was Bafumi and Herron (2010). Bafumi and Herron linked the responses of legislators and constituents by asking survey respondents to take positions on items that legislators had voted on. By assuming that these responses are equivalent to roll call votes, they link the two populations, and made proximity comparisons between them. This work drew from Jessee's (2009) elegant solution to the problem of spatial voting using similar data. However, the reservations expressed by past works in this genre (Aldrich and McKelvey, 1977; Poole and Rosenthal, 1997) was notably absent from these papers.

The promise of joint scaling is evident, and it is only natural that it should be extended to a wide variety of fields, as it has been. The innovation in many of these examples is to recognize that preferences can be revealed by a wide variety of behaviors, not just binary choices between clearly defined options. However, the joint scaling assumption is even bolder, and the degree of circumspection has often been less.

We begin by demonstrating that joint scaling has very apparent problems in the case of scaling media outlets and legislators. Then we move to the more complicated case of legislators and citizens.

3 Joint Scaling Media Outlets and Legislators or Districts

A question of enduring interest in the study of journalism is whether and to what degree to news reports contain an ideological bias or "slant." Are news stories simply presentations of fact or do they contain an (implicit) ideological point of view? If they have a viewpoint, what is it and how does it compare to that of the public? And, to what degree can variation in slant across news outlets be attributed to variation in consumers' tastes within the markets that the media outlets serve. Answers to these questions require a reliable measure of slant on an ideological or partisan scale that is comparable to similar measurements for voters or other political actors.

Recent papers by Groseclose and Milyo (2005) and Gentzkow and Shapiro (2010) provide similar and ingenious strategies for making slant measurements in a rigorous and objective way. Both approaches begin by positing stable mappings between ideology and the use specific words or terms in speech or writing. In the case of Groseclose and Milyo, the choice of authority (think tank or interest group) to cite is modeled by a standard random utility formulation in which the utilities associated with citing each think tank or group are taken to be a linear functions of the speaker's (or writer's) ideology. In Gentzkow and Shapiro, the relative frequency with which speakers employ partisan-charged phrases, such as "death tax" or "oil companies," are modeled in a reduced form as linear functions of a speaker's ideology plus a stochastic shock. In both cases, the parameters of the mappings between ideology and speech are identified through observations of the citations made or phrases used in the speeches of members of Congress.

Because the ideologies of members of Congress can be directly measured by their roll call voting (as in Groseclose and Milyo) or the support in their districts for the Republican presidential candidate (as in Gentzkow and Shapiro), the mapping between members' ideologies and their use of citations or partisan phrases can be inferred. Armed with a means of establishing a mapping from ideology to speech and assuming that this mapping is similar for media outlets and legislators, the ideologies of news outlets (which are not otherwise observable) can be backed out from their (observable) use of citations and partisan phrases.

The assumption of a common mapping allows these authors to build bridges between the known ideologies of legislators and the ideologies of the news outlets that they wish to infer. This is a strong assumption that goes untested by both Groseclose and Milyo and Gentzkow and Shapiro. Gentzkow and Shapiro explicitly describe their slant estimates in terms of this assumption in an as-if way writing that their estimates answer the question "if a given newspaper were a congressperson, how Republican would that newspaper's district be?" (2010, p. 46). They focus on the ordering of the newspapers from left to right describing their estimates as "index[ing] newspapers by the extent to which the use of politically charged phrases in their news coverage resembles the use of the same phrases in the speech of a congressional Democrat or Republican" (2010, p. 36). Here and in their subsequent analyses, they do not assert direct comparability of the legislator and media positions and rely on the bridging assumption only as a means of arraying the media outlets. Groseclose and Milyo lean more heavily on the bridging assumption by directly comparing estimated media ideologies to those of members of Congress in both interval and ordinal terms. They find that "Our results show a strong liberal bias: all of the news outlets we examine, except Fox News' Special Report and the Washington Times, received scores to the left of the average member of Congress." (2005, p. 1191). The difference is subtle, but important and conceptual. In one case, the bridge is taken as providing a way to identify which legislator a newspaper *most sounds like* and in the other the estimates reveal which legislator a newspaper is *most ideologically similar to*.

In either case, however, the bridging assumption is central and it can be explored empirically. The bridging assumption is equivalent to the assumption that the legislative data and the media data can be pooled and are governed by the same parameter values. In the alternative, the sets of observations do not pool in which case a distinct set of parameter values govern each data set. These two alternatives are nested allowing the validity of the pooling assumption to be measured by the loss of fit associated with restricting the model parameters to be common across the two groups. Interestingly, as we will see, this fit comparison can be made despite the fact that the unpooled alternative estimates of media ideology lie on an arbitrary scale and are, therefore, not directly comparable to the observable ideologies of the legislators.

In order to examine the admissibility of the bridging assumptions in Groseclose and Milyo and Gentzkow and Shapiro, we first lay out the underlying models and their estimators and then explore their data in turn. We begin with some definitions. Let i index a set of speakers ($\mathcal{I}$) that include both legislators ($\mathcal{I}_c$) and media outlets ($\mathcal{I}_m$). Let j index a set of J think tanks in Groseclose and Milyo or partisan-phrases in Gentzkow and Shapiro. The number of times that the ith speaker mentions the j think tank or phrase is $c_{ij} = \sum_{ik} y_{ijk}$ where $y_{ijk} = 1$ if member i mentions think tank or phrase j at her kth opportunity and 0 otherwise. The range of the citation opportunity index, k, depends on i and is equal to $i = 1, \ldots, K_i$ where K_i is the total number of mentions or citations made by the ith speaker. Finally, let x_i be the observed ideology (ADA score or 2004 two-party vote share for George W. Bush in the district) for $i \in \mathcal{I}_c$ (legislators) or the latent ideology that we seek to estimate for $i \in \mathcal{I}_m$ (media outlets or newspapers).

Groseclose and Milyo develop a simple structural model of citation choice. In their model, the utility that legislator i receives from citing the jth think tank at her kth opportunity is

$$a_j + b_j x_i + e_{ijk}$$

implying that the utility associated with citing each think tank is a think-tank specific linear function of ideology. The random utility components, e, are assumed to be i.i.d. and Weibull implying the well-known multinomial logistical (MNL) choice model of McFadden (1974) in which the outcome of each citation choice $y_{ik} = (y_{i1k}, \ldots, y_{iJk})$,

$$y_{ik} \sim \text{Multinomial}(\pi_i)$$

and the elements of the vector of choice probabilities π_i are

$$\pi_{ij} = \frac{\exp(a_j + b_j x_i)}{\sum_{j'=1}^{J} \exp(a_{j'} + b_{j'} x_i)}.$$

For the legislators, the as and bs can be estimated by the multinomial logit regression of y on x. In order to identify the underlying utility scale, Groseclose and Milyo follow convention in normalizing a_1 an b_1 (corresponding to the Heritage Foundation) to 0. Because the media xs are parameters to be estimated, Groseclose and Milyo cannot apply canned MNL estimation routines. Rather, they program the likelihood as a function of the full parameter vector ($a_2, \ldots, a_J$, $b_2, \ldots, b_j$, and x_i for $i \in \mathcal{I}$) which they then directly maximize by numerical optimization. With 50 think thanks and 20 media outlets and over 22 thousand observed citation choices, maximizing the likelihood directly in this way is computationally intensive and time consuming.[1] In probing the admissibility of the bridging assumption, we will need to refit the model many times and require a more efficient estimation approach. Fortunately, the estimation problem can be reformulated in a way that is equivalent, less computationally intensive, and makes clear that the model is much more similar to the approach of Gentzkow and Shapiro and others that have scaled ideology based on text (e.g., Slapin and Proksch, 2008) than is otherwise apparent.

Using the multinomial Poisson transformation that is often used to estimate multinomial models (see Baker, 1994), the Groseclose and Milyo model can be rewritten in terms of citation counts:

$$c_{ij} \sim \text{Poisson}(\lambda_{ij})$$

where

$$\lambda_{ij} = \exp(\mu_i + a_j + b_j x_i).$$

The Poisson and MNL formulations are equivalent in the sense that the parameter estimates and associated uncertainty estimates of the as, bs, and xs are identical. If x is observed, the other parameters can be estimated by a panel Poisson regression. The additional parameters, μ_i for each speaker, account for differences in the total number of citations (K_i) made across speakers. Interestingly, the random utility model of individual citation choices leads to citation counts that can be treated as a linear function of ideology and a Poisson link. The connection between the MNL and Poisson formulation usefully provides a compelling behavioral foundation for those who have modelled counts of words as linear functions of ideology with a Poisson link (e.g., Slapin and Proksch 2008 or Bonica 2013).

Estimates of the model can be found iteratively via the following steps:

[1] Groseclose and Milyo report that finding estimates and calculating standard errors by inverting the numerically-approximated Hessian required nearly 24 hours [replication materials]).

Step 0: Initialize unobserved media ideologies (x_i for $i \in \mathcal{I}_m$) as missing values.

Step 1: Fit a panel Poisson regression of c on x to recover $\hat{\mu}$s, $\hat{a}$s, and $\hat{b}$s omitted observations with missing values of x.

Step 2: Estimate media xs via Poisson regressions of of $c_{i\cdot}$ on $\hat{b}$ including a as an offset for each media outlet, $i \in \mathcal{I}_m$.

Step 3: Set $x_i = \hat{x}_i$ for $i \in \mathcal{I}_m$ and repeat from (1) until convergence.

A single pass through the steps provides consistent estimates of all model parameters (as the number of legislators and groups being cited grows large). Iterating until convergence provides the maximum likelihood estimates reported by Groseclose and Milyo. In the ML approach the media observations contribute to the estimation of the think tank parameters (as and the bs). In a simpler, two-step approach they do not.

Rather than motivate the mapping between ideology and the frequency which various words are used via a utility model, Gentkow and Shapiro simply assert the linear reduced form,

$$\tilde{f}_{ij} = a_j + b_j x_i + e_{ij}$$

where $\tilde{f}_{ij}$ is the frequency with which speaker i uses phrase j relative to other phrases ($\tilde{f}_{ij} = c_{ij}/\sum_{j'} c_{ij'}$ for all $i \in \mathcal{I}$). Normalizing the word counts to fractions that sum to one for each speaker accounts for differences in the amount speech each speaker engages in thus no speaker effects are included in the model. Gentzkow and Shapiro make no specific assumption about the form of the es other than they are mean independent of the xs.

They estimate their model via least squares using a two-step procedure:

Step 1: Regress $\tilde{f}_{\cdot j}$ on x for each phrase $j \in \{1, \ldots, J\}$ to recover $\hat{a}$s and $\hat{b}$s.

Step 2: Regress $\tilde{f}_{i\cdot} - a$ on $\hat{b}$ (with no intercept) for each newspaper, $i \in \mathcal{I}_m$, to recover $\hat{x}$s.

If the ("missing") newspaper xs are replaced by $\hat{x}$s from (2), the algorithm can be iterated to convergence. If we further assume that the es are independent and distributed normally with constant variance then this iterated solution is also ML.

Two differences between the models of Gentzkow and Shapiro and Groseclose and Milyo are now clear: the link (Poisson versus linear) and the decision to iterate a two-step estimation algorithm to convergence. Of these, the second is perhaps more consequential. Under the assumption that the speech of legislators and media outlets pool, the decision to iterate to convergence is not consequential and Groseclose and Milyo, who bring information from the media observations to bear on the estimation of the as and bs, more efficiently use the information available in the data. However, if the two data sources do not pool then Gentzkow and Shapiro still provide what they promise, an imputation of which sort of legislator each newspaper reads like (that is, an imputation that employs the mapping between ideology and speech that is estimated to hold for legislators). If the pooling fails, Groseclose and Milyo's full ML solution generates estimates force a common mapping set of parameters that average between those holding for the legislators and those that hold for the media and estimated ideologies of the media and the legislature are not directly comparable. They cannot be interpreted in an as-if way.

4 Media Slant Data

As discussed above, the Groseclose and Milyo (2005) data consists of citation made by both legislators and media outlets to think tanks. The legislator citations are made during floor speeches, and the media citations are made in news articles (not editorials, letters or reviews). Legislators speeches are drawn from the years 1993 through 2002, and citations from media outlets are drawn from varying lengths of time needed to generate at least 300 citations per outlet. Scores from Americans fo Democratic Action (ADA), adjusted for overtime comparability, are taken as measures of legislator ideology. These scores are based on the percentage of times a member votes on the liberal side of a set of votes selected by the ADA each year.

The dataset that results includes 22,170 citations to 168 think tanks made by roughly 800 legislators and 20 media outlets. As shown above, these think tank citations can be thought of much as words are thought of in Gentzkow and Shapiro (2010): each think tank is an indicator of ideological affiliation. Groseclose and Milyo (2005) attempt to capture the universe of influential think tanks rather than subsetting beforehand to ones that are cited more often for ideological reasons.

In contrast, Gentzkow and Shapiro (2010) limit the phrases they use to the 1,000 that show the most partisan patterns of use. The full set of potential phrases is also limited to two-word phrases that appear in newspaper headlines between 200 and 15,000 times and three-word phrases that appear between 5 and 1,000 times. All newspaper data comes from the period from 2000 to 2005. The two-party presidential vote share going to George Bush in a politicians' district in 2004 is used as the measure of congressional ideology.

4.1 Assessing the bridging assumptions in Groseclose and Milyo and Gentkow and Shapiro

The above discussion hints at related ways in which bridging assumption can be tested. First, we see that in either case, we could fit the media and legislative data sets separately and compare the fit using a likelihood ratio test or information criterion such as AIC. Alternatively, one could construct a Hausman-like test in which the less efficient two-step estimates are compared to the full ML estimates. However, given the large amount of data employed in these studies, we might be prone to reject bridging even in cases in which the substantive effect of the violation is trivial. Accordingly, we suggest several related approaches to assessing the admissibility of the bridging assumption that do not rely on formal hypothesis testing.

We begin with a simple visualization that allows direct comparison of the structure of the media and legislator data. Figures 1, 2, and 3 present heat-map representations of the Groseclose and Milyo data with only the ordering of the rows an columns differing across the figures. In each figure, each row represents the data from a media outlet or from a grouping of legislators. Legislators are grouped by ADA score ranges. Each column represents the data for a think tank. To create the chart, the (aggregated) count data are converted into relative row frequencies (so that each row sums to one), then each column is z-scored. Darker colors reflect greater z scores.[2]

In Figure 1 the rows and columns are organized to emphasize the structure of the legislator data. In particular, we order the rows by the estimated ideology value when only legislator data

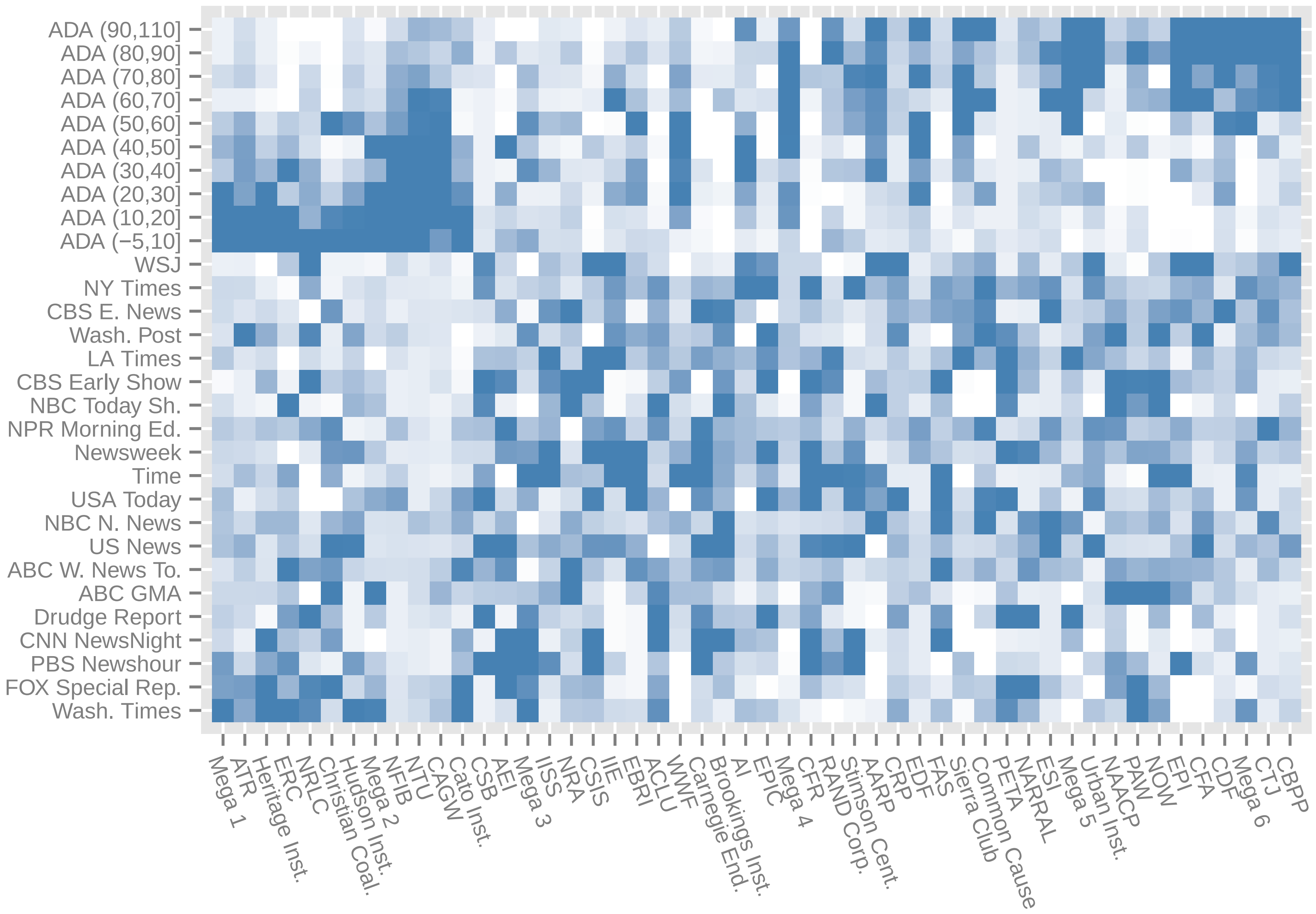
[2] The color ranges are truncated such that scores below -1 are white and scores above 1 are the darkest blue.

are used to estimate the mapping from ideology to think tank citations. The think tank columns are similarly ordered by the value of the slope (b) parameters in that estimation. The structure of the assumed model implies there should be a diagonal strip of blue from the bottom left to the top right of the plot within both the legislator data (placed at the top of the plot) and the media data (placed at the bottom of the plot). Figure 1 reveals a cluster of dark blue in the lower left and upper right corners of the legislator data, but a much weaker patter in the media data. While the lower left and upper right corners of the media data are darker than the upper left and lower right corners, the pattern is less pronounced. There is also a great deal of blue in the middle of the media map that is not found in the legislator map. While the less pronounced corners could reflect a set of media outlet ideologies with much narrower range than is found in the legislature, we also see that the concentration of blue in the middle of the media outlet data is not found for any range of ADA values in the legislator distribution.

Figure 2 reorganizes the rows and columns of the matrix to best highlight the structure of the media outlets' citations. Note that the shading of each cell has remained the same only the positions of the cells have been altered. Now we see a much stronger band of blue moving from the lower left to the upper right of the rows that present the media data and a much weaker, though still apparent, pattern in legislator data. Many of the rows and columns have been shift substantially. For example, People for the American Way (PAW), NARAL, PETA and NOW are more cited by conservative media outlets than by conservative legislators (ceteris paribus) as is revealed by the positions of their columns father to the left in Figure 2 than in Figure 1. Similarly IEI and the Stimson Center are much more commonly cited by liberal media outlets than by liberal legislators (ceteris paribus). We also observe substantial changes in the ordering of rows between Figure 1 and 2.

Under the assumption that the legislator and media data pool, we would expect these two figures to be very similar, but we see substantial differences. Beyond difference in the mapping from ideology to citations, we also note that there numerous think tanks that legislators are much more likely or much less likely to cite than is the media across the ideological spectrum. For example, the National Taxpayers Union (NTU) is more likely to be cited by legislators and the

Figure 1: Groseclose Milyo Data Organized to highlight the structure of legislators' citations



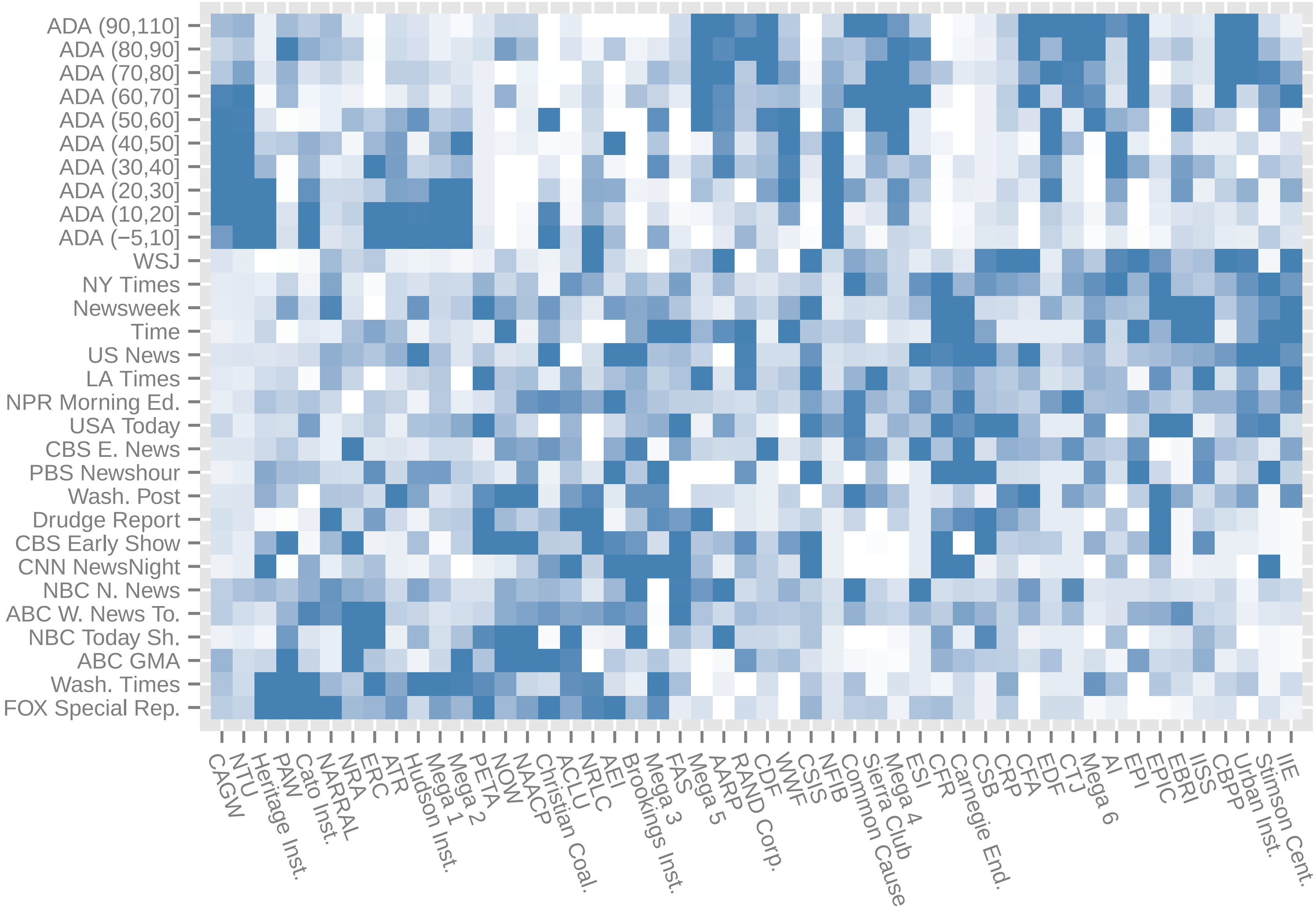
Note: Rows and columns are sorted to highlight the structure of the legislator citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

ACLU is much more likely to be cited by the media across the board. Taken together these plots raise doubt about the bridging assumption in this case.

Figure 3 presents the same map with the rows and columns organized to highlight the structure of the overall data under the pooling assumption (i.e., the columns and rows are sorted by the b and x estimates presented by Groseclose and Milyo). This figure reveals a compromise between the patterns presented in the first two. The degree to which the legislator or media observations dominate this compromise is a function of how sensitive the fit of the media and legislator data are to the model parameters and the relative amount of legislator and media data. Under the pooling assumption used to bridge the legislator and media ideology estimates, the relative amount of data collected for each group should have no systematic effect on the estimates, but if the data are governed by different parameter values, increasing the amount of legislator data will pull the estimates toward the legislator parameter values and increasing the amount of media data will pull the estimates towards the media parameter values.

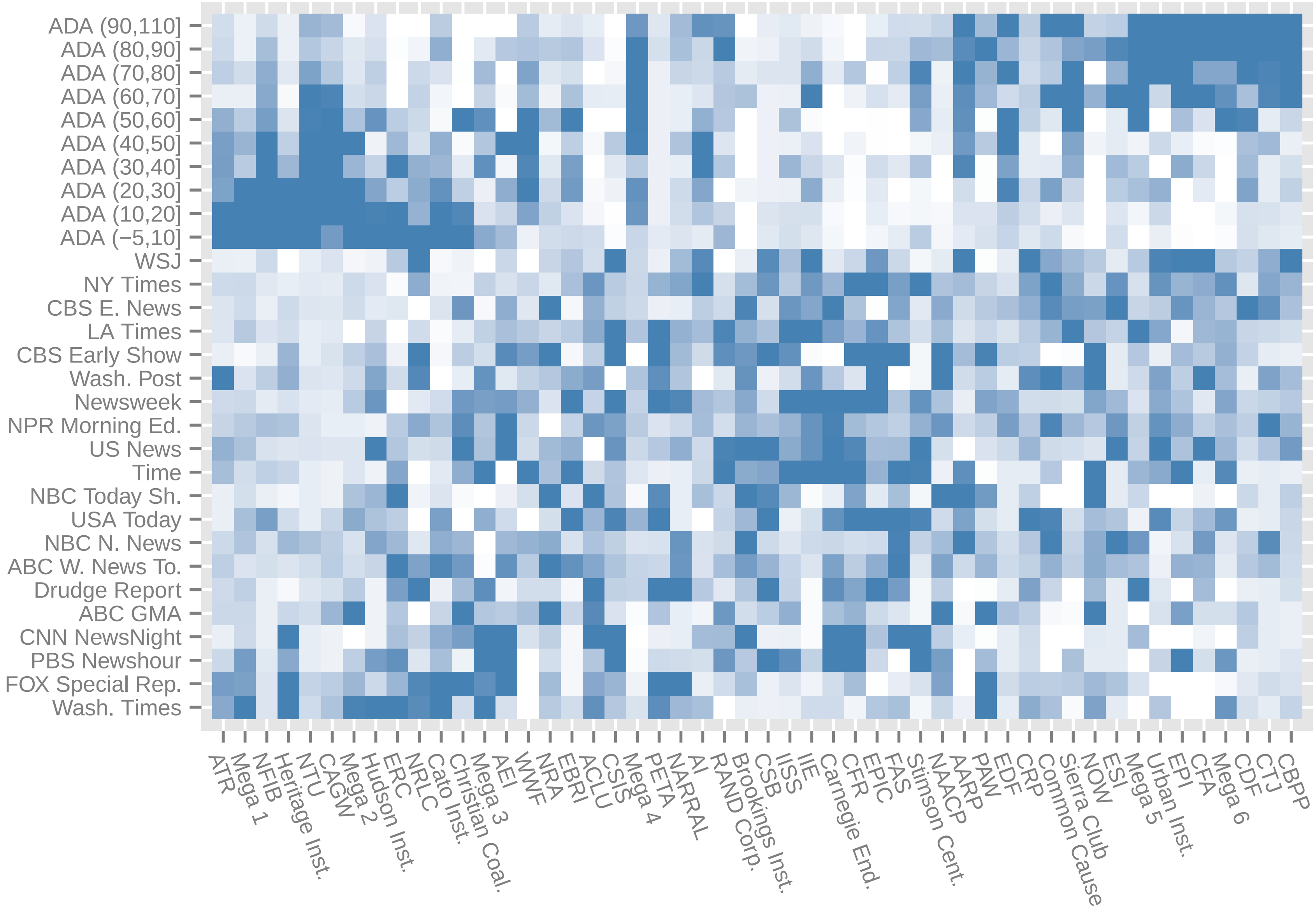
Figures4, 5 and 6, show similar data plots for the Gentzkow and Shapiro data. Here the full data matrix has 970 rows (including 536 legislators and 434 newspapers) and one thousand columns (partisan phrases). In the heat maps, the legislators rows are aggregated by similar district vote for Bush in 2004 and the columns are first sorted by their value of b and then collapsed into adjacent groups of 25 phrases. While the large number of phrases and media outlets make it difficult to assess how particular phrases and media outlets are organized across the three plots, it is very striking that the patterns of phrase use are quite different for the legislators and the newspapers. Figure 4 shows a very strong ideological sorting of the legislators use of the phrases, but little similar pattern is apparent in the newspaper rows of the figure. When the data are reorganized to maximize the amount of structure in the newspaper data in Figure 5 a stronger band of dark shading can be seen from the lower left to the upper right of the newspaper rows of the figure, but the apparent structure of the legislator rows has been largely lost. When the plot is organized to highlight the ideological structure across both the legislator and the newspaper data, the manifest structure is largely to separate the phrases into those used (relatively) more commonly by legislators and those used more commonly by newspapers. Here we see little to

Figure 2: Groseclose Milyo Data Organized to highlight the structure of media outlets' citations



Note: Rows and columns are sorted to highlight the structure of the media citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

Figure 3: Groseclose Milyo Data Organized to highlight structure under the pooling assumption



Note: Rows and columns are sorted to highlight the structure of the pooled set of citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

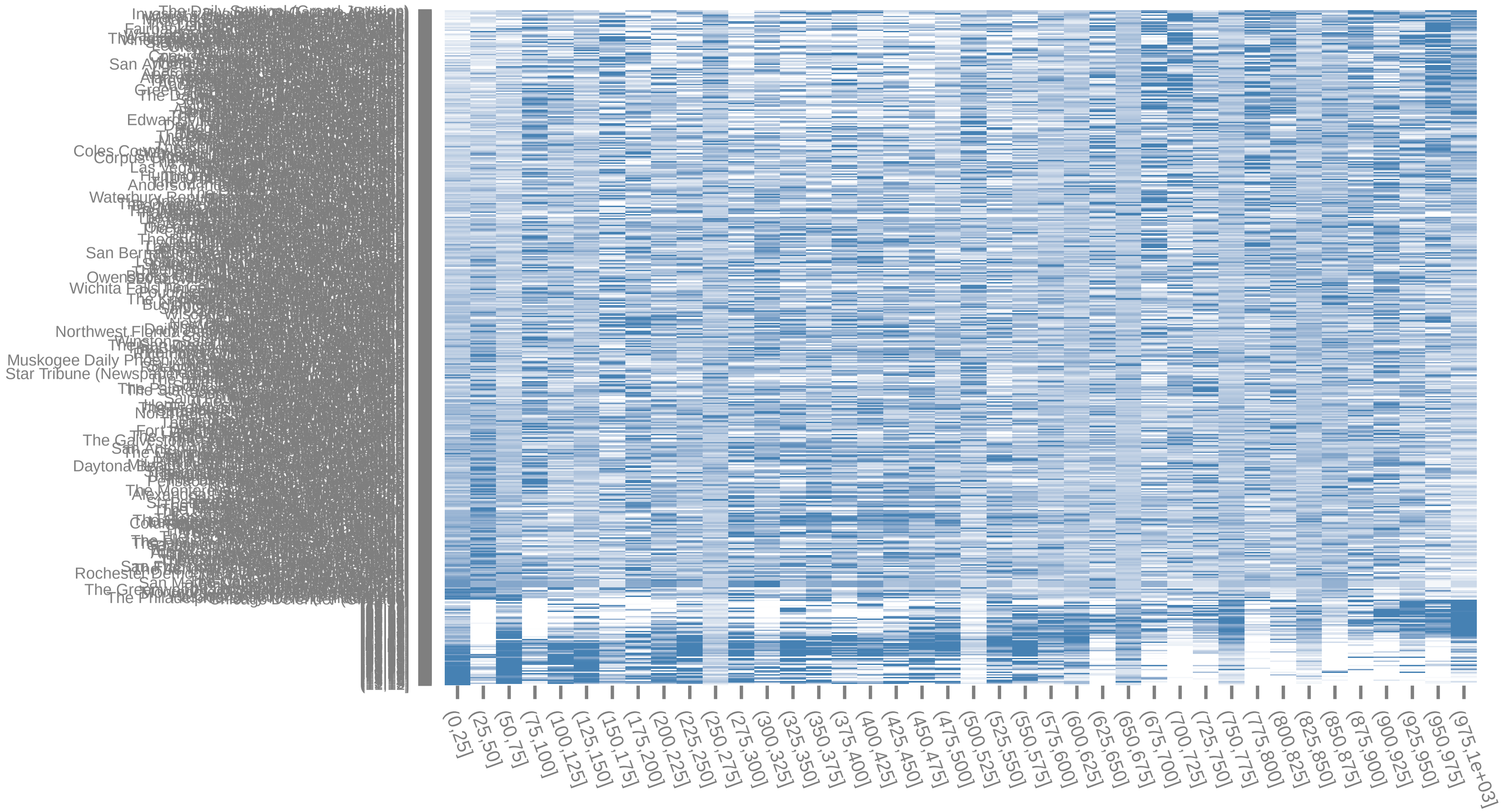
support the assumption that the legislator and newspaper responses are generated by the same model.

While looking at the heat maps provides a direct visual assessment of the apparent degree of pooling, the interpretation is some to some extent in the eye of the beholder. Showing that the optimal structure for each dataset is visually dissimilar is not quite the same as showing that they are incompatible with a particular model. In the next section we engage with the actual models used in each of these papers to recover estimates of slant. We show that pooling the data reduces the fit of the model for both groups.

As noted above, if the critical bridging assumptions hold then the estimates of the parameters that link ideology to the manifest indicators (citations or partisan phrases) should not systematically vary as a function of the balance of legislator and media data used to recover them. Similar results and fit should obtain if the as and bs are obtained exclusively from the legislator data (as in Gentzkow and Shapiro) or if the as and bs are informed by both the legislator and media data (as in Groseclose and Milyo). Similarly, the estimates should not systematically differ if we add more media outlets or legislators to the analysis. This suggests a simple way to probe the pooling assumption that can be assessed directly in terms of the parameter values that we seek to infer. We can consider counterfactually how the ML estimates of the model would vary as the balance of legislator and media data used in the estimation is manipulated. We do this by applying what amount to survey weights to the data. At one extreme, we downweight the media observations to the point that the as and bs are determined almost exclusively by the legislator observations (approximating the two-step estimation approach of Gentzkow and Shapiro). At the other extreme, we up weight the media observations so that the as and bs are almost exclusively determined by the media observations. We also fit intermediate values including unity which produces the unweighted ML estimates. We then plot key model quantities as a function of the degree to which the media data are up or downweighted. The more systematic variation is observed the less tenable is the pooling assumption.

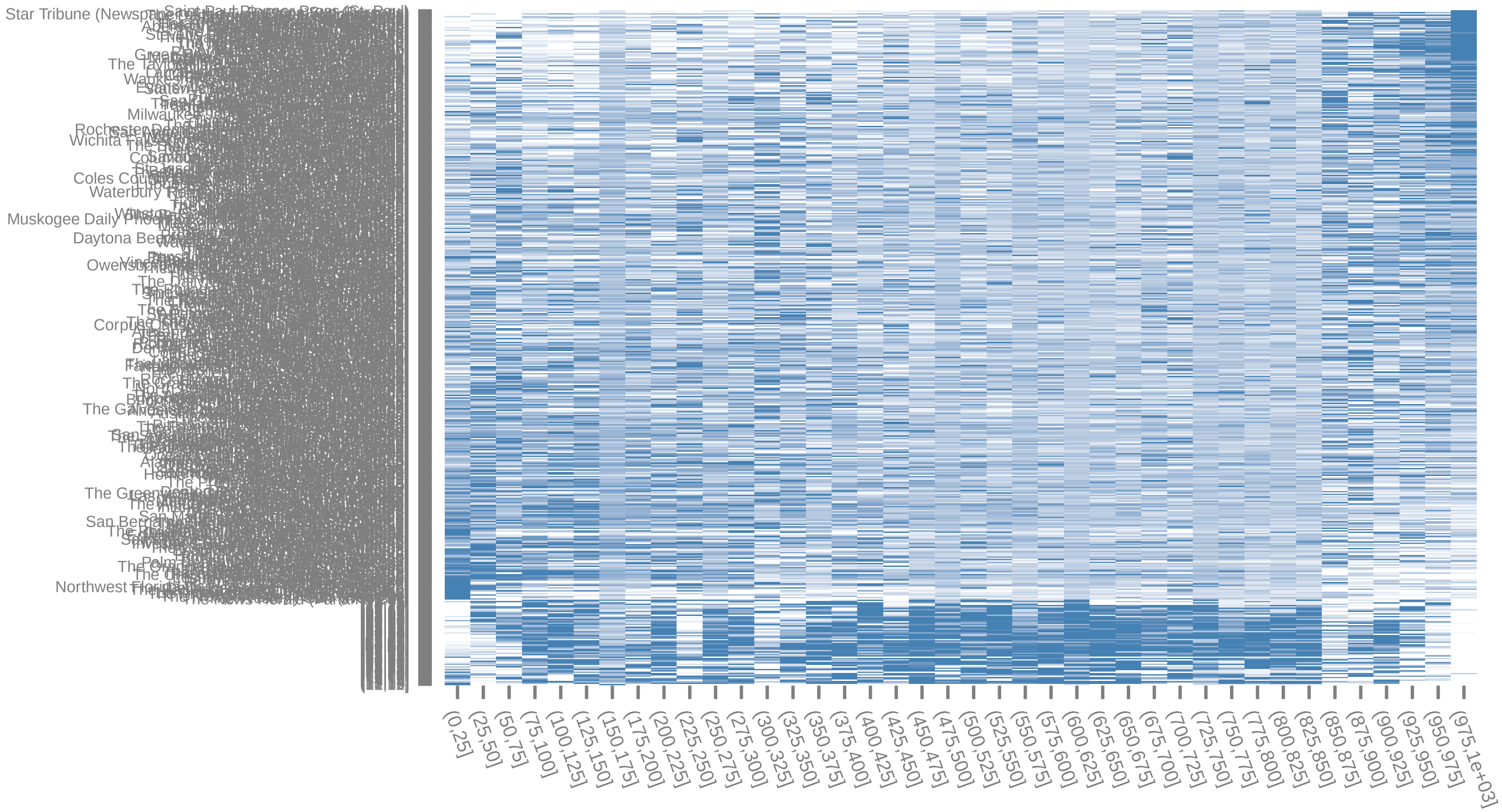
In Figure 7, we plot the **unweighted** log likelihood of the Groseclose and Milyo data as a function of the degree to which the media observations were up- or downweighted in the estimation.

Figure 4: Gentzkow Shapiro Data Organized By Legislator Citations



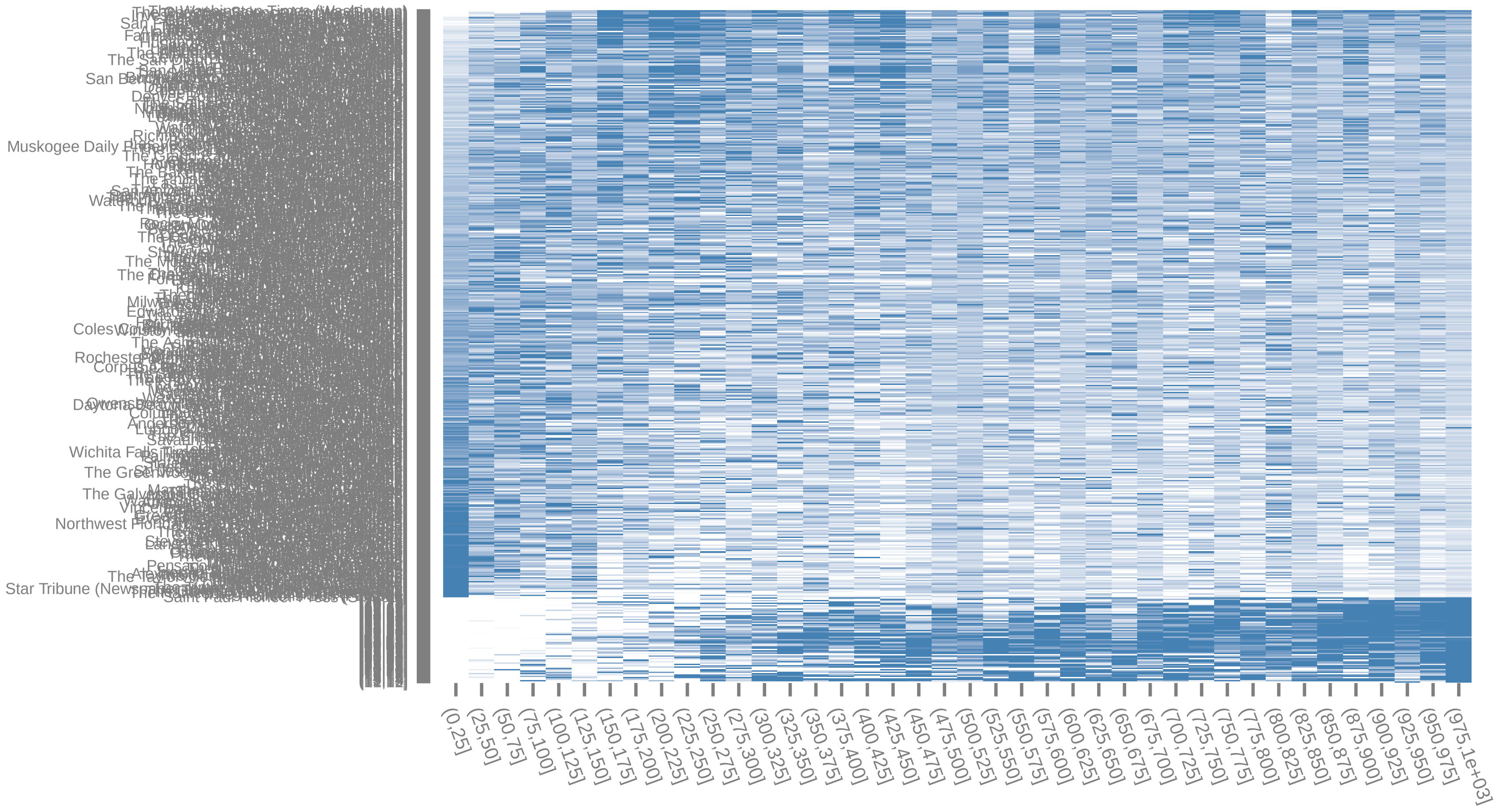
Note: Rows and columns are sorted to highlight the structure of the pooled set of citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

Figure 5: Gentzkow Shapiro Data Organized By Media Citations



Note: Rows and columns are sorted to highlight the structure of the pooled set of citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

Figure 6: Gentzkow Shapiro Data Organized By Both Legislator and Media Citations

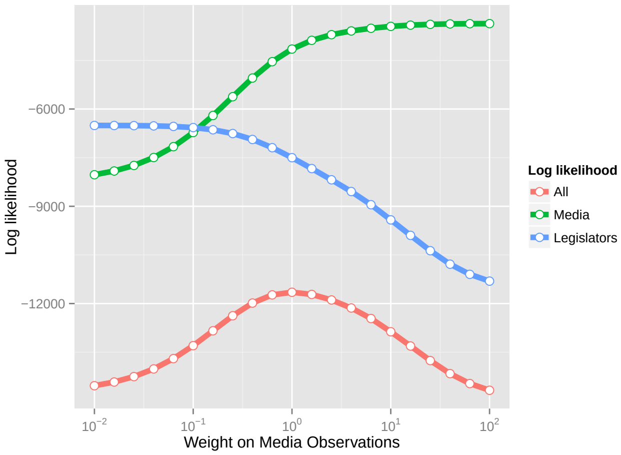


Note: Rows and columns are sorted to highlight the structure of the pooled set of citation patterns. Cells reflect reflect the relative frequency with which legislators and media outlets cite various think tanks. As described in the text, the original count are normalized by row so that each legislator and media outlets citations sum to 1. Each column is then z-scored. Darker shadings represent larger z-scores. The darkest shade represents z-scores above one. The lightest shade represents z-scores below -1.

In addition to showing the total log likelihood, we also show the log likelihood of the media and legislator data separately. The x-axis is the degree to which the media data were up- or downweighted on an exponential scale (so 10^0 is 1:1, 10^1 is 10:1, 10^2 is 100:1, 10^{-2} is 1:100 and so on). The y-axis shows the value of the unweighted likelihood. As the weight placed on the media increases, the likelihood of the fitted media observations increases and the likelihood of the legislator observations decrease by construction. Similarly, the maximum for the total likelihood must be achieved when equal weight is applied to the legislators and media because the maximum of the unweighted likelihood is precisely the objective when the media and legislator are given equal weight in the estimation. What is unexpected if the bridging assumption holds is the magnitude of the variation in the likelihood across the range of values. The media data is fitted dramatically better when the model is weighted towards the media data. If it were the case that the joint scaling assumptions were true, then setting parameter values that fit both groups would not constrain the media data from achieving close to its maximum. Giving more weight to the media data also has a severe negative effect on the likelihood of the legislator data. Similarly, we can consider the standard likelihood ratio test of the unconstrained model in which the media and legislator data are fit separately to the alternative in which the bridging assumption is enforced by summing the largest log likelihood achieved by the media observations (when media weight equals 100) and the largest log likelihood achieved by the legislator observations (when media weight equals 0.01) to the largest total log likelihood (when media weight equals 1), here we see a huge loss of fit associated with pooling the media and legislative observations.

This large change in likelihood suggests that the model would choose very different parameters if it were fitting the media data only rather than the media and legislators jointly, because different parameter values are leading to different likelihood values for the two groups. However, given that these are large datasets the substantive difference in the parameter estimates may be rather modest. In particular, we would not want to reject the opportunity to make the powerful substantive inferences that the pooling assumption allows us to make simply because it is unlikely that the models governing the legislators and media are not identical. Rather, we should consider how different the two data-generating processess are and how misleading the result of the pooling

Figure 7: Log likelihood by weight on media data in Groseclose-Milyo



Note:

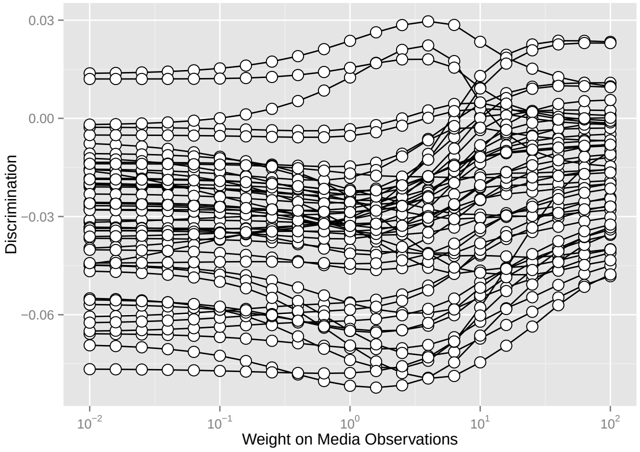
estimation might be. Because the relative amount of legislator and media data used is arbitrary and should not systematically alter the estimates, we should have little confidence in estimates that vary widely as function of relative weight given to the media observations.

Figure 8 shows the instability in the interest group discrimination parameters (the b_js) as the weight on the media observations is varied. If the joint scaling assumption were true, these discrimination parameters should be close to the same for both groups, and therefore the weighting should not affect the discrimination. Instead, the discrimination varies substantially. Some think tanks that were positively discriminating (wherein a citation indicates conservatism) become negatively discriminating (wherein a citation indicates liberalism), and some think tanks that were negatively discriminating become positively discriminating. Others parameters vary greatly in magnitude.

Just as the weight on each group affects the think tank parameters, it also affects the slant estimates for media outlets. Figure 9 shows the variation in slant parameters over the set of weights. Conclusions about the relative location of the media outlets as well as their dispersion depend greatly on whether we choose the structure indicated by media observations or by legislators. If great weight is put on the legislators, then the media appears to be more conservative on average, with relatively little dispersion in the space of ADA scores. As the weight on the media observations increases, the ordering changes as well as the dispersion and location of the mean.

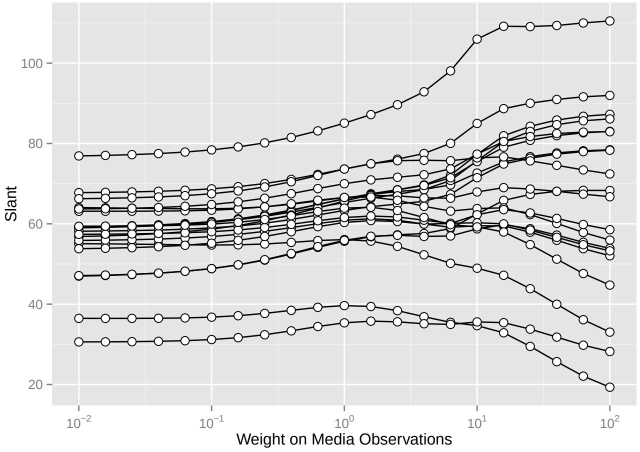
Figure 10 presents two different scenarios: one where the weight is 100:1 on legislators, and

Figure 8: Groseclose-Milyo: Discrimination parameters by weight on media data



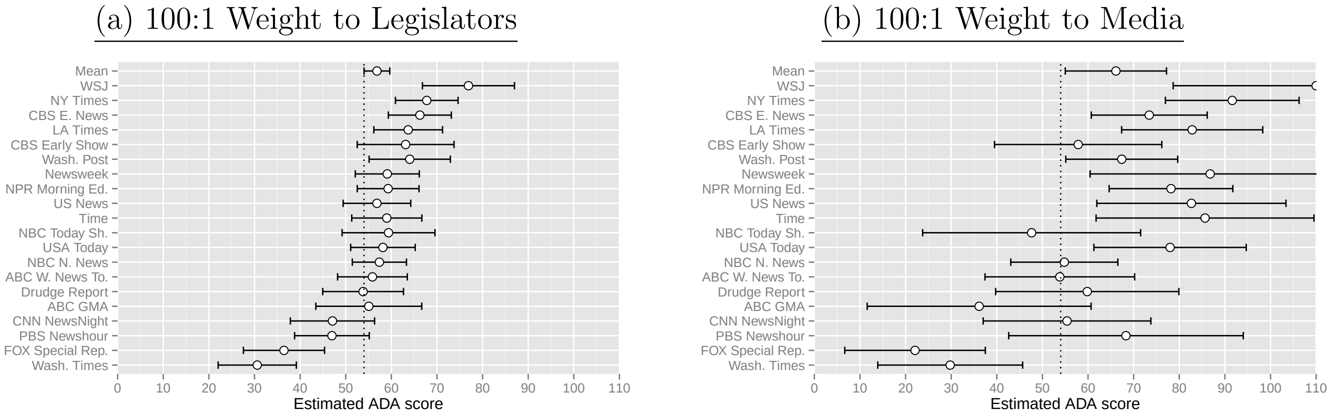
Note:

Figure 9: Groseclose-Milyo: Slant estimates by weight on media data



Note:

Figure 10: Media Ideal Points Under Two Different Weighting Assumptions (Groseclose Milyo)


(a) 100:1 Weight to Legislators

(b) 100:1 Weight to Media

Note: Each point represents the ideological score (on the Americans for Democratic Action scale) for each media outlet given the associated weighting assumption. The black lines are 80% credible intervals.

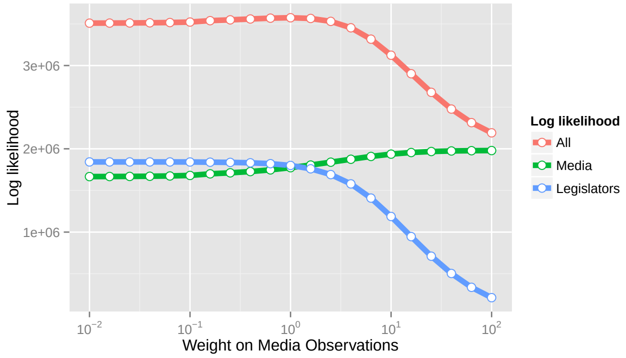
another where the weight is 100:1 on media outlets. The resulting ADA scores for media outlets, and their 80% confidence bounds, are shown. The conclusions that can be drawn about the relative position of the media outlets vary dramatically depending on which set of weights are chosen.

Figure 11 shows the effect on the likelihood for the Gentzkow and Shapiro data when the weights are varied. This graph shows the same pattern as Figure 7: when the weight on the media is increased, the model fits the media observation much better, but fits worse the legislators substantially worse, particularly as the weight on the media gets large.

Figure 12 shows how varying the weight affects the slant parameters. Due to the very odd pattern on the ride side of the range of weights, this graph is divided into two parts, where the right panel has a truncated x-axis. This graph helps explain why the likelihood for the legislators drops off more rapidly as the weight on the media observations is increased. In the right panel, we can see that the slant locations are changing as we go from weighting the legislators very heavily to only weighting them somewhat heavily. However, this trend is dwarfed by the what occurs when we the media observations get the lion's share of the weight. Suddenly, the location of the slant estimates are reflected over 0. This is because the model no longer attempts to fit the legislators: fitting the media better involves abandoning the constraints imposed by fitting the legislators.

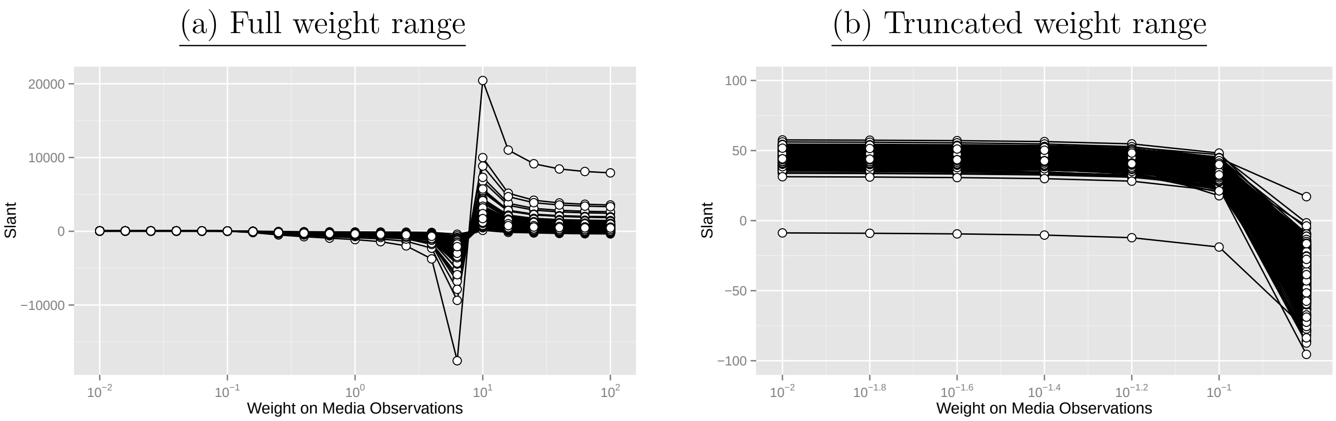
This is made clearer in Figure 13, showing the discrimination parameters, once again divided

Figure 11: Log likelihood by weight on media data in Gentzkow-Shapiro



Note:

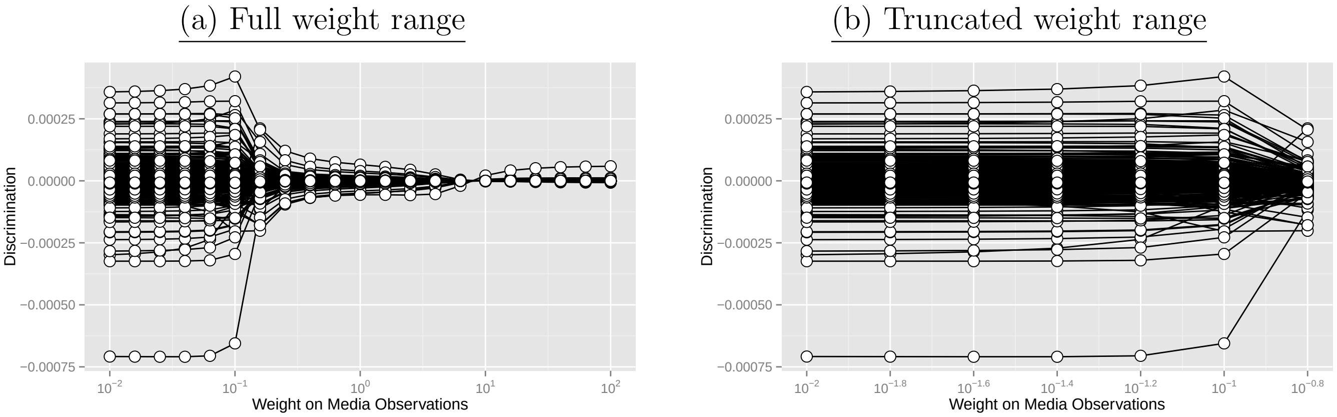
Figure 12: Gentzkow-Shapiro: Slant estimates by weight on media data


(a) Full weight range

(b) Truncated weight range

Note:

Figure 13: Gentzkow-Shapiro: Discrimination parameters by weight on media data


(a) Full weight range

(b) Truncated weight range

Note:

into two panels. As the weight on the media observations increases, the discrimination collapses towards 0.. The slant estimates become more dispersed, but the legislator estimates remain close to 0. The maximum likelihood estimates reflect a situation where legislators choose words almost randomly, and media outlets choose them systematically according to the model.

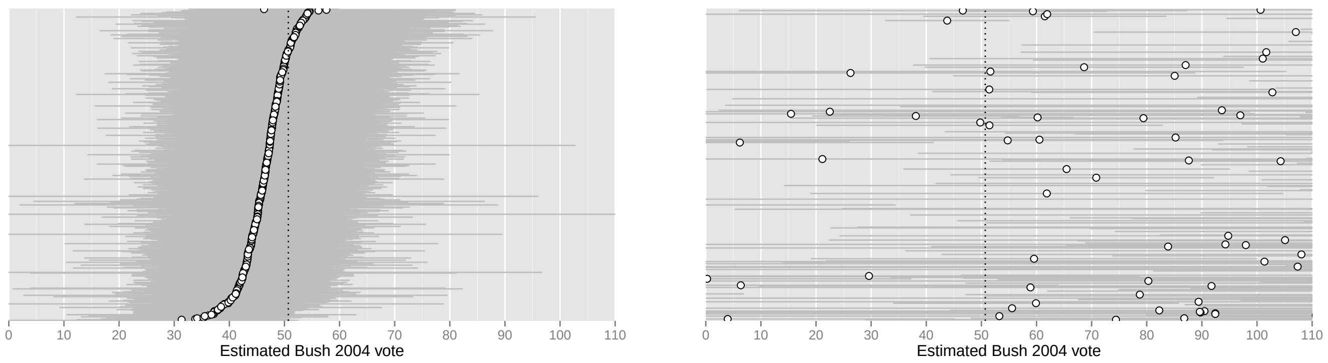
Figure 14 shows the affect of two different weighting regimes on the media idealpoints. As in Figure 12 above, the recovered locations are dramatically different—so much so that in the second panel the estimates are for the most part outside the range of the x-axis.

In the Appendix, we show that the log likelihood, discrimination parameters, and slant estimates are relatively constant if we weight by random subsets of the data rather than selecting the two groups we know to be dissimilar. This provides further evidence that the changes we observe when we vary the weights of each group are not the product of uncertainty or heterogeneity, but of specific differences between the structures of responses for these two groups.

5 Joint Scaling Legislators and Constituents

Assuming constant parameter values across groups is a strong assumption in many contexts. However, in the context of comparing legislators and constituents, this assumption seems particularly strong. Survey respondents make snap judgements in a low-information, low-stakes environments

Figure 14: Media Ideal Points Under Two Different Weighting Assumptions (Gentzkow Shapiro)



(a) 100:1 Weight to Legislators

(b) 100:1 Weight to Media

Note: Each point represents the ideological score (on the scale of Bush 2004 vote) for each media outlet given the associated weighting assumption. The grey lines are 80% credible intervals.

(Zaller, 1992). Legislators face a situation that is almost entirely the opposite. They are carefully and painstakingly informed by trained staff about the consequences of their choices, as well as being inundated by information from other legislators, outside groups, and the media. These choices often have important features that even careful outside observers could be forgiven for missing. A bill is not a simple representation of a policy view, but a collection of disparate policies and signals. A bill that appears to be on environmentalism may be *about* distribution of energy spending across districts; another bill that appears to be about emergency relief may be *about* a signal of support for a particular legislator.

Existing work that scales legislators and constituents jointly assumes that the response functions to particular survey questions and corresponding bills unders consideration in Congress are not just similar but exactly the same. As Jessee (2012) puts it: “respondents are treated as ‘guest senators,’ stopping in to vote on a small number of Senate votes.” Luckily it is possible to test the assumption that these response functions are the same, and to examine the sensitivity of the results to alternative assumptions. In particular, models of preferences for each group seperately should not drastically outperform models that restrict both groups to have the same response functions for particular items. Unlike the context of the media, it is not as simple to change the weight of each group, however it is simple to “unbridge” the groups entirely.

Much of the scaling literature that uses simple choice (rather than behavior) data uses an item response model akin to that familiar to political science from Clinton, Jackman and Rivers (2004)[3]. Responses to items (either survey questions, roll call votes, or codes for known political positions) are a function of each individual's latent political preferences. Let x_i denote the latent political preferences of person $i = 1, \ldots, N$, and y_{ij} denote person i's response to question $j = 1, \ldots, M$, where $y_{ij} = 1$ indicates a "yes" response and $y_{ij} = 0$ indicates a "no" response.[4] The probability of person i answering "yes" to question j is taken to be

$$\Pr(y_{ij} = 1) = \Phi(\beta_j x_i - \alpha_j)$$

where α_j and β_j are parameters, and Φ is the standard normal cumulative distribution function. In the educational testing literature, α_j is referred to as the "difficulty parameter" because a higher value of α indicates a lower probability of a "correct" answer (in our case, a yes answer), *ceteris paribus.* It is easier to think in terms of the "cut point" α_j/β_j, which is the value of x_i at which the probability of answering "yes" equals the probability of answering "no." β_j is referred to as the "discrimination" parameter because it captures the degree to which the latent trait affects the probability of a yes answer. If β is 0, then question j tells us nothing about an individual's preferences (x_i). We would expect β to be close to 0 if we ask question that which ideology or policy is irrelevant; for instance, a question about the respondent's favorite flavor of ice cream.

The complete log-likelihood is simply the sum of all of the individual log-likelihoods for each vote choice:

$$\ell(\theta; y, \mathcal{I}, \mathcal{J}) = \sum_{i \in I} \sum_{j \in J(i)} y_{ij} \ln(\Phi(\beta_j x_i - \alpha_j)) + (1 - y_{ij}) \ln(1 - \Phi(\beta_j x_i - \alpha_j))$$

where $\theta = (x_1, \ldots, x_N, \alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_M)$ is a vector of model parameters, $\mathcal{I}$ is the set of all people, $\mathcal{J}$ is the set of all items, and $\mathcal{J}(i)$ is the set of items responded to by the ith person. Following Jessee (2010) and Bafumi and Herron (2010), we assume that all non-responses are

[3] Poole and Rosenthal's (1997) W-NOMINATE and DW-NOMINATE are similar and also in wide use.

[4] Most of the questions used are dichotomous. Where they are not dichotomous, we use the rules given by the providers of the data in order to dichotomize them. Typically this involves choosing a sensible cut off point in an ordered question, and coding all prior items as "yes" and all later items as "no."

missing at random.[5] In this sense, we treat items upon which an individual did not have an opportunity to respond and items upon which they abstained or answered "Don't Know" similarly. Although the parameters are identified relative to one another, they lack a scale. We establish an arbitrary scale by normalizing the xs to have mean zero and standard deviation one.

Item response models allow us to estimate comparable measures of latent traits for each person because we assume homogeneous response functions.[6] That is, we assume that α_j and β_j do not vary by respondent (i) for any $j \in \mathcal{J}$. The systematic differences in patterns of responses across individuals are only due to differences in their political preferences, (xs). For many applications, the assumption of homogeneous response may seem innocuous. For instance, there is an extensive literature which shows that a one-dimensional item response model that assumes homogeneous response can account for much of the roll call voting in the US Congress. Joint scaling applications call this assumption in to question.

To characterize the identification challenge presented by joint scaling, we consider two groups of individuals: survey respondents (r) and senators (s). Let $\mathcal{I}_k$ and $\mathcal{J}_k$ be the sets of individuals and response items for groups $k \in \{r, s\}$.[7] Now we can rewrite the log-likelihood above as

$$\ell(\theta; y, \mathcal{I}, \mathcal{J}) = \ell(\theta; y, \mathcal{I}_r, \mathcal{J}_r) + \ell(\theta; y, \mathcal{I}_s, \mathcal{J}_s).$$

That is, the total log-likelihood is the sum of the contributions made by the respondents' responses and the senators' responses.

Figure 5 shows examples of four types of $N \times M$ response matrices, $[y_{ij}]_{ij}$. Each matrix has a column for each item in $\mathcal{J}$ and a row for the each individual in $\mathcal{I}$. The matrices are organized so that all members of the first group (say respondents) are listed before all members of the second group (say senators). Individuals belonging to both groups (if any) are listed in between those who are only senators and those that are only respondents. Similarly, all of the items responded to only by the second group are listed before items answered only by the first group. Items answered

[5] Almost all of the work we are aware of makes this assumption, but for some counterexamples see Poole and Rosenthal (1997), Lo (2013), and Powell (2015)

[6] While, as discussed below, homogeneous response is sufficient to establish scale comparability it is not strictly necessary see Carroll et al. (2013).

[7] An item j is a member of $\mathcal{J}_k$ if $j \in \mathcal{J}(i)$ for at least one $i \in \mathcal{I}_k$. That is, $\mathcal{J}_k = \cup_{i \in \mathcal{I}_k} \mathcal{J}(i)$.

by members of both groups are placed in between those answered by only one of the two groups. This arrangement reveals four basic types of joint data matrices: ones in which there is no overlap between the individuals or the items (type I); ones in which their are common items, but no overlap in group membership (type II): ones in which there is common group membership, but no common items (type III); and, ones in which there is both common group membership and items across the two groups (type IV).

We begin with a consideration of data matrices of Type I. Here items and individuals are distinct across groups ($\mathcal{I}_s \cap \mathcal{I}_r = \emptyset$ and $\mathcal{J}_s \cap \mathcal{J}_r = \emptyset$) and it is immediately clear that there is nothing "connecting" the respondent and senator estimation problems. The parameters associated with respondents only depend on the responses of respondents and the parameters associated with senators only depend upon the responses of senators. Consequently, the total likelihood can be maximized by maximizing the senator likelihood and the respondent likelihood separately. Similarly, the posterior distribution over each groups' ideal points are only functions of the responses observed within each group.

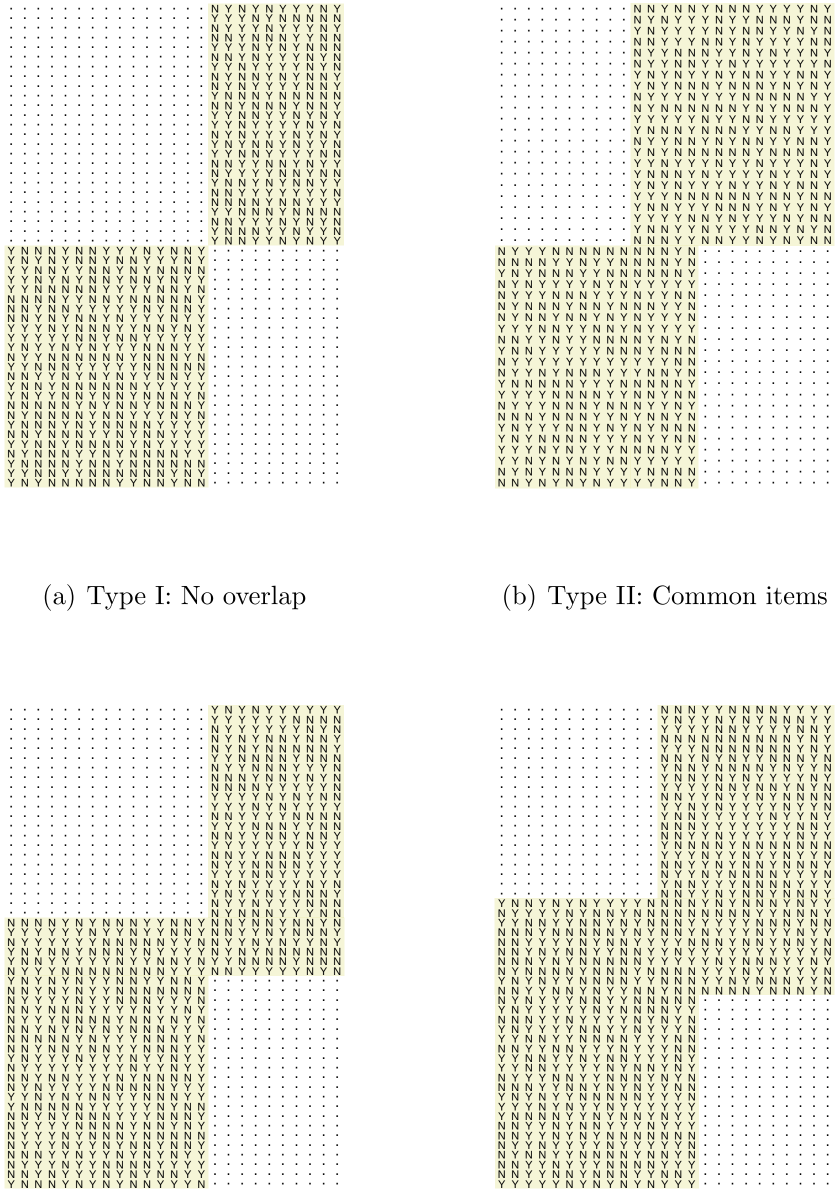
Because there is nothing linking the senator and respondent problems, the resulting ideology or policy scales are not comparable across groups. To see this, suppose that senators ideal points are placed on an arbitrary scale with ideal points x_i^* for all $i \in \mathcal{I}_s$. Suppose that those ideal points can be transformed back to a common scale as

$$x_i = \delta_1 + \delta_2 x_i^*$$

where δ_1 and δ_2 are fixed constants and $\delta_2 \neq 0$. Substituting x^* for x, the log-likelihood for senators' responses can be written as

$$\ell(\theta^*; y, \mathcal{I}_s, \mathcal{J}_s) = \sum_{i \in I_s} \sum_{j \in J_s(i)} y_{ij}(\Phi(\beta_j^* x_i^* - \alpha_j^*)) + (1 - y_{ij})(1 - \Phi(\beta_j^* x_i^* - \alpha_j^*))$$

where $\alpha_j^* = \alpha_j - \delta_1 \beta_j$ and $\beta_j^* = \delta_2 \beta_j$. Thus, the same likelihood values can be achieved for the senators' responses under any choice of scale without knowledge of, or possibility of learning, the values of the transformation parameters δ_1 and δ_2. This is, of course, nothing more than the



(a) Type I: No overlap

(b) Type II: Common items

(c) Type III: Common people

(d) Type IV: Common people and items

Figure 15: Possible joint-scaling data matrices. *Each panel shows a matrix of responses with columns equal to the number of items and rows equal to the total number of individuals across the two groups (senators and respondents). Dots in the matrices represent missing values. In panel (a) senators and respondents are disjoint as are the items on which each of the two groups vote. In panel (b), there are some individuals who are both senators and respondents. In panel (c), there are some items in common between senators and respondents. And, in panel (d), there are is overlap in items and membership between the two groups.*

usual problem of establishing the scale of any latent measurement. However, in the case of joint scaling with no overlap in items or individuals, any choice of scale for the senators can be fully accommodated by offsetting shifts in the item parameters that have no effect on the likelihood of the respondents' responses (because the two groups respond to different items). Because, the transformation of the senators' scale has no effect on the likelihood of the respondent data, the choice of senators' scale can be made independently of the choice of the respondents' scale. Thus, while senators can be located relative to one another and respondents can be located relative to one another, we cannot identify where senators are located relative to respondents if the sets of respondents and senators are disjoint and the set of items answered by respondents and senators are disjoint.

In order to place different groups on the same scale, there must be overlap either in the items that each group responds to or in group membership (data matrices of types II, III, or IV). That is, the data matrices containing responses of the two groups must be "glued" or "stitched" together through common rows and/or common columns.

In the context of an IRT model, overlap in group membership means not simply that the same individual i was a member of both groups ($i \in \mathcal{I}_r$ and $i \in \mathcal{I}_s$), but that she expressed the same preference (x_i) when a member of each group. Similarly, for two items to be common to each of two groups, it is not sufficient (or even necessary) that the same question was posed to each group, rather it is the manner in which members of each group translate their underlying ideology into a response to that item that must be the same (that is, a common item's α and β must be the same for members of each group).

These requirements are restrictive and are often an impediment to answering central research questions. For example, in order to establish comparable estimates of legislator ideologies across time, researchers often assume that the ideal points of individual legislators do not change over time (creating data matrices with common rows). However, if we are interested in studying how members of Congress might alter their ideological positions following a midterm landslide, we cannot identify the magnitude of each member's ideological shift without first assuming that some specified members' positions remained unchanged after the landslide.

In studying representation and it is, of course, unlikely that we could ever identify survey respondents who are also elected representatives. Thus, in order to stitch together respondent and senator data matrices, we require common items. Bafumi and Herron (2010) and Jessee (2010) work hard to construct survey items that probe how the public would vote on particular roll call votes.[8] However, one might wish to ask how and if voters translate their underlying ideological dispositions into opinions about specific votes and, in particular, if they do this a way that differs from how legislators make the same translation. For example, opinions offered in a survey might be poor substitutes for the counter-factual in which a respondent actually becomes a senator who has to cast roll call votes on the same questions. However, such a comparison requires that the commonly-posed items are allowed to have response parameters that differ between respondents and senators ripping loose our identifying stitches. Once again, we have to assume (at least some part of) the answer that we are seeking in order to pose the question.

As will be shown shortly, while we cannot identify a common scale for respondents and senators without at least as single common item, if we have more than one (possibly) common item, we can test hypotheses and form posterior-beliefs about the whether those (over)identifying assumptions are correct. While this approach can also be taken when the data matrix is of type II, III, of IV, we will focus here on the case of (possibly) common items (data of type II).

Continuing the notation developed above, we can now partition the set of items, $\mathcal{J}$ into three subsets: senate-only items, $\mathcal{J}_s$; respondent-only items, $\mathcal{J}_r$; and items that "bridge" both senators and respondents, $\mathcal{J}_b$. The log-likelihood can then be broken into three parts,

$$\ell(\theta; y, \mathcal{I}, \mathcal{J}) = \ell(\theta; y, \mathcal{I}_r, \mathcal{J}_r) + \ell(\theta; y, \mathcal{I}_s, \mathcal{J}_s) + \ell(\theta; y, \mathcal{I}, \mathcal{J}_b).$$

Because of the existence of the "bridging" items, we can no longer arbitrarily alter, for example, the scale of the senators ideal points, without affecting the likelihood. The log-likelihood of the

[8] [Bafumi and Herron also stitch these population together with the President by using public position taking data, and they stitch together the two chambers of Congress by using common votes on conference reports.

bridged items can be written as

$$\begin{array}{rcl}\ell(\theta^*, \theta; y, \mathcal{I}, \mathcal{J}_b) & = & \sum_{i \in I_r} \sum_{j \in J_r(i)} y_{ij}(\Phi(\beta_j x_i - \alpha_j)) + (1 - y_{ij})(1 - \Phi(\beta_j x_i - \alpha_j))+ \\ & & \sum_{i \in I_s} \sum_{j \in J_s(i)} y_{ij}(\Phi(\beta_j(\delta_1 + \delta_2 x_i^*) - \alpha_j) + (1 - y_{ij})(1 - \Phi(\beta_j(\delta_1 + \delta_2 x_i^*) - \alpha_j))\end{array}.$$

Now if we arbitrarily set the scale of the senate ideal points, we can estimate the δ_1 and δ_2 required to place respondents even if we only have a single bridged item. In the case of the single bridging item, δ_1 and δ_2 are exactly identified. That is, the respondent and senator estimation problems could be solved separately with each group's members being located on their own arbitrary scale (for example, both senators' x^*'s and respondents' x's could be normalized to have mean zero and standard deviation one). The transformation parameters δ_1 and δ_2 could then be recovered by noting that the item parameters for the common item on the scale used for the senators are by definition, $\alpha_b^* = \alpha_b - \delta_1 \beta_b$ and $\beta_b^* = \delta_2 \beta_b$ where α_a and β_b are the values of those parameters on the scale used for the respondents. After rearranging, we see that $\delta_1 = \frac{\alpha_b^* - \alpha_b}{\beta_b}$ and $\delta_2 = \frac{\beta^b}{\beta_b}$. Thus, the likelihood of the responses for senators and the responses for respondents can be calculated separately and then the values of δ_1 and δ_2 needed to place senators and respondents on the same scale can be found using these formulas. Note that the maximum log likelihood of the joint estimation is simply the sum the maximum log likelihood over the senators' choices and the maximum log-likelihood over the respondents' choices.

Because, we can always choose a δ_1 and a δ_2 that will equate any α_b and α_b^* and any β_b and β_b^*, we are not able to test the assumption that senators and respondents react in the same way to the common item: We can always transform the senators (or the respondents) scale such that their item parameters are on the common item are equivalent to those of the respondents (or senators) with no effect on the fit of the model.

If instead of a single bridging item, we have several bridging items, the model becomes overidentified. For every $b \in \mathcal{J}_b$, the ratio of β_b^* to β_b and $\alpha^* - \alpha_b$ to β_b must be constant (and equal to δ_1 and δ_2 respectively). Now we can no longer calculate maximum likelihood estimators or posterior distributions for the senators' and respondents' data sets separately. The restrictions on the common item parameters cause the senator data to affect all of the respondent parameters and *vice versa*.

Because assuming that the common items function in the same way for respondents and senators places constraints on the log-likelihood function, we can in the usual way compare the fit of a model that relaxes these constraints to one that imposes them. If the common items function similarly across the two groups, imposing the constraints will have little effect on model fit. If, however, the parameters associated with the common items differ substantially between the two groups then the model that allows for different common item parameters will fit the data substantially better and we will be able to reject the assumption that the common items can be used to identify the senators and the respondents on a common scale.

Each of our models is estimated using common Bayesian Markov Chain Monte Carlo methods, as implemented in the software package JAGS (Clinton, Jackman and Rivers, 2004; Plummer et al., 2003).

6 Legislator and Constituent Scaling Data

We use two datasets as cases to evaluate the joint scaling of survey respondents and legislators. The first comes from Jessee (2009). Jessee's paper is the first prominent paper to apply joint scaling to legislators and voters. Our second dataset is from Bafumi and Herron (2010). Bafumi and Herron's paper is the most widely read example to date of the use of joint scaling to examine the relative positions of voters and their representatives in Congress.

These data sets link responses to particular survey questions to roll call votes or other positions taken by elected officials. Jessee's data has a very simple structure. It consists of responses to a single national survey and roll call votes taken by Senators in a single session. Respondents answer a subset of 27 questions, each of which was meant to simulate a roll call vote taken in the Senate. The respondents were presented with a description of an actual bill voted on in the Senate, and asked how they would vote on that bill. For instance, here is the description that was used for a bill to require child safety locks on guns:

> S AMDT 1626 to S 397: Child Safety Locks Amendment

- Requires gun manufacturers and sellers to include child safety locks on all guns sold or transferred.

For more complex bills, the question might provide three or four bullet points of description. The responses to these roll call questions are the only survey responses included in the data set. Respondents received a random subset of the 27 questions that were asked. The average respondents answered "yes" or "no" to 11 questions. The legislators data includes 582 roll calls, of which the average legislator responded to 506, including 23 of the 27 roll calls that respondents were asked about.

In the resulting matrix of responses, legislator votes on roll calls and the survey respondent answers to the corresponding roll call questions are included in the same column. A "yay" vote is given the same code as a "yes" to the question and a "nay" vote is given the same code as a "no." The 582 roll calls that the respondents were not asked about are treated as missing for the respondents. Jessee's data resembles the type II example from Figure 5. The only difference is that there are no items on which respondents take a position but legislators do not have a corresponding roll call vote.

The data from Bafumi and Herron (2010) has a more complex structure. It includes members of the House and the Senate over two sessions of Congress (the 109th and the 110th), the President, and respondents to three different public opinion surveys (although all had a shared component). Most of the items used to link the respondents to members of Congress were asked on one of the three public opinion surveys, the Dartmouth module of the Cooperative Congressional Election Study. The questions that link the public opinion surveys to legislative roll call votes are primarily linked to votes taken in either the House or the Senate. In one case, an item is linked directly to the assumed position of the President. The public opinions surveys were linked to each other by common questions, and the House was linked to the Senate by conference committee votes, which are identical in both chambers (at least in subject matter). The political preferences of four members who graduated from the House to the Senate are assumed fixed, further linking the two chambers.

This "spiderweb" or "swiss cheese" structure links groups to each other by many different avenues. For instance, if the common item assumptions were justified then the fact that survey respondents take positions on both House and Senate roll calls would ensure that the estimated positions of Senators and Members of the House were comparable in the same policy space. It would be difficult to examine all of these relationships, so we focus on the assumptions underlying the questions which link survey respondents to legislators. We will take as granted that all other assumptions about common item or person parameters are correct. In other words, even though Bafumi and Herron's data has the structure that appears as type IV in Figure 5, we treat it as type II, ignoring links other than the items that link respondents to legislators.

Bafumi and Herron's data includes 8219 survey respondents and 629 elected officials. In order to save computing time, we keep 1100 roll call votes (including all of the ones that are linked to survey respondents) and discard the rest. There are 64 unique questions asked to survey respondents, although the average respondent only answers 33. There are 17 questions that are linked to legislator responses, of which the average survey respondent answers 8. These responses are concentrated among the respondents to the Dartmouth survey.

It is important to look at both of these examples, because any joint scaling exercise will be affected by the choices of that particular researcher. If questions are poorly phrased, or roll calls are wrongly portrayed, this will affect the results, as it should. By using both of these data sets, we examine two independent sets of assumptions about which choices by respondents are appropriate to equate with votes cast or positions taken by elected officials.

7 Legislator and Constituent Scaling Results

For each dataset, we estimated two different models with different assumptions. In the first "joint" model, we preserve the common-item structure of the data and estimate the model on the type II matrix, as in Figure 5. In the second "not joint" model, we sever the connection between the two groups, estimating them seperately. This is equivalent to the type I matrix in Figure 5. Table 1 shows the fit statistics for the constrained and unconstrained models estimated using the data from Jessee (2009). For the "not joint" or unbridged model, we can calculate our fit statistics by summing across two seperate analyses. The total number of responses is the same as in the "joint" or bridged case, and so we can compare these likelihoods even though one is the sum of two essentially seperate analyses.

Table 1 contains four different measures of fit. Two of these are widely-used statistics for model selection: the Deviance Information Criterion (DIC) and the Bayesian Information Criterion (BIC). Models with lower BIC and DIC are preferred. The Deviance Information Criterion is readily calculated based on our Bayesian estimation. To calculate the Bayesian Information Criterion, we require the values of the parameters that maximize the likelihood function. We use a simple expectation-maximization algorithm to calculate this quantity. The other two are measures of out of sample predictive accuracy. We withold 10% of the data at random from all of our analyses and classify the percent of these outcomes that are correctly predicted in each case. In particular, we focus on the percent of the bridged items that are correctly predicted for voters and legislators respectively.

For three out of our four statistics, we focus only on the "bridging" items, even in the "unbridged" case where there are twice as many items in this set, but the same number of responses. The reason for this is simple. Due to the arbitrary scale of the items and estimated preferences, both the bridged and unbridged models may estimate similar relative locations for the non-bridged items. In fact, in the case where bridging is entirely untenable, the bridging items will simply have 0 discrimination, and the model will fit each group as if they were estimated seperately. Alternatively, the bridging items may fit well for one group but the other group may be scaled such that discrimination is very low. This is analogous to what we observed in the case of the media scaling for the Gentzkow and Shapiro (2010) data.

Table 1 shows exactly this. First, both the DIC and BIC are substantially higher for the bridged model than the unbridged model. We should keep in mind that for a less constrained model, in this case the unbridged version, the likelihood will always be higher. However, both the DIC and BIC contain penalty terms to account for the large number of additional parameters that are used in the unbridged model. A difference in the BIC of 10 is considered "very strong" evidence in favor of the superior model (Kass and Raftery, 1995). In this case, the unbridged model is preferred by a difference of 502.

We can see from this table where the difference in fit comes from. For voters, the difference in model hardly changes the fit, which is just above 66% in both cases. However, for legislators, the fit degrades substantially. Recall that classifying all choices as the more popular response will classify choices at a rate greater than 50%. Joint scaling reduces the out-of-sample accuracy for legislators by 13 percentage points. This is at least a 46% reduction in the portion of the fit explained by the model, rather than a naive "always yes" or "always no" classification. The reason for this reduction is that choosing item parameters for these items that fit voters is not compatible with choosing parameters for these items that fit legislators. The structure of decision making on these items is fundamentally different in these two groups.

	DIC	BIC[9]	CV:voters	CV:leg
Bridged	84,540	133,305	66.3%	73.1%
Unbridged	83,062	132,803	66.5%	86.1%
Difference	1,478	502	-0.2%	-13.0%

Table 1: Fit Statistics for Data from Jessee (2009)

We replicate all of the above analyses using the data from Bafumi and Herron. As explained above, the "unbridged" model maintains the constant item parameter assumption for all items except the ones that are shared by legislators and survey respondents. Table 2 shows the results when we calculate the fit statistics for the Bafumi and Herron data. Once again, both the DIC and BIC strongly prefer the unbridged model. This time the difference in the BIC is 1,538. Once again, this difference comes from a severly reduced ability to predict legislator choices on the bridged items. The accuracy is reduced by 9.3 percentage points.

	DIC	BIC[10]	CV:voters	CV:leg
Bridged	286,656	193,154	75.2%	73.3%
Unbridged	284,629	191,616	75.5%	82.6%
Difference	2,028	1,538	-0.3%	-9.3%

Table 2: Fit Statistics for Data from Bafumi and Herron (2010)

In both of these cases, bridging legislators and constituents degrades the fit according to conventional model selection statistics. Further, it reduces the classification for legislators on the bridged items substantially. The exercise of comparing the location and spread of the ideal points of legislators to those of survey respondents is now in doubt. However, suppose that despite the large difference in likelihood between the joint and not joint models, we still wish to procede with our joint scaling. How certain should we be about the relative position and spread of the two groups?

We approach this question in two ways. First, we re-bridge the items in a manner that is clearly incorrect. For simplicity, we will focus on the Jessee (2009) data. Rather than assuming that the item parameters are constant for each of the "bridged" items provided, we instead connect the voter responses to arbitrary legislative votes, chosen at random. We assume that *these* item parameters are the same for legislators and citizens, even though we know that they are in no way related except by random chance. Table 3 shows the DIC that results (in this context it makes no sens to compute statistics for the bridging items only, because they differ). Although the DIC prefers Jessee's bridging scheme to our completely random one, the difference is less than half of the difference between the bridged and unbridged model. In other words, if the difference in DIC between the unbridged and bridged models in Table 1 is not enough to support the unbridged model over the bridged model, then it is not enough to support the bridged model over one where the bridges are made completely at random.

	DIC
Randomly Bridged	85,273
Bridged	84,540
Difference	733

Table 3: Comparison of random bridging scheme to the one in Jessee (2009)

We demonstrate the extent of this problem in another way as well. We compare the bridged model to one where the estimates are further constrained — in this case, constrained to a different location. Recall our notation for the shift and spread of one group relative to the other, δ_1 and δ_2, respectively. We can alter these quantities, and hence the location of the voters relative to the legislators. Then we can re-estimate optimal locations for the items, and compare the effect that this has on the likelihood.

Using a grid of values for δ_1 and δ_2, we re-estimate a restricted version of our model using an Expectation Maximization (EM) algorithm to find the optimal item parameters for a given δ_1 and δ_2. Starting from the parameters of the joint model, we apply the given adjustments and re-estimate the parameters of the items to achieve the best possible fit. Each cell in table 4 corresponds to an increase in the BIC relative to the bridged model resulting from a given change to δ_1, δ_2 or both. The rows correspond to changes in δ_1 (shift) and the columns correspond to changes in δ_2 (stretch). When $\delta_1 = 0$ the parameters are not shifted, and when $\delta_2 = 1$ they are not stretched, so the top left cell corresponds to 0 change in BIC. As these parameters change, the BIC increases. Interestingly, the more that citizens are shifted, the less harm is done by stretching them, and in fact this can reduce the degradation in the fit by increasing the overlap between the two groups and the items that are meant to fit them both.

The main takeaway from Table 4 is that even substantial shifts and stretches of relative position do less harm to the BIC than the increase of 502 that is achieved by going from the unbridged model to the bridged model. If this were not sufficient to reject the bridged model, then we also cannot reject large moves in the relative positions of the two groups.

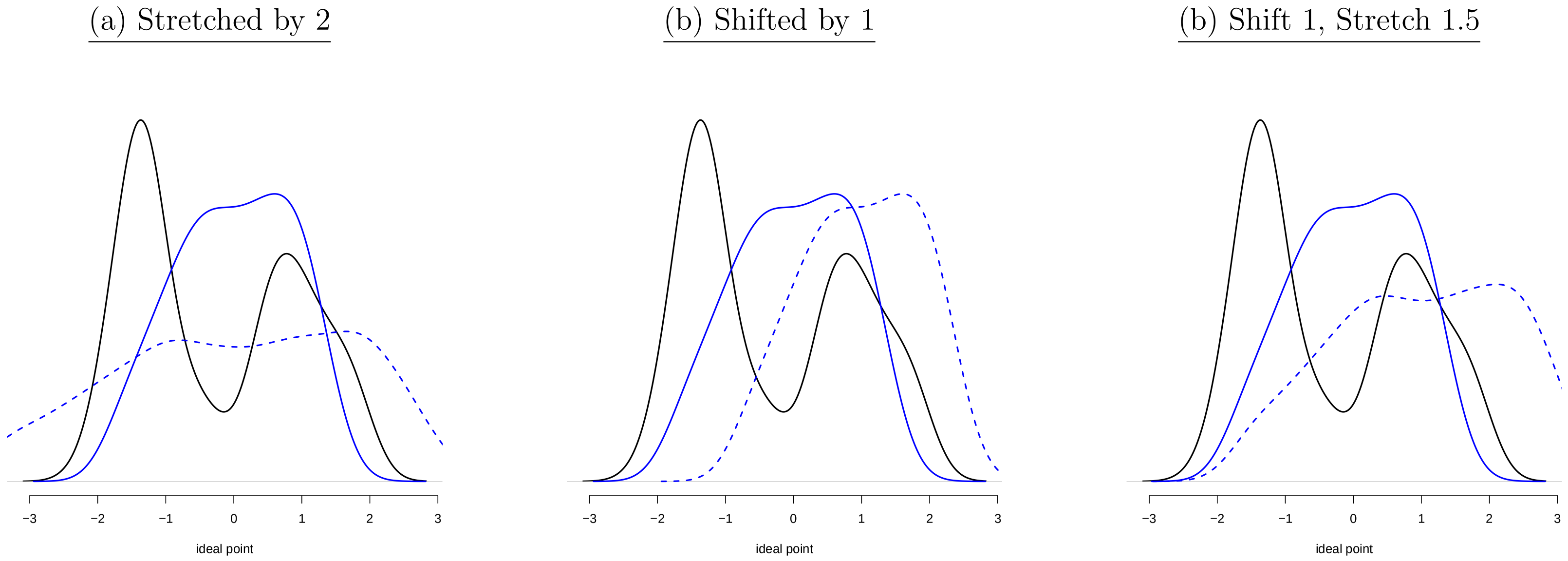
		“Stretch”			
		1	1.1	1.5	2
“Shift”	0	0	35	239	578
	0.5	72	79	213	514
	1	423	373	351	556
	1.5	1065	948	687	719
	2	1915	1749	1252	1031

Table 4: Artificial changes in BIC

In order to visualize these differences, Figure 16 shows graphs of three different hypothetical changes in the distribution of citizens relative to legislators. Each of these changes would cause an increase in the BIC over the bridged model that is close to 502 (they are 578, 423, and 351, respectively). The solid black line in each panel is the estimated distribution of positions of the legislators, and the solid blue line is the estimated distribution of positions of the voters. The dotted blue line in each panel is the hypothetical alternative set of positions of the voters. If each of these cases is a plausible alternative, then we can say very little about the representativeness of the legislators, or whether voters are choosing legislators who are close to them in terms of policy positions.

So far, in every example we have analyzed, including the media examples, the joint scaling

Figure 16: Jessee (2009) Re-Estimates



(a) Stretched by 2

(b) Shifted by 1

(b) Shift 1, Stretch 1.5

Note: The black line shows the estimated distribution of the legislators and the blue line shows the estimated distribution of the voters when bridged. The dotted blue line shows a series of alternatives in which the voters are either stretched, shifted, or both. The corresponding loss in terms of the BIC is 578, 423, and 351 for panels a, b and c respectively. By comparison, moving from the unbridged model to the bridged model comes at a loss of 502 to the BIC.

has failed. In order to show that this is not *necessarily* the case, we present two cases where it succeeds. In the most trivial case, if we simulate data from a data generating process where the common item assumption is true, the bridged model is preferred to the unbridged model. We also demonstrate this is true in the case of a single session of congress, where we "unbridge" legislators by seperating them into two groups at random and disconnecting their votes on shared roll calls.

For the simulation, we draw 300 ideal points, and 60 sets of item parameters, each from a standard normal distribution. We simulate responses where all people respond to all items according to the quadratic item response model shown above. In the "unbridged" model we seperate people into two groups at random and seperate their responses to each item. Table 5 shows the results. The BIC and DIC are lower for the bridged model than the unbridged model, showing a clear preference for pooling. In terms of out-of-sample prediction, the two models are virtually identical.

	DIC	BIC[11]	CV:group 1	CV:group 2
Bridged	15,201	18,384	69.1%	63.6%
Unbridged	15,285	19,401	69.5%	63.9%
Difference	-84	-1,017	-0.4%	-0.3%

Table 5: Fit Statistics for Data from Simulations

Rather than relying completely on data from a Monte Carlo simulation, we test a case where the bridging assumption is widely believed: bridging between legislators in the United States Congress. We take the members of the 112th House of Representatives and randomly divide them into two groups. We estimate a one-dimensional model where roll call votes are either bridged or not bridged. Table 6 shows the fits statistics for the two respective models. Once again, the bridged model is preferred by both the BIC and DIC and the out of sample fit is practically identical.

	DIC	BIC[12]	CV:group 1	CV:group 2
Bridged	74,009	89,098	88.7%	88.3%
Unbridged	74,521	100,534	88.8%	88.5%
Difference	-512	-11,437	-0.1%	-0.2%

Table 6: Fit Statistics for Data from Random Unbridging of the 112th House of Representatives

For those datasets where the bridging assumptions do not hold, solutions to this problem are not immediately apparent. The fact that errors are concentrated in the legislator choices on the bridged items does not mean that data should be dropped in order to avoid this problem. Fit could possibly be improved by dropping legislator-only items, because the model would be less constrained in the positions of the legislators. However, this is precisely why the common item parameter assumption does not hold. If these assumptions were justified, then additional data would only decrease uncertainty over the parameter values.

8 Conclusion

Overcoming the comparability challenge in preference data is one of the most important tasks for those of us seeking to understand how and to what extent the preferences of different groups relate to one another. Unfortunately, the method of joint scaling has not yet achieved this goal in every context. The common item parameter assumption is a strong constraint that does not pass empirical tests in the context of joint scaling of voters and legislators or comparing media slant to legislator or constituent ideology.

It is tempting to overlook the problems of joint scaling when the results seem to confirm our intuition. Unfortunately, to do this would be to turn these empirical projects on their heads. Until we have provided a model resting on convincing assumptions, we cannot draw firm conclusions about the proximity of the preferences of disparate groups.

However, just because joint scaling has not yet succeeded does not mean it will not be the foundation for methods that will succeed. Some possible avenues include using more general models or coming up with theoretically motivated reasons for item selection that beget items that pass the sort of tests presented here. The psychometrics and measurement theory literatures are rich with alternative models which include higher dimensional spaces, varying errors and hierarchical structures.

There are also some alternatives. Joint scaling is often undertaken under the assumption that, for instance, political representation or media slant can be thought of as a mapping from a space of preferences onto itself. It is no doubt the case that for some groups the preference spaces are distinct, and we should think of the relationships between these groups as mappings between two different spaces. In the context of media slant, we should recognize that the editorial and journalistic decisions have a complex relationship with political values. In the context of representation, it is worth revisiting the proposals of Achen (1978) and broadening our conception of how accountability might work.

References

Abramowitz, Alan I and Kyle L Saunders. 2008. “Is polarization a myth?” *The Journal of Politics* 70(02):542–555.

Achen, Christopher H. 1977. “Measuring Representation: Perils of the Correlation Coefficient.” *American Journal of Political Science* 21(4):pp. 805–815.
URL: `http://www.jstor.org/stable/2110737`

Achen, Christopher H. 1978. “Measuring Representation.” *American Journal of Political Science* 22(3):pp. 475–510.
URL: `http://www.jstor.org/stable/2110458`

Aldrich, John H and Richard D McKelvey. 1977. “A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections.” *American Political Science Review* 71(01):111–130.

Ansolabehere, Stephen, Jonathan Rodden and James M Snyder. 2008. “The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting.” *American Political Science Review* 102(02):215–232.

Bafumi, Joseph and Michael C. Herron. 2010. “Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress.” *American Political Science Review* 104:519–542.
URL: `http://dx.doi.org/10.1017/S0003055410000316`

Bailey, Michael A. 2007. “Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency.” *American Journal of Political Science* 51(3):pp. 433–448.
URL: `http://www.jstor.org/stable/4620077`

Baker, Stuart G. 1994. “The Multinomial-Poisson Transformation.” *Journal of the Royal Statistical Society. Series D (The Statistician)* 43(4):pp. 495–504.
URL: `http://www.jstor.org/stable/2348134`

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23(1):76–91.

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Medias Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(01):62–78.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace." *Working Paper* .

Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole and Howard Rosenthal. 2013. "The Structure of Utility in Spatial Models of Voting." *American Journal of Political Science* forthcoming.

Clinton, Joshua D, Anthony Bertelli, Christian R Grose, David E Lewis and David C Nixon. 2012. "Separated powers in the United States: The ideology of agencies, presidents, and congress." *American Journal of Political Science* 56(2):341–354.

Clinton, Joshua D, Molly C Jackman and Saul P Jackman. 2013. "Characterizing Chief Executives: Comparing Presidential and Congressional Preferences and Their Effect on Lawmaking, Agency Budgeting, and Unilateral Executive Action, 1874–2010." *manuscript* .

Clinton, Joshua D., Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(02):355–370.

Converse, Philip E. 1964. "The nature of belief systems in mass publics." *Critical Review* 18(1-3):1–74.

Erikson, Robert S. 1978. "Constituency Opinion and Congressional Behavior: A Reexamination of the Miller-Stokes Representation Data." *American Journal of Political Science* 22(3):pp. 511–535.

URL: `http://www.jstor.org/stable/2110459`

Gabel, Matthew J and John D Huber. 2000. "Putting parties in their place: Inferring party left-right ideological positions from party manifestos data." *American Journal of Political Science* pp. 94–103.

Gentzkow, Matthew and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78(1):35–71.

Groseclose, Tim and Jeffrey Milyo. 2005. "A Measure of Media Bias." *The Quarterly Journal of Economics* 120(4):1191–1237.
URL: `http://qje.oxfordjournals.org/content/120/4/1191.abstract`

Hill, Seth J and Chris Tausanovitch. 2014. "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization." *manuscript* .

Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103:59–81.
URL: `http://journals.cambridge.org/article_S000305540909008X`

Jessee, Stephen A. 2010. "Partisan Bias, Political Information and Spatial Voting in the 2008 Presidential Election." *The Journal of Politics* 72:327–340.
URL: `http://dx.doi.org/10.1017/S0022381609990764`

Jessee, Stephen A. 2012. *Ideology and spatial voting in American elections.* Cambridge University Press.

Jessee, Stephen A. 2015. "(How) Can We Estimate the Ideology of Citizens and Political Elites on the Same Scale?" *manuscript* .

Kass, Robert E and Adrian E Raftery. 1995. "Bayes factors." *Journal of the american statistical association* 90(430):773–795.

Kousser, Thad, Jeffrey B Lewis and Seth E Masket. 2007. "Ideological adaptation? The survival instinct of threatened legislators." *Journal of Politics* 69(3):828–843.

Lo, James. 2013. "Voting Present Obama and the Illinois Senate 1999-2004." *SAGE Open* 3(4):2158244013515684.

Lo, James, Sven-Oliver Proksch and Thomas Gschwend. 2014. "A common left-right scale for voters and parties in Europe." *Political Analysis* 22(2):205–223.

Miller, Warren E. and Donald E. Stokes. 1963. "Constituency Influence in Congress." *The American Political Science Review* 57(1):45–56. ArticleType: primary_article / Full publication date: Mar., 1963 / Copyright 1963 American Political Science Association.

Plummer, Martyn et al. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. Vol. 124 Vienna p. 125.

Poole, Keith T. 2007. "Changing Minds? Not in Congress!" *Public Choice* 131(3-4):435–451.

Poole, Keith T and Howard Rosenthal. 1984. "The polarization of American politics." *The Journal of Politics* 46(04):1061–1079.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress : a political-economic history of roll call voting*. New York: Oxford University Press.

Poole, Keith T and Howard Rosenthal. 2001. "D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting." *Legislative Studies Quarterly* pp. 5–29.

Powell, Eleanor Neff. 2015. "Pure Position-Taking in the US House of Representatives.".

Romer, Thomas and Howard Rosenthal. 1979. "The elusive median voter." *Journal of Public Economics* 12(2):143 – 170.
URL: `http://www.sciencedirect.com/science/article/pii/0047272779900100`

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(3):417–448.

Shor, Boris and Nolan McCarty. 2011. "The ideological mapping of American legislatures." *American Political Science Review* 105(03):530–551.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *Journal of Politics* p. Forthcoming.

Treier, Shawn. 2009. "Where does the president stand? Measuring presidential ideology." *Political Analysis* p. mpp035.

Volden, Craig, Alan E Wiseman and Dana E Wittmer. 2013. "When Are Women More Effective Lawmakers Than Men?" *American Journal of Political Science* 57(2):326–341.

Zaller, John. 1992. *The nature and origins of mass opinion.* Cambridge university press.