Has Joint Scaling Solved the Achen Objection to Miller and Stokes?*

PRELIMINARY DRAFT

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

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

February 6, 2013

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 of legislators and constituents proposes a solution to this problem. However, we show that this solution is insufficient. Jointly estimated item response models of legislator and constituent positions rest on the assumption that some items have the same parameter values for both groups. We show that this assumption fails statistical tests. Legislators take positions in a context that is fundamentally different from the context of public opinion polling.

*This paper is an early stage draft. Feedback is greatly appreciated. We would like to extend our thanks to Michael Herron and Stephen Jessee 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.

1 Introduction

Miller and Stokes (1963) are often credited with beginning the era of large-scale, empirical studies of ideological representation in political science [CITE]. Constituents are usually too numerous to observe directly, so scholars before Miller and Stokes typically drew conclusions based on the extent to which legislators appeared to be reacting to constituent pressure. 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. After centuries of theorizing about democracy, their paper opened the door to empirically examining political representation in a systematic way.

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 (1978) 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 position 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. 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 difference between the two measures of preferences. We could even use the absolute value of the difference. 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 the individual preferences are much more prone to error than legislator preferences, and may have a different structure altogether. Ansolabehere, Rodden and Snyder (2008) point out that scales of voter ideology based on multiple measures demonstrate much more ideological behavior than measures based on only one measure, such as the measures

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, the literature has proposed a solution to the problem of making proximity comparisons between legislators and constituents. The methodology that has been put forward is to create a *joint* scale by assuming that votes or positions taken by public office-holders can be linked to survey responses on particular questions answered by their constituents. Item response scaling is used to create common scales for both legislators and constituents. 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 have the same functional form as responses to roll call votes, they link the two populations, and make proximity comparisons between them. This method shows a way forward out of Achen's dilemma.

Many influential papers in a variety of areas have used the method of joint scaling to compare disparate populations. In the legislative politics literature, it has been used to compare the positions of legislators in different chambers or different legislative sessions (e.g. Poole and Rosenthal, 1997). Before Bafumi and Herron used this method to examine representation, Jessee (2009) used it to test theories of spatial voting. The method of joint scaling has been used to compare the positions of judges to those of elected officials (Bailey, 2007), respondents to different public opinion surveys (Tausanovitch and Warshaw, 2013), legislators and donors (Bonica, 2013), state legislators and members of Congress (Shor and McCarty, 2011), and legislators to members of the media (Groseclose and Milyo, 2005).

Although there is a growing list of influential papers that use the method of joint scaling, none of them have tested the key assumption on which joint scaling rests: that certain positions or item response functions are constant across groups (or in some cases, that responses are a particular function of support for a particular position or person). This is the first paper to test this assumption. We focusing on the case of proximity of legislators to voters, and in particular on the use of so-called "common" or "bridging" items to link the two populations.

The assumptions that underlie joint scaling are very strong. It is assumed functional form

of the dependence of item responses on underlying preferences is exactly the same across groups for a certain set of items. In the case of legislators and voters, it is typically assumed that for a given question, a constituent's preferences determine her response in exactly the same way that a legislator's preferences determine her support for a roll call vote on a similar subject.

Assuming the exact same functional form 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 environment [CITE ZALLER]. 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. What is worse, these choices often have important features that even careful outside observers could be forgiven for missing. A bill is a 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 are not just similar but exactly the same. As [CITE JESSEE BOOK HERE] 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.

In this paper, we show that two datasets from recent work that jointly scale the preferences of legislators and the public fail statistical tests of the common-item assumption. We can reject the hypothesis that the parameters of items are the same across groups with a very high degree of certainty. In case any doubt remains, we also show that the disparity in likelihood between the "common items" or "joint" model and the non-joint model is greater than the disparity in

likelihood that would result from big shifts in the relative positions of the two groups. In other words, if we accept the common item assumption, we must also accept a very wide range of relative locations for the two groups. As a result, works based on joint scaling fail to overcome the criticism that was leveled against Miller and Stokes.

We begin by explaining our model and the general model of joint scaling in an item response framework. In this discussion we also establish the statistical tests which will be used to evaluate the joint models. In the next section, we explain the datasets used for these tests. The following section explains our results. In conclusion, we briefly comment on the state of the literature and speculate on the possibility of overcoming Achen's objection.

2 Estimating Preferences on a Common Scale

Following most of the recent papers in the joint scaling literature, we employ the item response model familiar to political science from Clinton, Jackman and Rivers (2004). 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.[1] 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). 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"

[1] 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."

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 liberalism or conservatism (x_i). We would expect β to be close to 0 if we ask a completely irrelevant question; 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 missing at random.[2] 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 $x_i s$ 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.[3] 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, (x_i's). 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. However, so-called "joint" scaling applications call this assumption in to question.

A "joint" scaling exercise is one that attempts to estimate a latent quantity on a common scale for two or more separate populations. These could be: legislators at two separate points in time

[2][XXX ADD CITES TO EVERYONE ELSE EVER MAKING THIS ASSUMPTION AND KEITH TALKING ABOUT NOT MAKING IT AND JAMES LO NOT MAKING IT IN HIS DISSERTATION.]

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

(as in XXX); legislators, presidents, and Supreme Court justices (as in Bailey, 2007); or legislators and politicians (as in Bafumi and Herron, 2010; Jessee, 2010). 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 $k \in \{r, s\}$.[4] 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 2 shows examples of four types of $N \times M$ response matrices, $[y_{ij}]_{ij}$. Each matrix has a columns 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 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 not 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. Sim-

[4] 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, $\mathcal{J}_k = \cup_{i \in \mathcal{I}_k} \mathcal{J}(i)$.

ilarly, 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

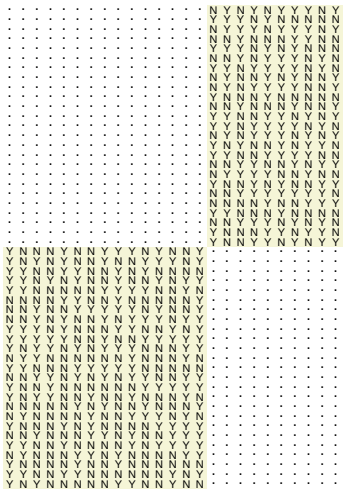
$$x_i = \delta_1 + \delta_2 x_i^*$$

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

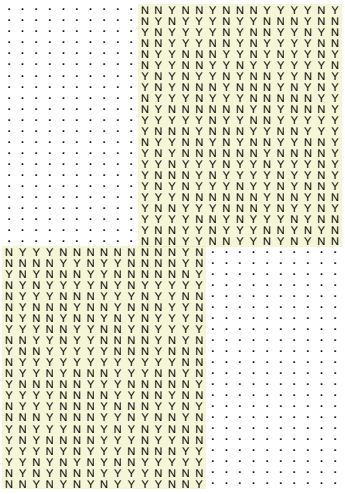
$$\ell(\theta^*; y, \mathcal{I}_s, \mathcal{J}_s) = \sum_{i \in I_s} \sum_{j \in J_s(i)} y_{ij}(\Phi(\beta_j^* x_i^* - \alpha_j^*)) + (1 - y_{ij})(1 - \Phi(\beta_j^* x_i^* - \alpha_j^*))$$

where $\alpha_j^* = \alpha_j - \delta_1 \beta_j$, $\beta_j^* = \delta_2 \beta_j$, is the entire vector of parameters transformed to the new scale. 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 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.

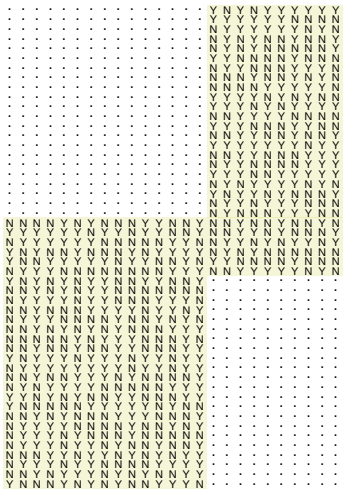
In order to place different groups on the same scale, there must be overlap either in the items that each group responds to or in group membership (data matrices of types II, III, or IV). That is,



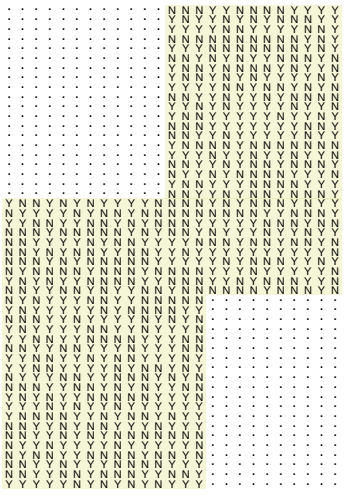
(a) Type I: No overlap



(b) Type II: Common items



(c) Type III: Common people



(d) Type IV: Common people and items

Figure 1: 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.*

the data matrices containing responses of the two groups must be "glued" or "stitched" together through common rows and/or common columns.[5]

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).

[XXX MAKE TABLE LISTING PAPERS IN WHICH JOINT SCALING IS CONDUCTED USING VOTE MATRICES OF DIFFERENT TYPES]

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.[6]

Here our interest is 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.[7] 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 the how legislators make the same translation. For example, opinions

[5] [XXX NOTE THAT KP INTRODUCED THE NOTION OF "GLUE" IN THIS CONTEXT]

[6] In order to get around this problem, interest groups and newspaper editorial board have sometimes been used as pseudo-legislators whose public positions of legislation are taken as their "votes" on corresponding roll calls. If these board or interest groups take positions ("vote") across time or venue, their votes can be used to cement otherwise disjoint voting matrices together (see [XXX ADD CITES].

[7] [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.

offered in a survey might be poor substitutes for the counter-factual in which an 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 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, 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 bridged items can be written as

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

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 both 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 arranging, 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^* 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 in such that their item parameters are on the common item are equivalent to those the respondents (or senators) with no effect on the fit of the model.

If instead of a single bridging item, we have several bridging item, 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.

In what follows, we take data from two seminal papers that attempt to jointly estimate the preferences of elected officials and the public. In both cases, we ask whether the data support the

identifying assumption that the common items function in the same way for members of the public as they do for members of the legislature (or the president).

3 Data

We use two datasets as cases to evaluate the joint scaling of survey respondents and legislators. The first comes from Jessee (2009). Although Jessee's focus is on spatial voting, not representation, his 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 responses 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 2. 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 much 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 othere words, even though Bafumi and Herron's data has the structure that appears as type IV in Figure 2, 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 responds to 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.

4 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 2. 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 2. Table 1 shows the fit statistics for the constrained and unconstrained models estimated using the data from Jessee (2009). This table contains four different measures of fit: the percent of choices that are correctly predicted by the model (PCP), the log likelihood, the Deviance Information Criterion (DIC) and the Geometric Mean Probability (GMP). A Bayes factor of e^{639} gives very strong support for the "not joint" model. The DIC, a measure which accounts for the additional parameters in the "not joint" model, also supports it.

It may seem puzzling that the likelihood difference between these two models is big, but the difference in classification is small. This may seem to beg the question of whether we are inferring to much from a very small slice of the data. Indeed, these differences come from a small part of the overall choice matrix, but one that is particularly consequential. Table 2 shows the percent of the items that are correctly classified in the joint model, the not joint model, and a naive

Table 1: Fit Statistics - Jessee Data

Statistic	Joint Model	Not Joint Model
Percent Correctly Predicted	0.831	0.833
Log Likelihood	-41906	-40775
DIC	88305	81686
GMP	0.708	0.711
Bayes Factor (Not Joint v. Joint)	e^{639}	

This table compares fit statistics for the joint and not joint models using the Jessee data.

model that simply assumes that ever person selects the majority choice, by person and item subsets. Rows represent the groups (legislators, survey respondents, and both), and columns represent sets of items (the "bridged" or shared items, items answered only by either group, and all items). What this table shows is that the differences in classification success some almost exclusively from legislator responses to the "bridged" questions. The difference here is big: about 7 percentage points higher classification success in the non-joint case. This is almost a third of the total improvement that the non-joint model made over the naive model.

Table 2: Percent Correctly Predicted Among Subsets - Jessee Data

	"Bridged" Questions Only		Other Questions		All Questions	
Joint Estimation	Naive PCP	PCP	Naive PCP	PCP	Naive PCP	PCP
Legislators	0.648	**0.814**	0.715	0.918	0.712	0.914
Survey Respondents	0.638	0.759			0.638	0.759
All	0.635	0.761	0.715	0.918	0.671	0.832
Not Joint Estimation	Naive PCP	PCP	Naive PCP	PCP	Naive PCP	PCP
Legislators	0.648	**0.883**	0.715	0.918	0.712	0.917
Survey Respondents	0.638	0.760			0.638	0.760
All	0.635	0.765	0.715	0.918	0.671	0.834

This table compares the percent of choices that are correctly classified for subsets of choices using the Jessee data.

It is notable that the classification success on the shared items for respondents does not substantially decrease in the joint model. These are the only items that survey respondents are asked, so they make up the complete likelihood for survey respondents. The model has estimated

item parameters that better explain the choices of the much larger number of survey respondents than they explain the choices of the much smaller number of legislators, and that has resulted in drastically decreased fit for one group but not the other.

The difference in likelihood between the two models is sufficient to conclude that the item parameters should not be assumed to be the same. However, to demonstrate this further we compare the different parameters that are estimated when the bridged questions are estimated seperately for each group. Because the parameters of these two models are on different scales, we project the item parameters for the respondents onto the item parameters for the legislators linearly. We transform the posterior densities of the parameters by the same linear projection, and compare these posteriors. If the item parameters are the same for the two groups, then we should find no difference between the posteriors of the transformed item parameters for each group.

Figure 2 shows the cutpoints of the common questions when they are "unbridged" for each group. Figure 3 compares the discrimination parameters. 95% credible intervals are shown. In each of these graphs, we can see that the posteriors of these item parameters rarely overlap- we can reject the null that they are the same most of the time for both the cutpoints and the discrimination parameters. In fact it is not correct to infer that *any particular item* is different between the two groups because the posteriors of the discrimination parameters or cutpoints do not overlap. What these graphs show is that the joint model rests on a wrong assumption, so no comparison between these items can be made.

Finally, we turn to the question of what effect a wrong specification has on our inferences about the relative location and spread of the two groups. Comparing the location and spread of the ideal points of legislators to those of survey respondents is a dubious endeavor now that we know that these estimates rest on incorrect constant item parameter assumptions. However,say for a moment that despite the large difference in likelihood between the joint and not joint models, we still believe our results. How certain should we be about the relative position and spread of the two groups?

Figure 4 is a contourplot that shows the likelihood of the model under different assumed values of the relative shift and spread of the two groups. d1 is an adjustment to the relative location of

the groups, where and adjustment of 1 would move all legislators to the right of their estimated position by 1 standard deviation of the scale of x. d2 is the relative spread of the groups, where a value of 2 implies that the spread of the legislators is doubled relative to their estimated spread, with the spread of the respondents held fixed.

Using a grid of values for d1 and d2, we re-estimate a restricted version of our model using Expectation Maximization (EM) algorithm to find the optimal item parameters for a given d1 and d2. 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 contour on the plot represents an area of d1 and d2 of equal likelihood, where the likelihood is calculated for the bridged items only.

As a reference point for this plot: the total log likelihood difference between the joint and not joint models for only the bridged items is about 637. That is the difference between not joint model and the likelihood of the joint model with 0 shift (d1=0) and no stretch (d2=1). In order to achieve that large of a reduction in likelihood from the joint model, we would need to shift the groups by 0.5 (d1=0.5 or 1.5) or stretch the groups by a factor of almost 2 (d2=2). These are very large differences, enough to substantially shift any conclusions about the relative proximity of legislators and their constituents.

For ease of interpretability, Figure 5 plots this figure again, but this times in terms of Geometric Mean Probability, not likelihood. It can be seen that even big changes in the stretch and shift of the two groups produces only relatively small changes in the probabilities of the responses.

We replicate all of the above analyses using the data from Bafumi and Herron. As explained in the data section, the "not joint" model maintins the constant item parameter assumption for all items except the ones that are shared by legislators and survey respondents. Table 3 shows the fit statistics for the two models. As before, the fit of the "not joint" model is much higher, safely rejecting the "joint" model.

Table 4 breaks down the classification percentages by model and subgroup. As before, the main difference between the joint model and the not joint model is that the latter does a much better job of explaining the votes of legislators on bridging items. Classification for this set of choices is

Table 3: Fit Statistics - Bafumi and Herron Data

Statistic	Joint Model	Not Joint Model
Percent Correctly Predicted	0.886	0.887
Log Likelihood	-154619	-152981
DIC	317916	306120
GMP	0.781	0.783
Bayes Factor (Not Joint v. Joint)	e^{1257}	

This table compares fit statistics for the joint and not joint models using the Bafumi and Herron data.

reduced by 6 percentage points, almost 25% of the reduction in error from the "not joint" model over the naive model.

Table 4: Percent Correctly Predicted Among Subsets - Bafumi and Herron Data

	"Bridged" Questions Only		Other Questions		All Questions	
Joint Estimation	Naive PCP	PCP	Naive PCP	PCP	Naive PCP	PCP
Legislators	0.634	**0.823**	0.732	0.941	0.730	0.939
Survey Respondents	0.608	0.819	0.639	0.816	0.631	0.817
All	0.601	0.820	0.732	0.895	0.686	0.886
Not Joint Estimation	Naive PCP	PCP	Naive PCP	PCP	Naive PCP	PCP
Legislators	0.634	**0.884**	0.732	0.942	0.730	0.940
Survey Respondents	0.608	0.820	0.639	0.816	0.631	0.817
All	0.601	0.827	0.732	0.895	0.686	0.887

This table compares PCP stats for subsets using the Herron data.

Figures 6 and 7 show the projection of the "not joint" item parameters on two each other for the common items. As before, most of these parameters as statistically distinguishable, which is inconsistent with the common item parameter assumption.

Finally, Figures 8 and 9 replicate Figures 4 and 5 above using the Bafumi and Herron data. The results are similar. The difference between the log likelihoods of the joint and not joint models is 1257. With no greater reduction in log likelihood, the dispersion of the legislators could be multiplied by 2, reduced by 0.5, or their entire distribution could be shifted by .25 relative to the respondents. Within this range of uncertainty, few conclusions can be drawn about the relative position of the legislators and the respondents.

There are is no easy solution that we know of to the problems posed in this section. 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, becuase the model would be less constrained in the positions of the legislators. However, this is precisely why the common item parameter assumptions do not hold. If theses assumptions were justified, then more data would only decrease uncertainty over the parameter values.

5 Conclusion

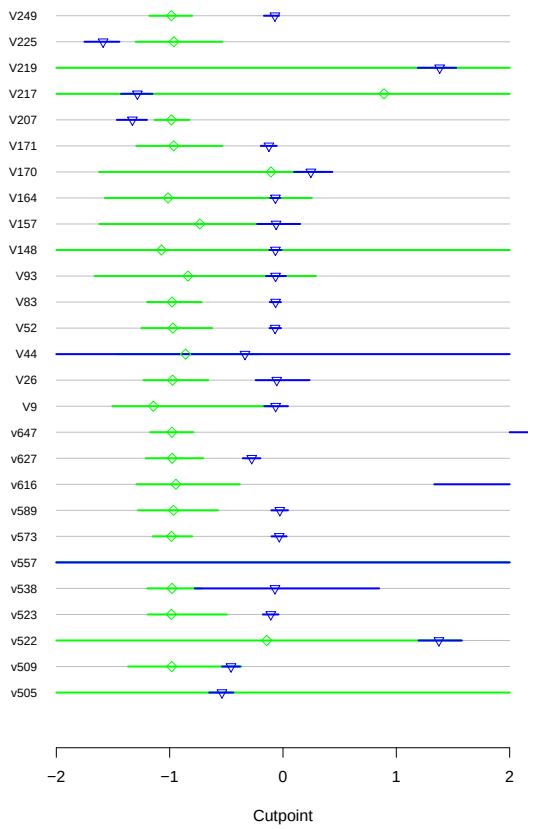
Overcoming Achen's challenge to Miller and Stokes is one of the most important tasks for those of us seeking to understand how and to what extent legislators represent their constituents. Unfortunately, the method of joint scaling has not yet achieved this goal. The common item parameter assumption is a strong constraint that does not pass statistical tests in the context of representation.

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 the empirical project of understanding representation on its head. Until we have provided a model resting on convincing assumptions, we can not draw firm conclusions about the proximity of voter positions to legislator positions.

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 analyses presented here.

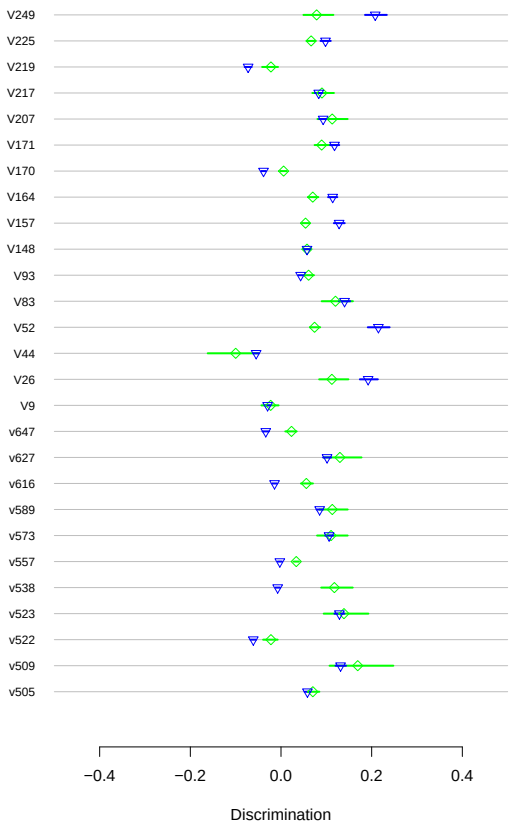
There are also some alternatives. Joint scaling assumes that representation is a mapping from a space of preferences onto itself. It may be the case that the space of respondents preferences and the space of legislator preferences are distinct, and we should think of representation as a mapping between these two different spaces. With this in mind, it is worth revisiting the proposals of Achen (1978) and broadening our conception of what is meant by representation.

Figure 2: Cutpoints - Jessee Data



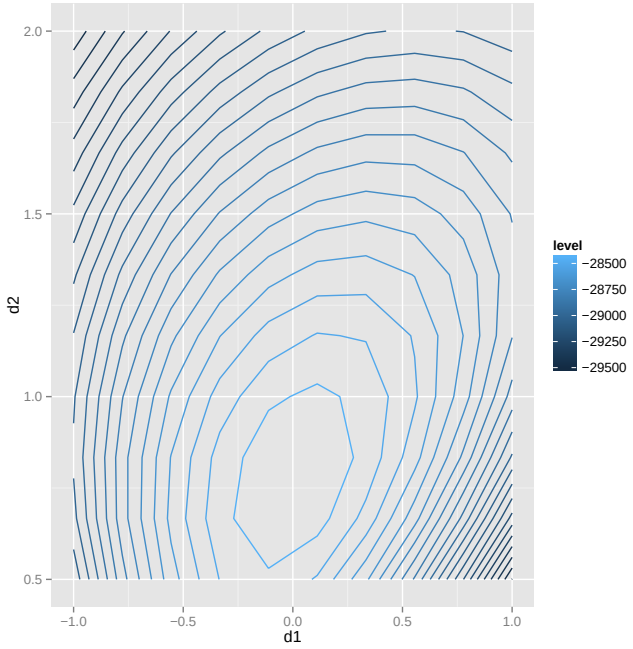
Estimated cutpoints of common items for legislators and respondents. *This plot shows the estimated cutpoints of the items that are assumed common in the "joint" model as estimated from the "not joint" model, using the Jessee data. The cutpoints for the legislators are projected onto the corresponding cutpoints for the respondents. The estimated cutpoints for the legislators are marked by a green diamond and the estimated cutpoints for respondents are marked by a blue triangle. Lines represent 95% credible intervals.*

Figure 3: Discrimination - Jessee Data



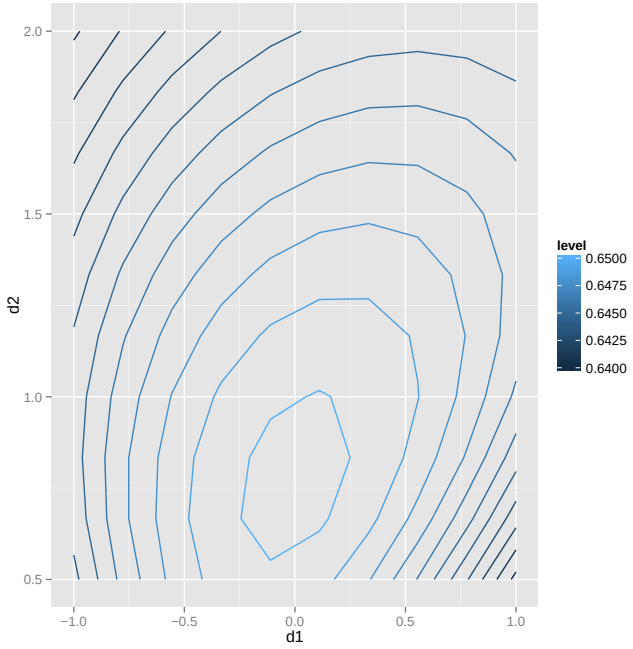
Estimated discrimination of common items for legislators and respondents. *This plot shows the estimated discrimination parameters of the items that are assumed common in the "joint" model as estimated from the "not joint" model, using the Jessee data. The discrimination parameters for the legislators are projected onto the corresponding discrimination parameters for the respondents. The estimated discrimination parameters for the legislators are marked by a green diamond and the estimated discrimination parameters for respondents are marked by a blue triangle. Lines represent 95% credible intervals.*

Figure 4: Likelihood Effect of Alternative Configurations - Jessee Data



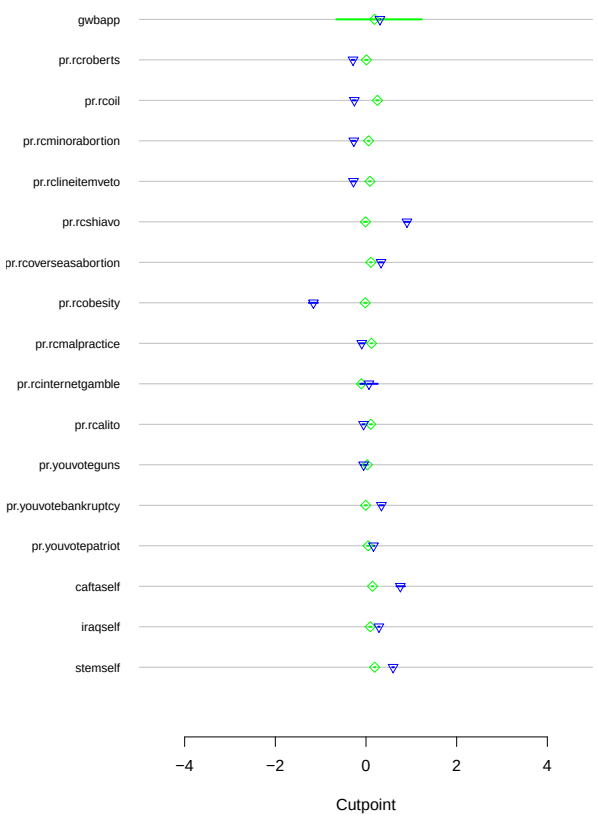
This plot shows the effect on the likelihood of altering the shift and stretch between the distribution of legislator and respondent ideal points.

Figure 5: GMP Effect of Alternative Configurations - Jessee Data



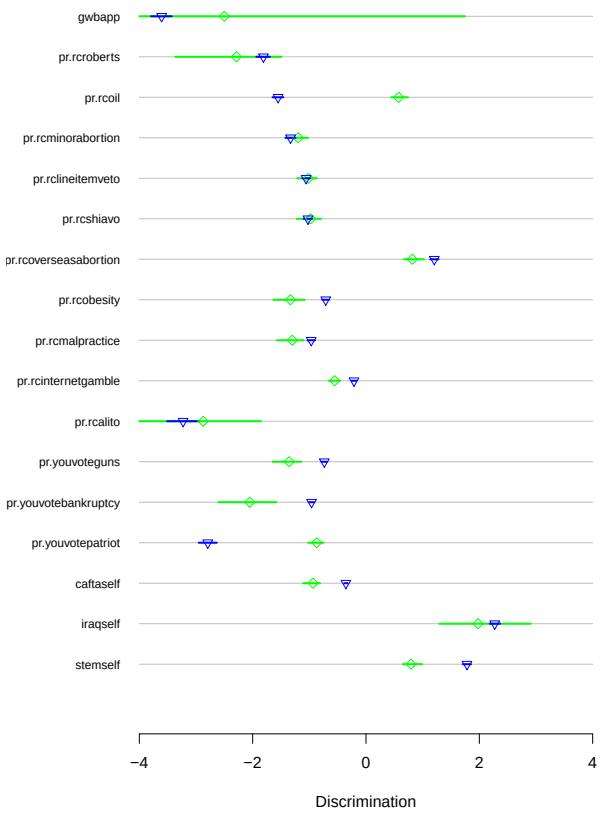
This plot shows the effect on the Geometric Mean Probability of altering the shift and stretch between the distribution of legislator and respondent ideal points. This is isomorphic to Figure 4.

Figure 6: Cutpoints - Bafumi and Herron Data



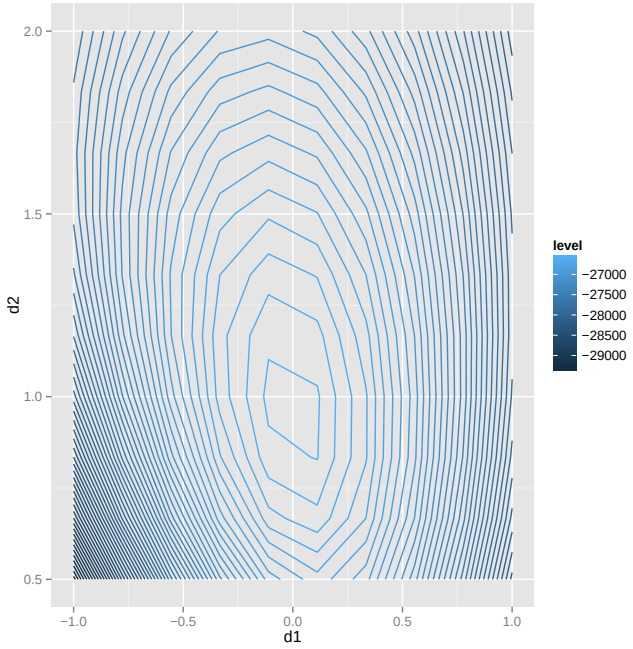
Estimated cutpoints of common items for legislators and respondents. *This plot shows the estimated cutpoints of the items that are assumed common in the "joint" model as estimated from the "not joint" model, using the Bafumi and Herron data. The cutpoints for the legislators are projected onto the corresponding cutpoints for the respondents. The estimated cutpoints for the legislators are marked by a green diamond and the estimated cutpoints for respondents are marked by a blue triangle. Lines represent 95% credible intervals.*

Figure 7: Discrimination - Bafumi and Herron data



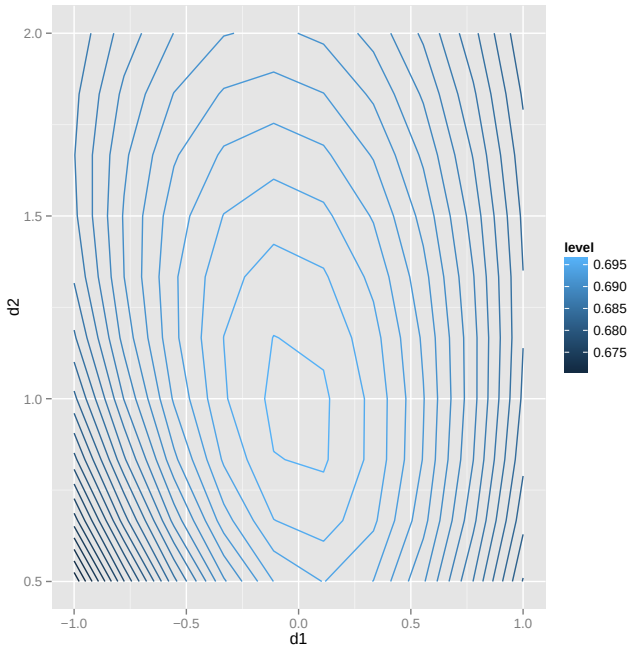
Estimated discrimination of common items for legislators and respondents. *This plot shows the estimated discrimination parameters of the items that are assumed common in the "joint" model as estimated from the "not joint" model, using the Bafumi and Herron data. The discrimination parameters for the legislators are projected onto the corresponding discrimination parameters for the respondents. The estimated discrimination parameters for the legislators are marked by a green diamond and the estimated discrimination parameters for respondents are marked by a blue triangle. Lines represent 95% credible intervals.*

Figure 8: Likelihood Effect of Alternative Configurations - Bafumi and Herron Data



This plot shows the effect on the likelihood of altering the shift and stretch between the distribution of legislator and respondent ideal points.

Figure 9: GMP Effect of Alternative Configurations - Bafumi and Herron Data



This plot shows the effect on the Geometric Mean Probability of altering the shift and stretch between the distribution of legislator and respondent ideal points. This is isomorphic to Figure 8.

References

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

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

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., 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

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

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

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.

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

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

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