Estimating Better Left-Right Positions Through Statistical Scaling of Manual Content Analysis

Thomas Däubler* Kenneth Benoit†

February 13, 2017

Abstract

Borrowing from automated "text as data" approaches, we show how statistical scaling models can be applied to hand-coded content analysis to improve estimates of political parties' left-right policy positions. We apply a Bayesian item-response theory (IRT) model to category counts from coded party manifestos, treating the categories as "items" and policy positions as a latent variable. This approach also produces direct estimates of how each policy category relates to left-right ideology, without having to decide these relationships in advance based on out of sample fitting, political theory, assertion, or guesswork. This approach not only prevents the misspecification endemic to a fixed-index approach, but also works well even with items that are not specifically designed to measure ideological positioning.

Key Words: Party manifestos, IRT, Bayesian estimation, Comparative Manifestos Project, policy positions, measurement.

*University of Mannheim and MZES, email: thomas.daeubler@mzes.uni-mannheim.de
†London School of Economics and Trinity College Dublin, email: kbenoit@lse.ac.uk

By a British mile, measures of *left-right* policy positioning outstrip all other measures of policy distance in the study of political party competition. With roots in early spatial descriptions of the seating in the Constituent Assembly following the French Revolution (see Carlyle, 1888, 92 in Benoit and Laver, 2006, 12–13), this orientational metaphor has proven to be one of the most resilient of all conceptual frameworks for distinguishing political actors by their policy differences. While many configurations of positioning on specific dimensions, such as economic, social or environmental policy, are possible (e.g. Benoit and Laver, 2006; Bakker et al., 2015), in practice these can be summarized using a single policy continuum (e.g. Gabel and Huber, 2000; Laver and Budge, 1992). It is meaningful, and certainly *useful*, to refer to this dimension as a "left-right" axis, ordering party and voter positions from left to right in a "manner agreed upon by all" (Downs, 1957, 142). As analysis of expert placements of parties on a "left-right" dimension — without specifying in advance what this should mean — has shown clearly that party placements on this dimension can be predicted from policy locations on more specific policy scales (Benoit and Laver, 2006, 141).

Widespread disagreement exists, however, over how to conceptualize and measure this common left-right dimension. This discord is witnessed in the numerous debates over the relative merits of expert surveys (Benoit and Laver, 2007), indices constructed from the content analysis of manifestos (Budge and Meyer, 2013), debates over how best to construct such indices (Franzmann, 2015; Jahn, 2014), the validity of scaling roll-call votes (Proksch and Slapin, 2010), and a growth industry using automated and statistical approaches to scale positions from political "text as data" (e.g. Laver, Benoit and Garry, 2003; Slapin and Proksch, 2008; Lowe, 2016; Grimmer and Stewart, 2013). Conceptually, these approaches divide into two main camps: those that define left-right *a priori* based on theoretical reasoning about its policy content, and those who treat left-right as a "superdimension" emerging from how actors bundle issues in the real world, and whose content can consequently only be inferred *a posteriori* in a given context.

We believe that the first approach faces problems that are practically impossible to overcome. First of all, it will be difficult to reach a consensus on the substantive content of left-right *a priori*

in any given context. Second, even if fit successfully to a given context, this measure developed *a priori* will not fit cases from other contexts, if the new contexts are in any way different. Finally, even if a common understanding can be established, it is unlikely to be precise enough for deriving specific measurement instructions such as how to weight different policy categories (Gabel and Huber, 2000, 95). Therefore, we propose a new method for inferring the left-right dimension *a posteriori* from high-dimensional measurements of party policy.

We apply a Bayesian measurement model rooted in Item-Response Theory (IRT) to the single largest source of evidence on cross-national party positions over time, the Manifesto Project's dataset, which is based on manual content analysis of over 3,200 manifestos in 55 countries. In contrast to approaches based on fixed definitions of left-right ideological content, our inductive method does not require any a priori assumptions about its substantive content, and permits all relevant information to be used for inference.[1] We also demonstrate that it produces position estimates from Manifesto Project data that better match expert placements than the very widely used "Rile" index. In addition, we show how the IRT approach can be extended to measure policy positions on multiple dimensions, and that it may uncover positions even based on data that were not specifically collected for this purpose, such as non-positional policy topics coded by the Comparative Policy Agendas Project.

Our method adds to existing approaches that inductively estimate left-right from content analysis data, as most prominently represented by the "vanilla" method of Gabel and Huber (2000). IRT approaches are more than data reduction techniques, since they explicitly model the relationship between subjects — in our case parties — and items (Reckase, 1997, 29). This allows for interesting empirical insights into the nature of left-right issue, and is amenable to formal testing of theories about its content. The Bayesian measurement model also has the advantage that it directly quantifies uncertainty in the inferred parameters.

Finally, we provide a bridge between *qualitative* coding methods and more automated, "text

[1] We note the analogy to the debate about how to measure democracy. Treier and Jackman (2008) point out that democracy is a controversial concept and that justifying weights for the more specific indicators from the Polity dataset is very difficult, which leads them to suggest a latent variable model. See also Pemstein, Meserve and Melton (2010) for a similar argument.

as data" methods for drawing inferences from political text. Our approach combines the best of both worlds: the high confidence in content validity that comes from expert human coders, with an extremely flexible measurement approach that infers lower-dimensional policy measurements. We recommend using the *full* set of coded statements, particularly when their degree of informativeness for left-right policy is unknown. The approach we suggest, however, also allows the user to incorporate in a formal and transparent fashion as much prior information about the policy content of left-right as desired.

Contrasting approaches to Measuring Left-Right Ideology

Two contrasting approaches exist to defining the "left-right" political dimension, with fundamentally different implications for measurement. The first perspective, often termed the *a priori* approach (Benoit and Laver, 2006), treats the substantive content of the left-right dimension as known *ex ante*, and then seeks to locate the policy positions of political actors on this dimension. Because left-right policy differences are rooted in political theory, linking ideologies to bundles of issue positions, proponents of the deductive approach tend to draw on authoritative thinkers to identify the ingredients of left-right ideology. For example, Jahn (2011, 750-751) builds on Bobbio (1996), who draws on classic distinctions between "left" and "right" attitudes toward (in)equality, as found in the thought of Rousseau and Nietzsche. The authoritativeness of the definition is then taken as conveying construct validity to the measurement of left-right according to the pre-defined content. Similarily, Budge and Meyer (2013, 89) argue that the construction of the Manifesto Project's left-right scale reflects opposing arguments by highly influential early modern theorists, including Marx and Engels on the left and Disraeli and Spencer on the right. Based on these associations, the Manifesto Project's "Rile" scale — the most widely used left-right measurement in political science — selects 13 of its policy categories as pre-defined "left" policy categories, another 13 as "right," and treats its remaining 30 policy categories as unrelated to left-right. More recent work such as Inglehart (1984) has updated these constructs, but takes essentially the same deductive

approach to identifying the issues that constitute the essential distinctions between left and right ideology in contemporary settings.

A second approach reverses the logic of inference about the left-right dimension, identifying its content as simply the sum of whatever parts it comprises (cf. Fuchs and Klingemann, 1990), such as preferences for taxes versus spending or social liberalism versus conservatism. Rather than debating what are the components of left-right, the goal of the inductive approach is to discover "the best-fitting empirical representation of the policy space under investigation, using techniques of dimensional analysis to infer latent policy dimensions and then interpreting the substantive meaning of these dimensions in terms of relative locations of key political agents on these" (Benoit and Laver, 2006, 59).

The best-known general application of the inductive approach is from Gabel and Huber (2000), whose "vanilla" method applied factor analysis to the manifesto category percentages and scored each manifesto on the first principal factor as a measurement of left-right position. From this perspective, there is no basis for establishing *a priori* the substantive meaning of left-right ideological differences; rather, "the left-right dimension is defined inductively and empirically as the 'super-issue' that most constrains parties' positions across a broad range of policies" (Gabel and Huber, 2000, 96). Franzmann and Kaiser (2006) and Franzmann (2015) also proceed in a primarily inductive manner, using regressions of policy category shares on party indicator variables to determine which categories differentiate between parties and thus provide information about positions.[2]

Proponents of the *a priori* approach argue that the grounding of its definition in a known, and fixed, frame of reference, facilitates comparison of like with like across different contexts. A scale with fixed components, so goes the argument, broadens its applicability to not just more, but also to *every possible*, context.[3] Advocates of the *a posteriori* approach, by contrast, see the content of

[2]Mixtures of *a priori* and inductive approaches are also possible. König, Marbach and Osnabrügge (2013), for instance, applies dynamic factor analysis on input data consisting of pre-selected and pre-scaled categories from the CMP's content analysis codes, as well as relying on prior information on party positions from expert surveys for inference. A similar use of expert surveys is made by Bakker (2009), who applied a two-parameter logistic IRT model to a subset of CMP categories that are pre-grouped into left and right items.

[3]Consider, for instance, the ambitious claim that the Rile index's "a priori, deductive nature is important in allowing its application in all places at all times without the qualifications about content or context which apply to inductive scales. It is a substantively invariant measure whose numeric values always carry the same meaning.. . . They apply

the left-right dimension as variable to such an extent that "it may be impossible for any single scale to measure this dimension in a manner that can be used for reliable or meaningful cross-national comparison" (Benoit and Laver, 2006, 143). The challenge then becomes one of producing valid measures of the high-level concept of left-right ideology that is meaningful for a specific sample of countries, in a way that facilitates comparison, without either hard-coding elements into that measure that may be inappropriate to that context or failing to include elements that are.

This challenge is not unique to political science. An analogous measurement problem exists in economics: the Consumer Price Index (CPI). Designed to capture the typical cost of a basket of goods and services consumed by households, the CPI consists of an index constructed from the prices paid by a designated consumer segment for a "market basket" of the goods and services purchased by a household. The CPI has to be time-invariant because its use is explicitly comparative: to track *changes* in inflation across different years. To achieve this comparability, the CPI must be adjusted in three ways. First, the sample of representative goods which the basket comprises must be updated to match changes in consumer consumption, technology, etc. It would be an invalid measure in 2016 to use a basket containing spurs or wax candles, for instance, because consumers no longer ride horses to work nor do scholars pen their papers under the flicker of sooty candles. Similarly, mobile phone applications, which were added to the (UK) market basket in 2011 for the first time (Gooding, 2011, 102), would have been unimaginable components even two decades ago. Second, even if unchanged, the weights of items must be adjusted to fit the same components appropriately to a new context. Finally, the index is only meaningful relative to a certain base, because the very nature of what is being measured (prices) is constantly changing. Valid measurement requires selecting appropriate components, appropriately weighting those components, and fixing an appropriate comparison point, none of which can be sucessfully achieved through an index fit out of sample.

When characterizing differences between political parties, the challenge is to place parties on a meaningfully comparable dimension even when we are not completely confident, *a priori*, what are

universally without having to be adjusted for particular contexts, and thus provide a promise of invariant and reliable measurement across limitations of time and space" (Budge and Meyer, 2013, 90).

the ingredients of that dimension. To address this challenge, we provide a method for determining positioning on the left-right "super issue" that also meets the "need [for] explicit criteria of how categories can be transferred to a left-right scale" (Franzmann and Kaiser, 2006, 166), by introducing an explicit model based on item-response-theory offering a number of advantages over existing inductive measurement approaches. Our IRT model constitutes a representation of the actual data generating process, i.e. manifesto writing, with an intuitive interpretation for each parameters, and model-based uncertainty measures. Using our measurement model approach, moreover, we are able to estimate not only the latent party positions, but also at the same time estimate the degree to which each policy category contributes to the content of the left-right dimension.

Scaling policy dimensions

Data: Category counts from manually coded party policy statements

Because of their regular publication and the "official" status, party manifestos have formed the main source of textual data for both manually coded content analysis research such as the long-standing Manifesto Project (e.g. Budge, Robertson and Hearl, 1987; Budge et al., 2001; Klingemann et al., 2006; Volkens et al., 2013), as well as numerous attempts to extract policy positions automatically using supervised (Laver, Benoit and Garry, 2003) or unsupervised (Slapin and Proksch, 2008) learning methods. By drawing on the rich dataset of coded policy statements from the Manifesto Project (Volkens et al., 2014), we are drawing on the same dataset used to estimate left-right ideology, both inductively and deductively, by numerous other researchers (e.g. Gabel and Huber, 2000; Franzmann and Kaiser, 2006; Laver and Budge, 1992; Jahn, 2011; Mölder, 2013).[4]

By applying unsupervised scaling models originally developed in the quantitative, computational tradition to manually coded content analytic data, we also demonstrate a way to bridge

[4]This consists of 56 core policy categories, plus an additional 51 extended categories added to cover policy in countries added since the 1980s. The Manifesto Project's coding method relies on qualitative content analysis using trained expert coders to classify the sentences of each text into a predefined set of policy categories spanning seven domains. For details see https://manifesto-project.wzb.eu/coding_schemes/1.

automated text analysis methods with more traditional, qualitative text analytic approaches. Automated scaling methods typically model "bags of words," by combining relative word frequencies to measure policy (e.g. Laver, Benoit and Garry, 2003; Slapin and Proksch, 2008). We show that this same approach is easily adapted to manually coded content categories of larger units of text such as sentences.

As a measurement model, our approach also fully satisfies the criterion that measurements should be accompanied by uncertainty estimates. In line with the general insight that more data provides more confident estimates than less data, our approach uses a more explicit model of the data generating process than Benoit, Laver and Mikhaylov (2009)'s approach based on simply simulating white noise to generate confidence intervals.

A Measurement Model for Unordered Categorical Outcomes

Our fundamental aim is to infer the "left-right" position of a political party at a specific time. All we observe, however, is a set of category codings for the manifesto text. Following Benoit, Laver and Mikhaylov (2009), we start from the notion that the party intends to communicate a certain position, called θ_i in the manifesto i. This position is fundamentally unobservable and uncertain, but will be communicated through the text. As writing proceeds, the party makes various policy statements referring to different issues, generating observable data in the form of counts of statements in different policy categories. The configurations of statements made in different party manifestos provide a basis on which we can measure their policy positions, because some policy issues are explicitly positional, or represent valence issues for which differences in emphasis represent differences in position (Stokes, 1963; Franzmann and Kaiser, 2006; Dolezal et al., 2013). Our model takes into account these considerations by modelling the latent variable as well as the left-right policy components directly using a model based in item-response theory (IRT). Policy statements form the "test items," parties correspond to the subjects, and the estimate of latent "ability" θ_i represents a party's left-right policy position. Each category's contribution to the observed outcomes is mapped via a series of item parameters that measure their association with the latent ideological

dimension.[5]

Applied to the current problem, consider for a single text i that it generates a series of statements x_k where $k = 1, \ldots, v_i$. Each statement coded from text i represents an "item," from which there are a fixed set of J possible unordered categories (statement types). Were there only two possible statement types with $J = 2$, then we could express for category $j = 1$:

$$P(x_k = 1) = \frac{e^L}{1 + e^L} \tag{1}$$

and for category $j = 2$,

$$P(x_k = 2) = 1 - P(x_k = 1) \tag{2}$$

and where the logit transformed quantity L is expressed in the familiar two-parameter logistic item-response formulation as $L = a_j(\theta_i - b_j)$.

In this formulation with just two response categories, θ_i is a latent measure of subject i's "ability" to answer the item with a response $j = 1$ (representing a "correct" answer) versus with a response of $j = 2$ (an incorrect answer). The parameters a_j and b_j represent the *discrimination* parameters and *difficulty* parameters of item j, respectively. It is only necessary to speak of a single response category for each item, since the only other response category can be expressed in terms of the probability of this item. For a text coding research design, this would be analogous to having a two category coding scheme, where a text unit might belong only to one category or the other.

Now consider the case where $J > 2$. In this situation, we replace the binomial logistic formulation with the multinomial generalization. Here,

$$Pr(x_k = j) = \frac{e^{L_k^j}}{1 + \sum_{j=1}^{J-1} e^{L_k^j}} \tag{3}$$

[5] A similar model is introduced in Elff (2013). That article focuses on estimating positions on separate dimensions, for which items are pre-selected, though. Albright (2008) applies a Bayesian binomial model to data for all the Manifesto Project categories, but does not consider results for the item parameters at all.

where L_k^j represents a "multivariate logit" (Bock, 1972), and there is a vector of J such logits for each item. This formed the basis for Bock (1972)'s *nominal response model* (NRM), generalising the two-parameter logistic IRT model from the binomial to the multinomial case of a multiple, unordered categorical response structure. In the multivariate formulation (for a given text i),

$$L_k^j = \zeta_k^j + \lambda_k^j \theta_i \tag{4}$$

where the quantities ζ_k^j and λ_k^j represent the item parameters for the jth category of response for the "item" k. (Note: k is a statement choice that is made, where this statement has to be assigned to one statement type category.) For estimation purposes, of course, the general function in Eq. 4 is not identified, because it is invariant with respect to the translation of L_k^j. We explain how we constrain and estimate this model in Appendix B. For a discussion of how to interpret item parameters in the context of multiple response categories we refer the reader to Appendix C.[6]

For a fixed and equal number of items $n_1 = n_2 = \ldots = n_I$, we could then observe counts of each category j across a set of i individuals, in the same way that we could tabulate response categories across test takers, with the important proviso that in this setup, each item would have identical response categories. This would give us $Y_{ij} = \sum_k x_{ik}$, corresponding to a matrix of counts for each response category j for each text i. In text analysis and many other settings, however, the number of items differs across cases, for example when texts differ in length. Put differently, the number of items $Y_{i.} = \sum_j Y_{ij}$ varies across i. To accommodate this, we can reformulate the NRM as a log-linear model for the expected value μ_{ij} of the counts Y_{ij} :

$$\log(\mu_{ij}) = \alpha_i + \zeta_j + \lambda_j \theta_i, \tag{5}$$

where the α_i is a parameter that represents variable text length, necessary because of the differential

[6]Implicit in our model is a quadratic spatial utility function (compare Elff, 2013): document parameter θ_i and item parameter λ_j are multiplied, which can be shown to result from the difference between a document location and a category location being squared. We do not try to uncover these locations, since we see little analytical value added by projecting both parties and items into the same space (which lacks a clear-cut interpretation, since there is no status quo and an alternative such as in the context of roll-call voting).

number of test items caused by the fact that manifestos vary (in practice, quite significantly) in length. Note that in this context α_i should not be considered an exposure variable as in standard count regression and be fixed to the log of text length; instead, the number of items constitutes itself a random variable that will recover the exact margins (cp. Agresti, 2013, 361-362).

To define a data generating process for the counts in this log-linear model, we can use the negative binomial distribution, with rate μ_{ij}, and an additional variance parameter linked to this expected rate. We use the probability mass function as defined by Cameron and Trivedi (1986, 32-33)

$$Pr(Y_{ij} = y_{ij}|\mu_{ij}, \phi_j) = \frac{\Gamma(y_{ij} + \phi_j)}{\Gamma(y_{ij} + 1)\Gamma(\phi_j)} \left(\frac{\phi_j}{\mu_{ij} + \phi_j}\right)^{\phi_j} \left(\frac{\mu_{ij}}{\mu_{ij} + \phi_j}\right)^{y_{ij}}$$

The expected value of the counts is given by $\mathrm{E}(Y_{ij}) = \mu_{ij}$, and the variance by $\mathrm{Var}(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi^{-1})$ (the "NB2" variance function). The parameter ϕ^{-1} represents the extra variance in the data relative to the special case of the Poisson ($\phi^{-1} = 0$), for which the variance is equal to the mean. One reason why we observe overdispersion may be unobserved heterogeneity at the level of the counts Y_{ij}. Put differently, in addition to the latent variable there are random effects that influence the counts in the cells of the table representing the frequency of category use for each document. To allow for the possibility that the variance systematically differs across categories, we infer a different ϕ_j parameter for each category. In the case of $\phi^{-1} = 0$, our model becomes algebraically equivalent to the "wordfish" Poisson scaling model for word counts first presented by Slapin and Proksch (2008) (see Appendix A for details). Their model, however, was not explicitly presented as an IRT model, instead using maximum likelihood to estimate θ_i by conditioning on fixed effects for the item and exposure parameters.[7]

[7] More recently, Lo, Proksch and Slapin (2016) extended their model introducing an overdispersion parameter at the document level.

Estimating Left-Right as a Latent Variable

In this section we fit the basic model to the core 56 Manifesto Project categories as "items" to estimate the single-dimensional latent variable θ_i, and compare this measure to other solutions.

Party locations on a single dimension

Fitting the main model (Eq. 5) to the 56 Manifesto Project policy counts, we are able to obtain estimates of the policy positions θ_i for each party i on a single dimension of policy.[8] We restricted our sample to manifestos issued in democratic countries after the first oil crisis which arguably changed politics considerably.[9] Table 1 presents these results for the IRT model, first for the Poisson model and second for the negative binomial model incorporating a separate variance parameter ϕ_j estimated at the policy category level.[10] For the set of all manifestos, we obtained estimates and confidence intervals ("Bayesian credible regions") for each manifesto. The top part of Table 1 presents estimates for selected parties from the UK, the United States, and Germany. In each context, the location of the parties has high face validity, ordered in a manner which would accord with any informed observer's understanding of party politics in each context. With the move to the centre of Blair's New Labour in 1997, for instance, the measure tracks Labour's move relative to the more traditional leftist position of Labour in 1987.

Figure 1 plots the party positions inferred by the negative binomial IRT model against measures based on expert surveys.[11] The graph shows the overall pattern, and results for three broad subsets of the sample: Western Europe, Eastern Europe, and the Pacific plus North America. (No expert surveys were available for other regions, so these are not reported.) Overall, there is a good linear

[8] We aggregated the extended four-digit category codes into their respective three-digit "parent" categories.

[9] To be precise, starting from the 2014b edition of the data (Volkens et al., 2014) we use post-1972 manifestos from countries with a Polity-IV rating of at least seven, or a Freedom House rating of at least nine (when no Polity-IV rating was available). We dropped duplicate data entries (manifestos associated with several parties) and cases based on estimates and those with missing document length information.

[10] The Gelman-Rubin-statistic (Gelman and Rubin, 1992) does not show any signs of non-convergence. See Appendix E.

[11] Expert data are from Benoit and Laver (2006); Steenbergen and Marks (2007); Hooghe et al. (2010); Bakker et al. (2015). We match the expert placement to the temporally closest manifesto, if a document is available within three years before or after the survey date.

Model	Poisson		Negative Binomial	
	Mean	*95% BCR*	*Mean*	*95% BCR*
Illustrative Estimates of θ_i				
UK 1987				
Labour	-0.56	[-0.78,-0.34]	-0.57	[-1.30,0.13]
Liberal Democrats	-0.28	[-0.43,-0.13]	-0.05	[-0.73,0.62]
Conservatives	1.12	[1.00,1.23]	1.27	[0.52,2.04]
UK 1997				
Labour	0.25	[0.10,0.40]	-0.12	[-0.87,0.62]
Liberal Democrats	-0.18	[-0.35,-0.01]	0.07	[-0.64,0.82]
Conservatives	0.91	[0.79,1.03]	0.86	[0.08,1.71]
United States 2012				
Democrats	-0.11	[-0.24,0.01]	-0.15	[-0.78,0.43]
Republicans	1.55	[1.46,1.64]	1.03	[0.45,1.63]
Germany 2009				
Left	-1.92	[-2.05,-1.78]	-2.21	[-3.04,-1.33]
Greens	-1.37	[-1.45,-1.27]	-1.31	[-2.09,-0.56]
SPD	-0.90	[-1.01,-0.79]	-0.90	[-1.64,-0.14]
CDU/CSU	0.36	[0.27,0.46]	0.40	[-0.28,1.17]
FDP	0.36	[0.28,0.45]	0.35	[-0.25,0.97]
Chains		2		2
Warmup		500		500
Samples per chain (after thinning by)		750 (2)		750 (2)
N manifestos		2288		2288

Table 1: Examples of estimated left-right positions $\hat{\theta}_i$ (basic model based on core 56 CMP category counts)

fit ($r = .75$). Our measure using IRT scaling of *all* policy categories thus corresponds well to what experts judged to be the left-right dimension, without having to use any *ex ante* expert judgment as to which policy categories should be associated with this dimension.

The best match of our model estimates to the expert survey scores is observed in the Pacific and North America, at r=.89 and r=.82 respectively. The correlation with the expert survey estimates was lowest in Eastern Europe (r=.55), indicating that the same patterns that fit overall did not fit particularly well in Eastern Europe, and that manifestos may also be a less reliable data source in that region. This finding is consistent with earlier research findings that the content of left-right policy is different in post-communist settings (Benoit and Laver, 2006; Mölder, 2013). While we do not investigate this further here, our results underscore that the assumption of identical and identically weighted ingredients set by the fixed-index approach is not only too strong, but

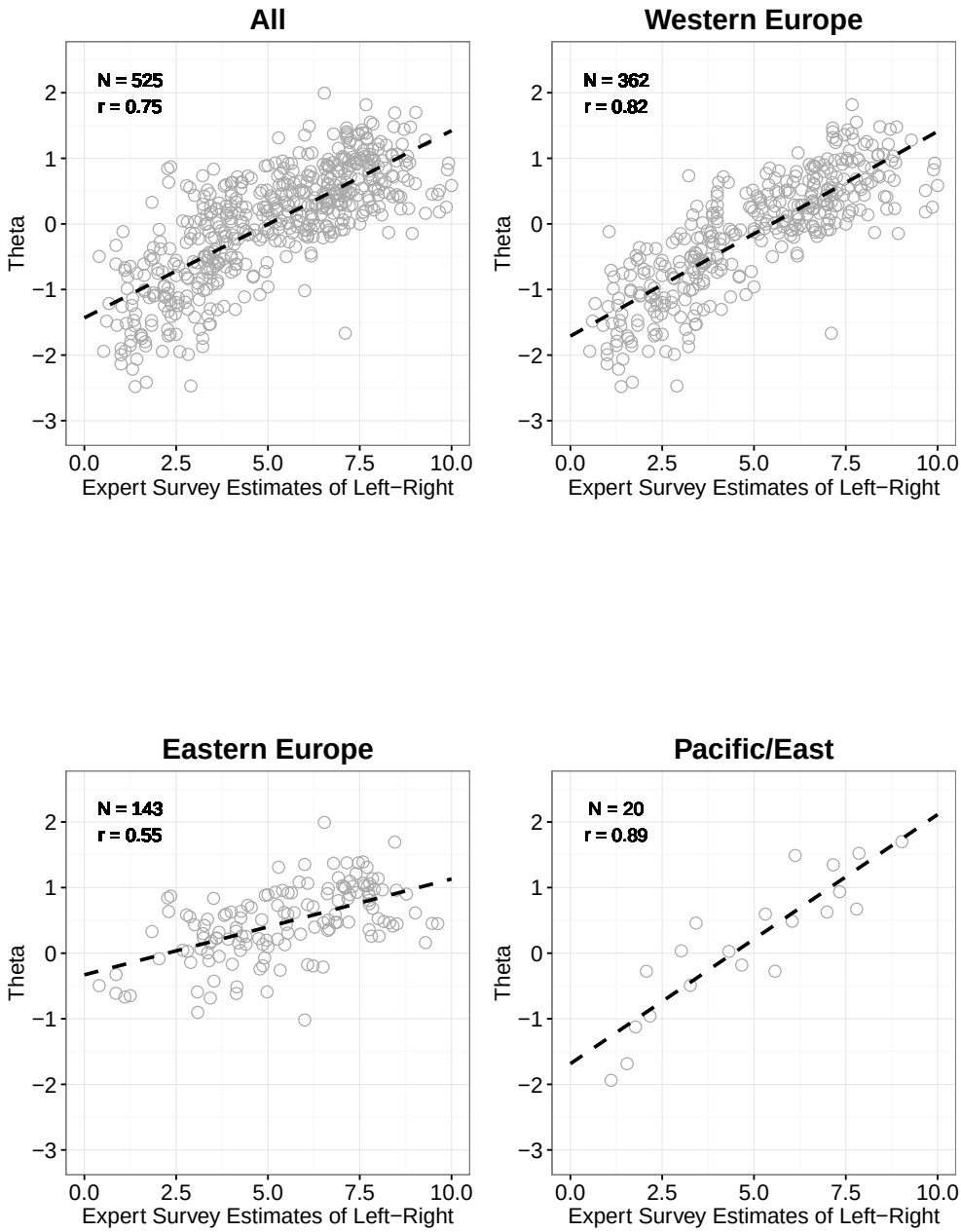


Figure 1: Comparison of $\hat{\theta}_i$ to expert survey estimates for the post-1972 sample (from Table 1).

impossible to test directly without using a more inductive, data-driven approach such as our IRT model.

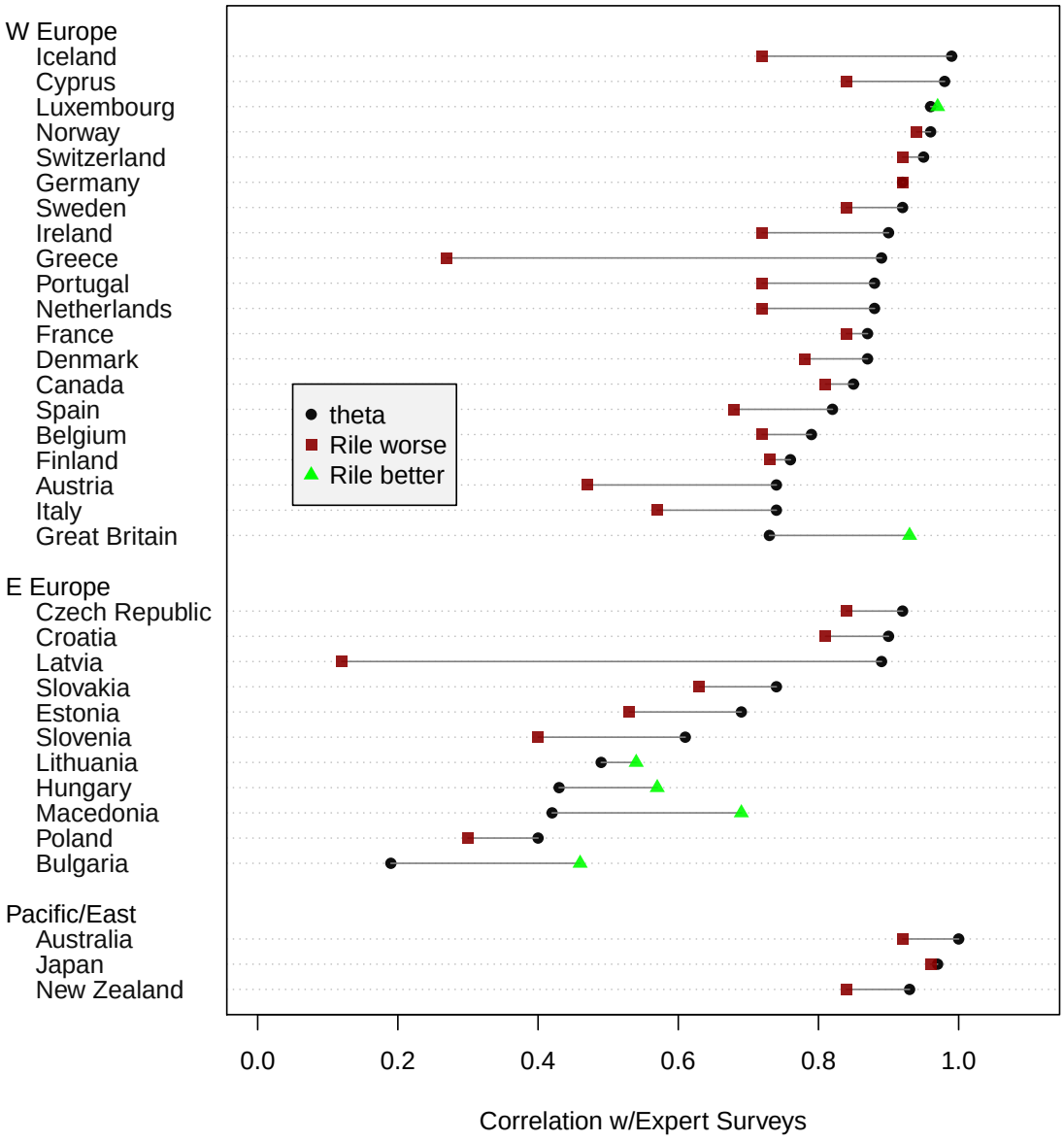


Figure 2: Comparison of IRT versus Rile correlations with expert survey estimates of left-right policy, by country.

Not only do the IRT estimates of left-right position perform well judged against independent expert surveys, but also they generally perform better than the very widely used measure distributed with the Manifesto Project data, *based on the same hand-coded manifesto content*. In Figure 2 we compare, by country, the correlations with expert survey scores of our IRT estimates of left-

right policy, to the correlations of the expert survey scores with the Rile index provided with the Manifesto Project's dataset. When the Rile score better matched variation in the expert scores, we plotted it as a green triangle, otherwise, we plotted it as a red square. Except for Great Britain and the near tie of Luxembourg, the only countries for which Rile's correlation with the expert scores was higher were four countries of Eastern Europe (Lithuania, Hungary, Macedonia, and Bulgaria), where the fit was among the poorest of all countries for both scales and for which the point estimates belie a large confidence interval due to the noise from estimating the positions of these cases from manifesto content. In all other country cases, our estimate provided better performance than the Rile measure.

A better estimate of uncertainty: Modelling policy shifts

Because the model's parameters are stochastic and inferred by sampling from the posterior, it is possible to estimate uncertainty over the left-right positional parameters $\hat{\theta}_i$ through simulating draws from the posterior distribution. A key feature of our approach, the ability to obtain uncertainty measures directly from posterior draws is something not provided directly by other methods. Estimating uncertainty from manifesto data, Benoit, Laver and Mikhaylov (2009) applied non-parametric simulation by assuming that the category frequencies were drawn according to a multinomial distribution and bootstrapped the category counts on this basis to compute a standard error for each Y_{ij} in addition to compound categories such as the additive Rile index. Our method allows a more direct and flexible approach in which the data-generating process may be characterized according to a more realistic model than the Poisson/multinomial, in our case through a negative binomial distribution. This makes it possible to implement parametric approaches to uncertainty, facilitating the sorts of comparisons against other benchmarks to judge which model is most appropriate, such as the comparisons to human coding of uncertainty in Lowe and Benoit (2013).

Using the results of our model estimated on the full sample, we can contrast our results to those of Benoit, Laver and Mikhaylov (2009, Table 1), who reported that using their non-parametric

Statistically Significant Change?	BLM (2009) Shifts	%	Poisson $\Delta\hat{\theta}_i$ Shifts	%	Neg Binomial $\Delta\hat{\theta}_i$ Shifts	%
No	859	54.2%	836	52.7%	1,533	96.7%
Yes	726	45.8%	749	47.3%	52	3.3%
Total adjacent	1585	100%	1585	100%	1585	100%
Non-adjacent	703	–	703	–	703	–
Total	2,288		2,288		2,288	

Table 2: Comparative over-time mapping of policy movement on Left-Right measure, taking into account the statistical significance of shifts – comparing change in IRT estimates to Benoit, Laver and Mikhaylov (2009) estimates from non-parametric bootstrapping.

bootstrapping procedure, only 38% of parties' observed left-right movements could be declared real rather than the result of stochastic features generating the text from underlying policy positions. Applying their method to our sample, we found that around 46% of parties' left-right movement from one election to the next could be considered real change not attributable to noise from the inherently stochastic process of generating observed text mentions from a fixed policy position. The Poisson model estimates match this rate quite well, at about 47%, not surprising given that the underlying stochastic process (an equivalent multinomial distribution) underlies both methods. For the the negative binomial model, with its much larger confidence intervals on the estimates of the policy shifts, only 3.3% of these could be deemed to be different from no change based on their 95% confidence intervals. While we offer these results more as a demonstration of how to measure policy shifts based on more direct models of policy measurements, rather than a definitive finding on the issue of the volatility of party policy, our findings suggest that in this context the real rate of policy change lies somewhere in between, and is overstated using the Benoit, Laver and Mikhaylov (2009) approach. By contrast, the IRT approach provides an estimate of real policy movement based not only on a more complete model that includes the baseline noisiness of stochastic text, but also incorporating the full information as to how the use of policy statements reflects the dimension of left-right politics based on all of the patterns found in the dataset.

The difference between the negative binomial and Poisson model results, furthermore, illustrates the significant consequences of different model assumptions. The restrictive and unrealistic

variance assumption of the Poisson model can lead to significantly underestimated parameter uncertainty in $\hat{\theta}_i$ (Lowe and Benoit, 2013). That the negative binomial model is the more appropriate choice for the data at hand is also the conclusion of a comparison of model fit across the two specifications. As Cameron and Trivedi (2013, 189) discuss, the Pearson statistic P points to overdispersion (or possibly misspecification of the mean) if it exceeds N minus the number of estimated parameters. In our application, $N = 128,128$ which in itself is far smaller than $P = 2,045,622$ of the Poisson model. The negative binomial model at least alleviates the problem, with $P = 682,993$. The negative binomial also outperforms the Poisson specification when it comes to accounting for zero counts in the data. The observed rate of 48.5% is closely matched by the expected rate of zeros by the former (46.3%), but drastically underpredicted by the latter (26.8%).

The policy content of the left-right dimension

In this section we examine the association between the policy categories and left-right dimension. The IRT approach allows all items to be incorporated into measuring left-right policy, while letting the scaling provide an appropriate weight for each specific policy component. We demonstrate how this procedure, in addition to producing better left-right estimates, also provides interesting insights into the substantive content of the main axis of party competition.

We also compare the results of our scaling to the original categorization of the Manifesto Project policy categories into left, right, and neutral by the "Rile" index (see Laver and Budge, 1992, 25–30). With the goal of locating parties in a single dimension, Laver and Budge applied country-by-country factor analysis to codings from ten Western European countries[12] in the period 1945–1985. Inspecting each set of results for face validity and making some decisions to combine or drop some categories based on their loadings, the result was the first version of the Rile scale, now the most widely used quantity in the Manifesto Project dataset. In fitting the manifesto data to the single dimension of difference that appeared meaningfully to differentiate parties, they em-

[12]These were: Austria, Belgium, Great Britain, Denmark, France, Ireland, Italy, Luxembourg, Netherlands, and Sweden.

phasize that this process was "based solely on the intrinsic plausibility and coherence of the sets of issues that define the underlying policy dimension" (25). Presumably, this is why categories such as "Political authority: Positive (305)" are considered "right-wing": because in the sample examined, this was the pattern of their association. Using our method that automatically includes and estimates the weighted contribution of each policy input to the left-right dimension, we can compare our scaling results to those of the original Manifesto Project's categorization.

Applying our one-dimensional IRT model to the coded policy category counts, pooled across countries, we observe a good association of our scaled discrimination parameter estimates values to the left and right categories of the Rile index. Figure 3 plots the item discrimination parameters $\hat{\lambda}_j$ for each policy category, fit to post-1972 coded manifestos. As can be seen by the positioning of the parameter estimates relative to the dividing line (which represents the mean λ across items), most of the the Rile left categories are indeed associated with left positions, and most of the right categories with right positions. Some are far less informative than others, however, and are not estimated as corresponding to left or right, including Political Authority: Positive (305), Social Harmony: Positive (606), Constitutionalism: Positive (203), Freedom and Human Rights: Positive (201), Education Expansion: Positive (506), and Protectionism: Positive (406). Some of these categories, such as Political Authority: Positive, are known to cause problems in indexing left-right positions by biasing leftist parties to the right, such as the Italian Communist Party, a far-left party erroneously scored as far right in the 2000s because of its high proportion of statements coded Political Authority: Positive (see Benoit and Laver, 2007, 97–98).

Among the 30 policy categories excluded from the Rile index, furthermore, we see several that are very strongly associated with left-right policy positions: Marxist Analysis: Positive (415) and National Way of Life: Negative (602) on the left, for instance, and Labour Groups: Negative (702) and Multiculturalism: Negative (608) on the right. *For this sample*, we see from Figure 3 that there are numerous categories not used to estimate left-right context that could have contributed productively to the measurement of party positions along this single dimension. By modelling category counts directly as a function of the responsiveness of the party's manifesto content to their

underlying latent position θ_i, the IRT approach uses all available information. Instead of requiring a (potentially arbitrary or controversial) list of "in" and "out" categories, the IRT approach uses them all and determines their relative contributions from the data.

Using IRT to Estimate Multiple Dimensions of Policy

For many applications, researchers are interested in measuring policy positions in more than one dimension. For instance, many party systems can be adequately characterized by competition along two dimensions, an economic and a "social" one related to moral questions such as abortion and homosexuality (Laver and Hunt, 1992; Hooghe, Marks and Wilson, 2002). One approach for obtaining policy positions in multiple dimensions is to assign categories *a priori* to the different policy fields/dimensions, and conduct the scaling on a dimension-by-dimension basis (Elff, 2013). Proceeding this way is not recommended, however, when we have reasons to believe that there are categories which are linked to more than one of the latent dimensions.

For this purpose, we can extend the model to two dimensions d. We now model the expected counts as

$$\log(\mu_{ij}) = \alpha_i + \zeta_j + \lambda_{1j}\theta_{1i} + \lambda_{2j}\theta_{2i} \tag{6}$$

in which we estimate two positions θ_{di} and two sets of discrimination parameters λ_{dj} (for two dimensions $d \in \{1,2\}$). The challenge, as with any multi-dimensional IRT model, is to impose appropriate constraints in order to reach statistical identification of all the parameters (Jackman, 2001; Rivers, 2003). Our solution is detailed in Appendix D. For the two-dimensional model, we set $\phi^{-1} = 0$, i.e. choosing a Poisson likelihood *a priori*.[13]

Figure 4 shows the λ_{dj} parameters of all items that were selected to contribute to the respective dimension (economic in the left panel of the graph, and "social" in the right one). Again, we can see

[13]Relaxing this assumption led to convergence problems. The way the constraints are set requires that both the variances of the policy positions and the variances of the discrimination parameters are inferred from the data. It seems that this already tricky inference problem is exacerbated when adding an error variance in form of the overdispersion parameter. For an assessment of convergence of the two-dimensional Poisson model see the respective section of Appendix E.

considerable variation in the extent to which the categories discriminate. For the purely economic items, "Labour Groups: negative" and "Education Limitation" are most rightist, whereas "Marxist Analysis: positive" "Nationalisation: positive" are the most leftist ones. Considering the "social dimension," the results also correspond to the expectations, with the contrasting pairs related to "National Way of Life" and "Traditional Morality" to be found at the opposing ends of the scale.

Particularly interesting insights can be gained from the results for the items that were allowed to contribute to both dimensions. We may expect that most of these items are either predominantly associated with only one dimension, or that economically rightist (leftist) discrimination parameters tend to go along with socially conservative (liberal) ones. Indeed, we find such categories. For example, positive references to the military represent positions on the right side on both dimensions. And "Environmental protection: positive" is one category whose usage in manifestos is mostly explained by a party's position in one of the dimensions, in this case notably the second, "social" rather than the first, economic dimension.

In addition, there are a number of items which follow a more complex pattern. The prime example is the "Political Authority: positive" category (which is one of the rightist items in the Manifesto Project's fixed "rile" scale). In the second dimension, its λ value is indeed positive, implying "socially" conservative positions. In the economic realm, however, the category is associated with the left end of the political spectrum. This result makes perfect sense, however, as political authority in the economic context corresponds to a more active role of the state in economic regulation. The extra analytical leverage we receive from the two-dimensional solution is also shown by the two categories referring to European integration. Pro-integration statements reflect positions that are economically rightist, or socially liberal. Conversely, negative remarks about European integration can result from views that are economically leftist, or socially conservative.

Next, we compare the Manifesto Project-based party positions on both dimensions to those from expert surveys as above (Figure 5, using the "taxes vs. spending" and "social policy/social lifestyle" party ratings. On the economic dimension, there is a quite strong positive association,

reflected in a correlation of $r = .73$. With regard to the social dimension, there is also a positive correspondence, although the observations are much more scattered around the regression line. Note, however, that we let a broad range of categories (rather than just a few related explicitly to morality issues) contribute to the second dimension, and that manifestos in general do not necessarily cover "social" issues as extensively as they do economic ones. This and the fact that this dimension is inherently more eclectic (Benoit and Laver, 2006) will make it harder to correctly place the parties on the second dimension.

Using alternative items: The Comparative Agendas Project

Thus far, we have fit our model to data from the Manifesto Project. To illustrate the flexibility of our approach, we also apply it to data from an entirely different manually coded content analytic scheme: the Comparative Policy Agendas Project (CAP) (Walgrave and De Swert, 2007; Baumgartner, Green-Pedersen and Jones, 2008). The CAP aims to identify the topic focus of policy documents, media coverage and political events such as cabinet meetings (Baumgartner, Green-Pedersen and Jones, 2008). Our analysis focuses on a set of Belgian political documents coded by the CAP — a sub-project described by Walgrave, Varone and Dumont (2006, 1025) as follows: "These agendas were encoded in their entirety in order to compute relative issue attention (saliency) in percentage of all issues appearing on these agendas." In the Belgian case, the coded documents also include party manifestos.

The general coding approach used for the Belgian manifestos resembles that of the Manifesto Project, as "(semi)sentences" (Walgrave and De Swert, 2007, 42) were hand-coded into one of 137 (in some cases 143) categories. An important difference between the CAP and the CMP, however, is that the CAP categories are exclusively based on content and thus not intended to be positional. The category scheme spans a wide array of very detailed topics including issues of political organisation (e.g. "State reform, political power and intercommunity conflicts," code 012), economic matters (e.g. "trade policy," code 148), social questions (e.g. "migration and integration

of immigrants," code 173), and environmental topics (e.g. "water," code 294). Additional various categories (to name a few) encompass "conception and contraception" (code 172) and "fishing" (code 318).

These data pose a challenge for measuring party positions, since they are not designed as positional items. Constructing a fixed index to measure left-right positions, in other words, would be very difficult, since we have few prior theoretical expectations as to which of the numerous policy content categories actually convey information in terms of left-right positioning. By contrast, our inductive scaling approach allows us to measure this latent dimension and estimate the contribution of each item to the scale. To model this, we apply the negative binomial scaling model to the 37 Belgian manifestos from 1991-2003 that were coded by the Belgian CAP team (2003 data include Flemish-speaking parties only).[14] Here we focus on the validation of the results for the set of cases that could be matched with temporally close expert survey estimates (positions for all CAP-coded documents are shown in Appendix F). Figure 6 plots the results, indicating high face validity. The Green and Socialist parties appear on the left side of the political spectrum, while the far-right *Vlaams Belang* is anchored on the extreme right, with the Christian Democrats, and the (both economically and "socially") Liberals as well as the regionalist parties in the center. The regression line indicates the close correspondence to independently measured left-right positions from expert surveys ($r = .87$).

One advantage of the IRT model is that we also learn something about the items and thus about the content of political contestation in the context to which it was applied. First, it is interesting to note that even with the non-positional CAP data and allowing for item-level overdispersion, we find that approximately one third of the 137 items discriminate on the left-right dimension (judged on the basis of whether 90% of the posterior distribution of a λ_j are to the left or right of zero). The three most leftist items are "Environmental problems with energy," "Forestry," and "Biohazard," while the three most rightist categories are "State Reform," "Migration," and "Asylum." This suggests that the category codings in the Belgian CAP represent issues that practically work as

[14] Probably due to the smaller sample, convergence is more difficult to achieve in this case. The model is run for a warmup period of 8,000 iterations, followed by another 2,000 for inference, with a thinning factor of four.

valence issues, where differences in emphasis are linked to parties' positions, and the scaling model picks up this information.

Taken together, using the IRT model on the CAP data provides strong evidence that the important thing about scaling positions is not so much the input in terms of particular items, as having a variety of viable candidate items coupled with a procedure that can estimate their contribution to the underlying latent policy dimension. As long the items contain *some* information that is indicative of policy differences, an appropriate scaling procedure can recover the positions on the latent variable even if a large part of the input data differentiates very little between parties.

Conclusion

We have proposed a measurement approach for left-right policy positioning based on item response theory, permitting the estimation of ideology as a latent "ability" variable, and for the contribution of each element of measured policy to act as "items" for which additional mentions are generated depending on their relationship to the underlying ideology variable. Not only does this approach allow a better estimation of uncertainty about these parameters than other approaches, but also it permits more realistic modelling of the stochastic process that generates the observed counts of specific categories of statements as a nominal response framework. Our approach was to model counts as a negative binomial process, estimating an additional variance component for each category, providing a better fit and more conservative error estimates for the resulting political quantities.

Our application of IRT to hand-coded sentences also underscores the great value of manually coding manifesto content. Manual content analysis schemes are designed to maximize *validity*, by using expert human judgment. We have shown how unsupervised scaling methods originally designed for including large numbers of features whose discriminatory power is unknown, can be adapted to scale low-level quantities of interest from manually coded content without precommitment to fixed schemes for measuring these lower dimensions. This frees content analytic

schemes to focus on their areas of strength, which is coding specific, high-level dimensions of content, while keeping open the maximum set of options at the design stage of the project. Our approach therefore combines the best of two worlds: valid measurements of specific policy statements performed by expert human coders, with a valid and flexible procedure that infers lower-dimensional policy measurements from the *full* set of coded statements.

Our replication of the left-right policy positions for Belgium using a completely different set of items, from the Comparative Agenda Project's coding of the same party documents, drives this essential point home. Just as in the classical testing framework from which IRT was developed, it is not the exact questions which are of interest, but rather the manner in which patterns of response inform us about the latent quantities of interest. These underlying quantities remain the same for individuals, while tests and their questions differ. Using our inductive approach, we focus directly on the essential quantity—latent ideological positions. We can make the most from the items, without having to make controversial theoretical assumptions about their individual contributions.

Here we have presented basic one- and two-dimensional models, but our framework is flexible enough to permit more sophisticated parameterizations. Exploiting the ability to add hierarchically dependent parameters in a Bayesian model, for instance, it is possible to estimate item parameters that vary across contexts. This would be analogous to a multi-level IRT model, with random effects partitioning the items (see Fox, 2010, section 6.3). In addition, the flexibility of the Bayesian IRT approach permits a more complete model of the political process, potentially including covariates to hierarchically model the parameters of the core model. Not only does this offer the potential to incorporate additional information as variables to improve the model fit in specific contexts, but also allows for the possibility of testing substantive explanations of parties' policy shifts directly within the measurement model. Using the flexibility of IRT scaling by treating coded categories as items to which the manifesto statements respond, extensions will permit the direct modeling of a variety of interesting political questions without having to decide *a priori* how the manifesto content relates to the quantities being estimated.

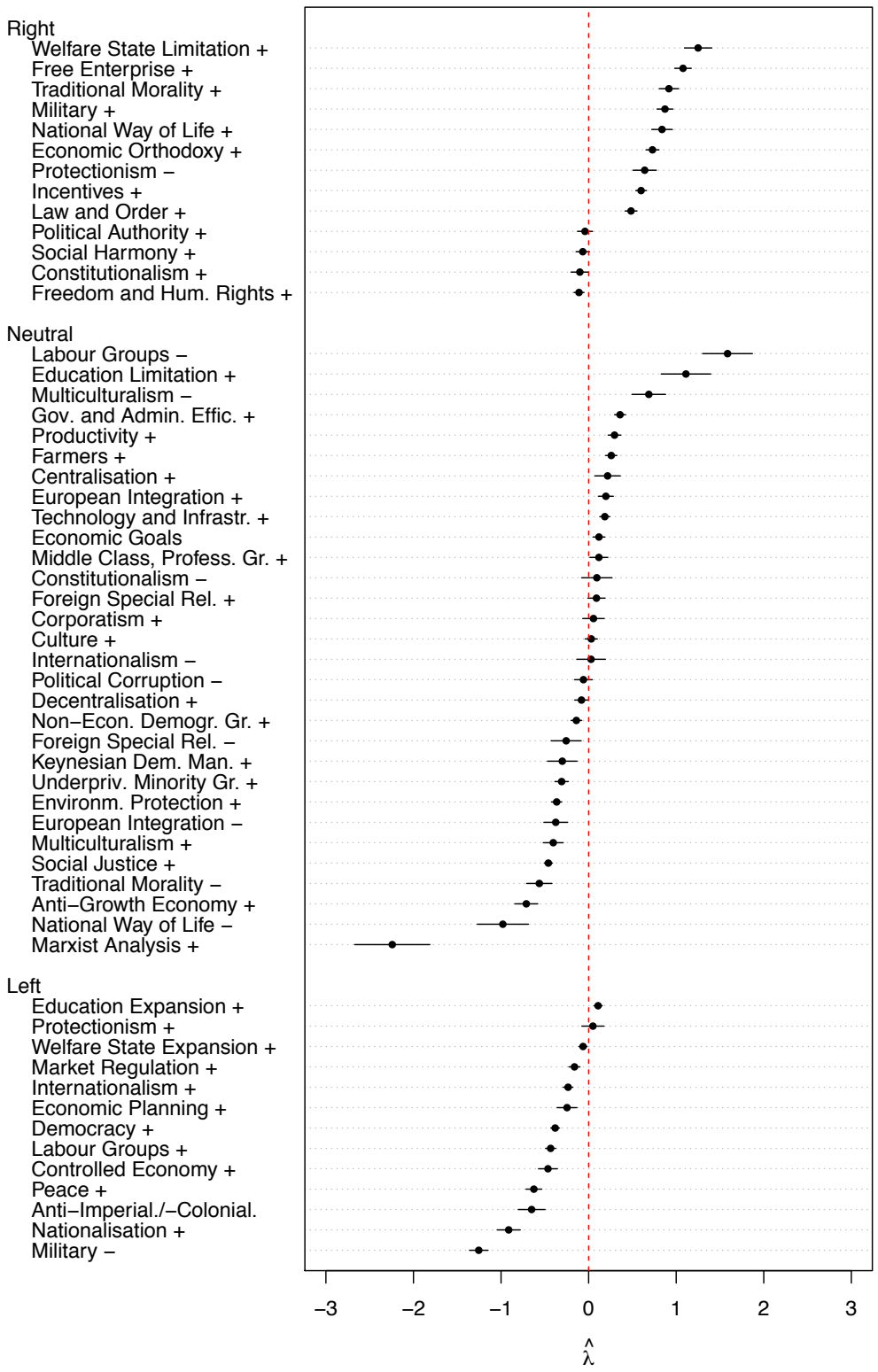
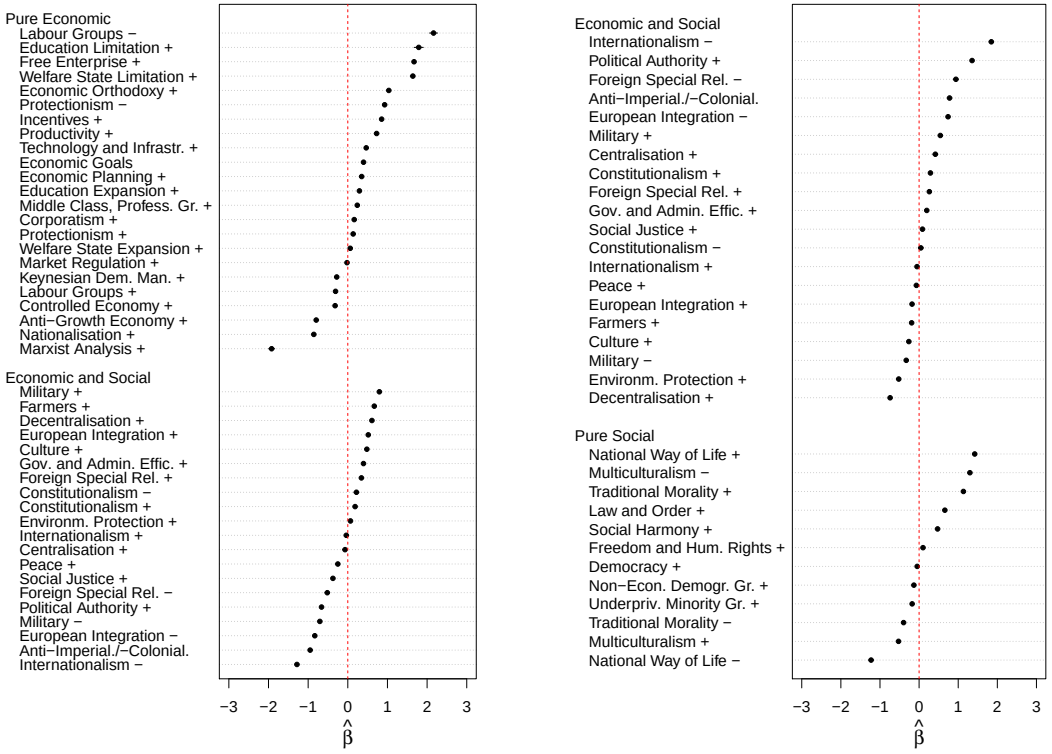


Figure 3: Item discrimination parameter estimates ($\hat{\lambda}_j$) by Rile category, for the post-1972 sample.



(a) Economic Dimension

(b) Social Dimension

Figure 4: Item discrimination parameter estimates $\hat{\lambda}_{dj}$ from the two-dimensional model fit with subsets of categories selected for possible economic and social policy content.

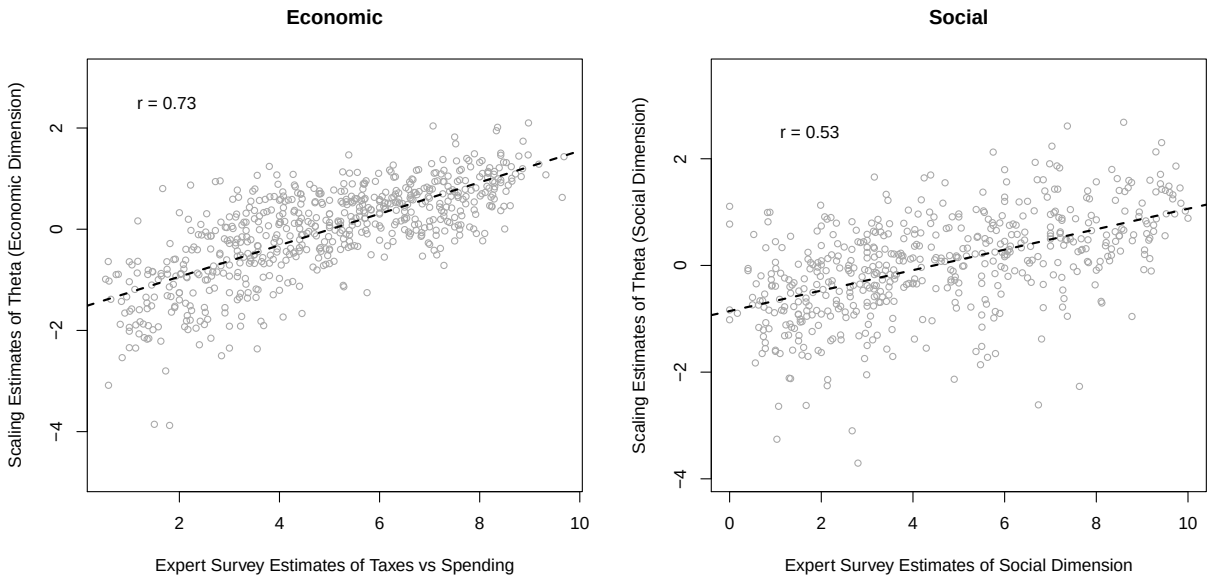


Figure 5: Correlations of two-dimensional model estimates for $\hat{\theta}_{\text{econ } i}$ and $\hat{\theta}_{\text{social } i}$ with expert survey estimates.

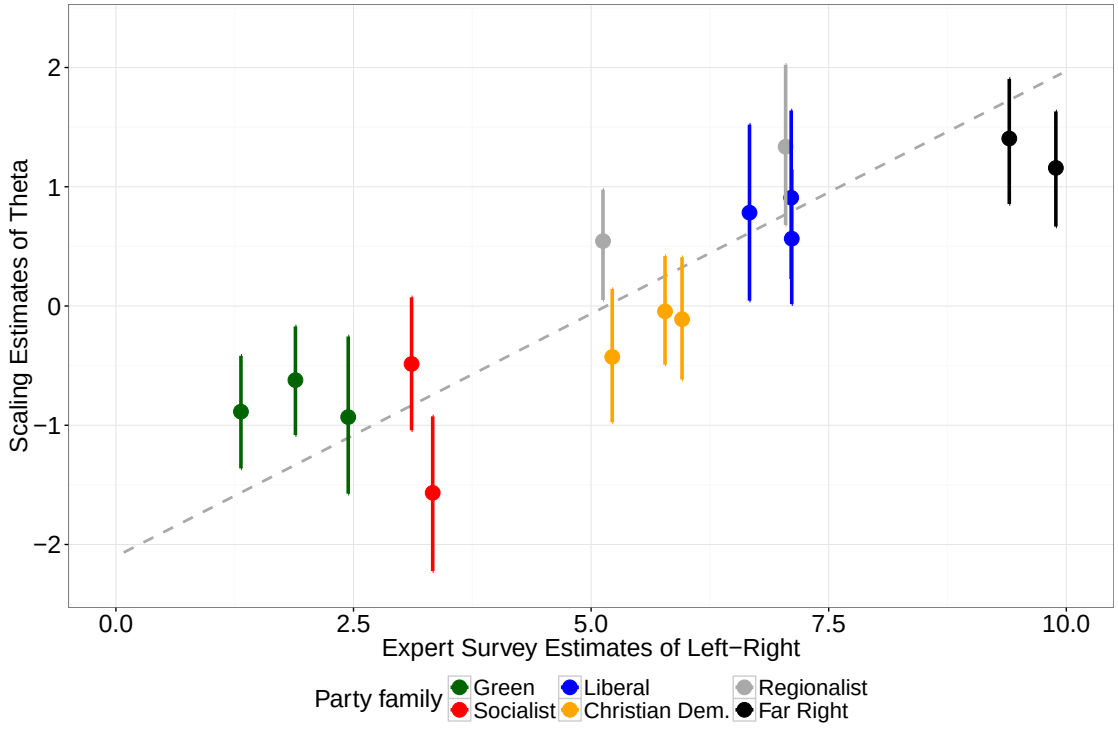


Figure 6: Left-right estimates for Belgian Parties from the Comparative Agendas Project dataset (1999–2003, with 90% BCIs) compared to expert survey placements. Shown are (from left to right): Agalev 2003, Ecolo 1999, Agalev 1999,PS 1999, SP 1999,VU 1999,PSC 1999, CVP 1999, CD&V 2003, PRL-FDF 1999, N-VA 2003, VLD 2003 and 1999, VB 2003 and 1999. Dashed grey line represents linear fit.

References

Agresti, Alan. 2013. *Categorical data analysis*. Hoboken: Wiley.

Albright, Jeremy J. 2008. "Bayesian Estimates of Party Left-Right Scores.". Unpublished paper, Indiana University and University of Michigan.
URL: *http://www.polmeth.wustl.edu/files/polmeth/polmeth.pdf*

Baker, F. B. 1992. *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bakker, Ryan. 2009. "Re-measuring left–right: A comparison of SEM and Bayesian measurement models for extracting left–right party placements." *Electoral Studies* 28(3):413–421.

Bakker, Ryan, Catherine de Vries, Erica E. Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen and Milada Vachudova. 2015. "Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999-2010." *Party Politics* 21(1):143–152.

Baumgartner, F. R., C. Green-Pedersen and B. D. Jones. 2008. *Comparative Studies of Policy Agendas*. Routledge.

Benoit, Kenneth and Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Benoit, Kenneth and Michael Laver. 2007. "Estimating Party Policy Positions: Comparing Expert Surveys and Hand Coded Content Analysis." *Electoral Studies* 26(1):90–107.

Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2):495–513.

Bobbio, Norberto. 1996. *Left and right: The significance of a political distinction*. University of Chicago Press.

Bock, R Darrell. 1972. "Estimating item parameters and latent ability when responses are scored in two or more nominal categories." *Psychometrika* 37(1):29–51.

Budge, Ian, David Robertson and Derek Hearl. 1987. *Ideology, Strategy and Party Change: Spa-*

tial Analyses of Post-War Election Programmes in 19 Democracies. Cambridge: Cambridge University Press.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.

Budge, Ian and Thomas Meyer. 2013. Understanding and Validating the Left-Right Scale (RILE). In *Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts*, ed. Andrea Volkens, Judith Bara, Ian Budge, Michael D. McDonald and Hans-Dieter Klingemann. Oxford University Press pp. 85–106.

Cameron, A Colin and Pravin K Trivedi. 1986. "Econometric models based on count data. Comparisons and applications of some estimators and tests." *Journal of Applied Econometrics* 1(1):29–53.

Cameron, A Colin and Pravin K Trivedi. 2013. *Regression analysis of count data*. 2 ed. Cambridge: Cambridge University press.

Carlyle, Thomas. 1888. *The French revolution: a history*. Chapman & Hall, Ld.

Dolezal, Martin, Laurenz Ennser-Jedenastik, Wolfgang C Müller and Anna Katharina Winkler. 2013. "How parties compete for votes: A test of saliency theory." *European Journal of Political Research* 53(1):57–76.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65(2):135–150.

Elff, Martin. 2013. "A dynamic state-space model of coded political texts." *Political Analysis* 21(2):217–232.

Fox, J P. 2010. *Bayesian item response modeling: Theory and applications*. Heidelberg: Springer.

Franzmann, Simon and André Kaiser. 2006. "Locating Political Parties in Policy Space: A Reanalysis of Party Manifesto Data." *Party Politics* 12(2):163–188.

Franzmann, Simon T. 2015. "Towards a real comparison of left-right indices. A comment on Jahn." *Party Politics* 21(5):821–828.

Fuchs, Dieter and Hans-Dieter Klingemann. 1990. The Left-Right Schema. In *Continuities in Political Action: A Longitudinal Study of Political Orientations in Three Western Democracies*, ed. M. Kent Jennings and Jan W. van Deth. Berlin: Walter de Gruyter pp. 203–234.

Gabel, M. J. and J. D. Huber. 2000. "Putting parties in their place: Inferring party left-right ideological positions from party manifestos data." *American Journal of Political Science* 44(1):94–103.

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Sciences* 7(4):457–472.

Gooding, Philip. 2011. "Consumer Prices Index and Retail Prices Index: the 2011 basket of goods and services." *Economic & Labour Market Review* 5(4):96–107.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Hoffman, Matthew D. and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15(Apr):1593–1623.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine De Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen and Milada Vachudova. 2010. "Reliability and validity of the 2002 and 2006 Chapel Hill expert surveys on party positioning." *European Journal of Political Research* 49(5):687–703.

Hooghe, Lisbet, Gary Marks and Carole J. Wilson. 2002. "Does left/right structure party positions on European integration?" *Comparative Political Studies* 35(8):965–989.

Inglehart, Ronald. 1984. The Changing Structure of Political Cleavages in Western Society. In *Electoral Change, Realignment and Dealignment in Advanced Industrial Democracies*, ed. Russell Dalton et. al. Princeton: Princeton University Press pp. 25–69.

Jackman, S. 2001. "Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference and model checking." *Political Analysis* 9(3):227–241.

Jahn, Detlef. 2011. "Conceptualizing Left and Right in comparative politics Towards a deductive approach." *Party Politics* 17(6):745–765.

Jahn, Detlef. 2014. "What is left and right in comparative politics? A response to Simon Franzmann." *Party Politics* 20(2):297–301.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara and Ian Budge. 2006. *Mapping Policy Preferences II. Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD 1990-2003*. Oxford: Oxford University Press.

König, Thomas, Moritz Marbach and Moritz Osnabrügge. 2013. "Estimating Party Positions across Countries and Time—A Dynamic Latent Variable Model for Manifesto Data." *Political Analysis* 21(4):468–491.

Laver, Michael and Ben W. Hunt. 1992. *Policy and Party Competition*. New York: Routledge.

Laver, Michael and Ian Budge. 1992. Measuring policy distances and modeling coalition formation. In *Party policy and government coalitions*. New York: St. Martin's Press.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Estimating the policy positions of political actors using words as data." *American Political Science Review* 97(2):311–331.

Lo, James, Sven-Oliver Proksch and Jonathan B Slapin. 2016. "Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos." *British Journal of Political Science* 46(3):591–610.

Lowe, Will. 2016. "Scaling things we can count." Unpublished paper, Princeton University.
URL: *http://dl.conjugateprior.org/preprints/scaling-things-we-can-count.pdf*

Lowe, William and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Mölder, Martin. 2013. "The validity of the RILE left-right index as a measure of party policy." *Party Politics* 22(1):37–48.

Ostini, Remo and Michael L Nering. 2009. *Polytomous Item Response Theory Models*. Sage.

Pemstein, Daniel, Stephen A Meserve and James Melton. 2010. "Democratic compromise: A latent variable analysis of ten measures of regime type." *Political Analysis* 18(4):426–449.

Proksch, S.-O. and J. B. Slapin. 2010. "Position taking in the European Parliament speeches." *British Journal of Political Science* 40(3):587–611.

Reckase, Mark D. 1997. "The past and future of multidimensional item response theory." *Applied Psychological Measurement* 21(1):25–36.

Rivers, Douglas. 2003. "Identification of multidimensional spatial voting models.". Unpublished manuscript, Stanford University.

URL: *http://polmeth.wustl.edu/files/polmeth/river03.pdf*

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Stan Development Team. 2015. "Stan: A C++ Library for Probability and Sampling, Version 2.6.0.".

URL: *http://mc-stan.org/*

Steenbergen, Marco R and Gary Marks. 2007. "Evaluating expert judgments." *European Journal of Political Research* 46(3):347–366.

Stokes, Donald E. 1963. "Spatial Models of Party Competition." *American Political Science Review* 57(2):368–377.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52(1):201–217.

Volkens, Andrea, Judith Bara, Ian Budge and Michael D McDonald. 2013. *Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts*. Oxford University Press.

Volkens, Andrea, Pola Lehmann, Nicolas Merz, Sven Regel and Annika Werner. 2014. "The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2014b.".

URL: *https://manifestoproject.wzb.eu/datasets*

Walgrave, Stefaan, Frédéric Varone and Patrick Dumont. 2006. "Policy with or without parties? A comparative analysis of policy priorities and policy change in Belgium, 1991 to 2000." *Journal of European Public Policy* 13(7):1021–1038.

Walgrave, Stefaan and Knut De Swert. 2007. "Where does issue ownership come from? From the party or from the media? Issue-party identifications in Belgium, 1991-2005." *The Harvard International Journal of Press/Politics* 12(1):37–67.

Supplementary Appendixes

A Equivalence to Slapin and Proksch's "Wordfish" model

Eq. 5 is equivalent to Slapin and Proksch (2008)'s unidimensional Poisson scaling model of document positions θ_i for a document-term matrix, expressed as:

$$\log(\mu_{ij}) = \alpha_i + \psi_j + \lambda_j \theta_i \tag{7}$$

These equivalencies are mapped in Table 3.

Our IRT Model		Slapin and Proksch (2008)	
Qty.	Interpretation	Qty.	Interpretation
θ_i	latent "ability"	θ_i	Ideological position
α_i	denominator for multinomial equivalence	α_i	Fixed document length effect
ζ_j	Conditional "difficulty" parameter	ψ_j	Fixed word effect
λ_j	Item discrimination parameter	β_j	Word sensitivity to position θ_i

Table 3: Equivalencies between IRT Model and "Wordfish"

B Statistical identification of the one-dimensional model

The model in Eq 5 requires additional constraints for identification, as there are five fundamental indeterminacies. First, shifts in the mean of the ζ_j can be compensated by shifts in the mean of the α_i – put differently, it is impossible to infer whether all categories are jointly more (less) frequent or whether the manifestos are longer (shorter) overall. So the location of the ζ_j needs to be fixed by a mean or corner constraint. Second, changes in the mean of the λ_j can be set off by changes in α_i – when all the categories are jointly more (less) responsive to position, the resulting addition (loss) of text can be offset by higher (lower) alpha values (with the specific amount for text i depending on θ_i). This issue can be addressed e.g. by imposing that the mean of λ_j equals zero. Third, the

mean of the positions θ_i is not fixed. Increases (decreases) can be set off by shifting the mean of the ζ_j, i.e. making the baseline frequency of all categories smaller (larger). We can resolve this indeterminacy by constraining the mean of the positions. Fourth, only the ratio of the variance of λ_j and the variance of θ_i is fixed. We can infer the variation in positions relative to the variation of the discrimination parameters, but not their absolute levels. This implies that either of the two variances needs to be constrained (for instance to be one), while the other is allowed to vary. Fifth, the polarity of the positions θ_i is not determined, since larger values can represent either more rightist or more leftist positions (we can multiply both λ_j and θ_i by negative one and get the same result). This reflection invariance can be prevented by constraining the order of either two positions $\theta_i < \theta_{i'}$ or discrimination parameters $\lambda_j < \lambda_{j'}$.

In a Bayesian framework, these restrictions can in practice be implemented as either hard constraints or as soft constraints through the choice of priors. We use a hard constraint (choosing a reference category j for which $\zeta_j = 0$) to address the first indeterminacy, and soft constraints through priors for the second to fourth:

$$\begin{aligned}
\alpha_i &\sim N(\mu_\alpha, \sigma_\alpha) \\
\zeta_j &\sim N(\mu_\zeta, \sigma_\zeta) \\
\lambda_j &\sim N(0, \sigma_\lambda) \\
\theta_i &\sim N(0, 1) \\
\mu_\alpha &\sim N(0, 5) \\
\mu_\zeta &\sim N(0, 5) \\
\sigma_\alpha &\sim \text{Half-Cauchy}(0, 5) \\
\sigma_\zeta &\sim \text{Half-Cauchy}(0, 5) \\
\sigma_\lambda &\sim \text{Half-Cauchy}(0, 5) \\
\phi^{-1} &\sim \text{Uniform}(0, 200)
\end{aligned}$$

We leave resolving reflection invariance to the post-processing stage, when we invert the scale (if necessary) so that increasing θ_i and λ_j represent more rightist positions (as judged on the basis of prior knowledge). We also "harden" the soft constraints post-hoc by mean-deviating λ_j and standardising θ_i in each draw, and mean-center the ζ_j (which makes the specific choice of the reference category irrelevant). We simulate all models using Hamiltonian Monte Carlo (Hoffman and Gelman, 2014), as implemented in the software package Stan (Stan Development Team, 2015), by sampling from the posteriors following a suitable warm-up period.

C Interpretation of item parameters

To start with, after a bit of algebraic manipulation we can see the relation of the item parameters to their counterparts in the standard 2PL-IRT model wherein $L = a_j(\theta_i - b_j)$:

$$a_j = \lambda_j \tag{8}$$

$$b_j = \frac{\zeta_j}{\lambda_j} \tag{9}$$

The λ_j therefore form the "discrimination" parameters a_j, indicating indicate how a particular policy category j's use varies in response to changes in the latent dimension θ_i. Put differently, the absolute size of λ_j reflects the degree to which a category is positional, and its sign shows whether the category is a "left" or a "right" item. Note, however, that the "difficulty" parameter b_j in the standard model is a combination of the values of the two item-level parameters in the NRM. This equivalency was also noted by Baker (1992), who cautioned that the values of the NRM parameters do not have a simple formulation in terms of the standard (e.g. 2PL) IRT model, because they describe the discrimination and location of specific item category response functions, whose shapes and locations depend on the way the parameter values from all the categories combine (Ostini and Nering, 2009, 18).

This complication is shown in Figure 7, which refers to a hypothetical example with five item categories that differ with regard to their λ_j value.[15] The α parameter was set to 1, and so were all five ζ_j parameters, implying that the baseline frequency of the five categories (i.e. the part unrelated to the latent position) is the same. The left panel shows how the expected number of items falling in the respective category varies with the latent position θ_i, depending on λ_j. The expected number of statements increases (decreases) monotonically over the range of the latent position if λ_j is positive (negative), and does so more strongly the more extreme λ_j. When $\lambda = 0$, the category is not responsive to the latent position at all, and the curve is flat at the level of the baseline frequency determined by the combination of α_i and ζ_j. Also note that in the case of the

[15] Elff (2013, 221) briefly discusses a plot similar to the right panel of Figure 7.

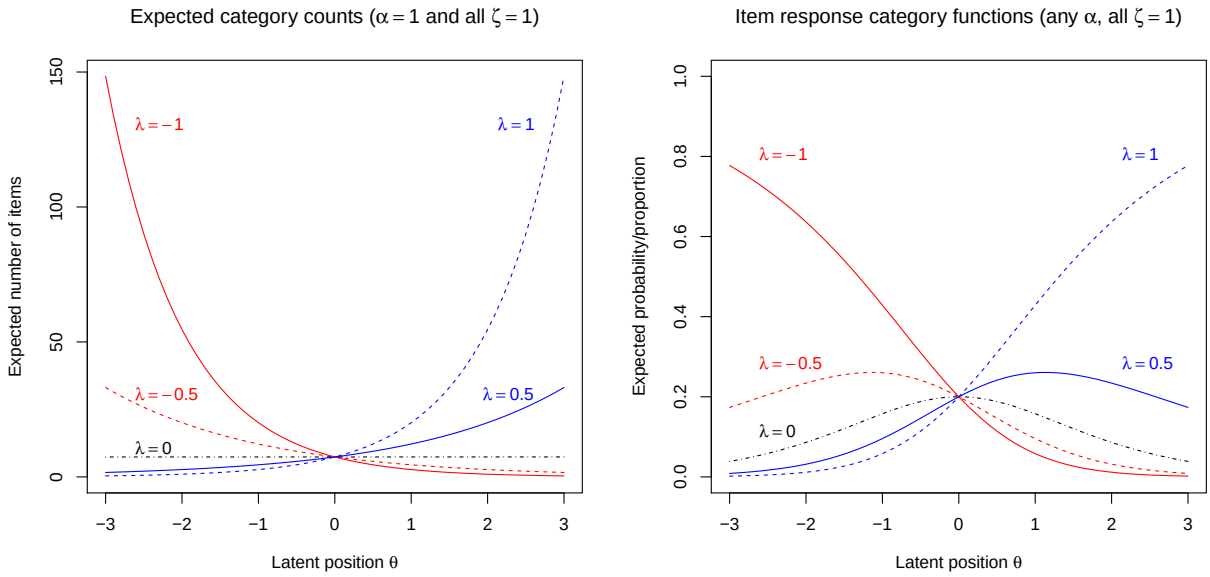


Figure 7: Item category characteristic curves for a hypothetical example with five categories.

example, all curves cross in the same point at $\theta = 0$, due to the baseline frequency being equal as all ζ_j were chosen the same. The right-hand panel now illustrates how the parameter values *jointly* form the expected probability that an item falls in a certain category (or equivalently, the expected proportion of items in a category). Here, due to the interdependence, only the item response functions for the two categories with the lowest/highest λ_j are monotonically decreasing/increasing, approaching zero and one in the limit, as the latent position ranges from plus to minus infinity. In other words, an infinitely rightist/leftist document would only consist of words/statements falling into the rightmost/leftmost category. The curves for the remaining $J-2$ categories, in contrast, follow a unimodal shape. As in the left panel, the equality of the ζ_j implies the curves intersect at $\theta = 0$, with the associated y-coordinate in the right graph equal to $1/J = 0.2$. Also note that the plot in the right panel holds for any α, i.e. regardless of the total number of items.

In this context, note another interesting implication of the IRT model: a certain position can be expressed in many different ways. For example, in order to communicate a markedly leftist position, a party may use one very leftist category a few times, it may use one moderately left category many times, or it may refer to various moderately left categories a few times each.

D Statistical identification of the two-dimensional model

In order to find a unique solution for Eq.6, we apply the following constraints. First, to fix the variances and the covariance of the discrimination parameters, we use a category j for which $\lambda_{1j} = 0$ and $\lambda_{2j} = 1$, and a category j' for which $\lambda_{1j'} = 1$ and $\lambda_{2j'} = 0$.[16] We also rescale the mean position in each dimension to zero. These six restrictions constitute the first set of constraints. For further identification, the λ_{dj} are set to zero on the economic dimension for items that we believe are certainly "non-economic," and to zero on the second dimension for items that we judge to be clearly unrelated to "social" questions as understood here. This implies that we also have a number of categories which do not discriminate on either of the two dimensions, but which are retained in the data and for which we infer a ζ_j parameter. In our estimation below, for example, we constrained "Economic Goals" to be of only economic discriminative power, and "Multiculturalism: Positive" to be of only social discriminative power. Of particular importance, however, is that we leave some items free to be associated with both dimensions.[17] The constraints are completed by setting one of the ζ_j to zero, as in the one-dimensional case.

[16] The categories whose λ_{dj} was set to one were "Free enterprise: positive" for the first and "Traditional morality: positive" for the second dimension. As item for which $\lambda_1 = \lambda_2 = 0$ we choose "Political corruption: negative."

[17] This means that our model is purposefully over-identified, since it is sufficient for identification to apply a single constraint per dimension on λ_{1j} and λ_{2j}, one double constraint on λ_1 and λ_2 for the same category, and two further constraints on any λ_{dj}. Note that these constraints also resolve reflection invariance.

Priors for the two-dimensional model are chosen as follows:

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha)$$

$$\zeta_j \sim N(\mu_\zeta, \sigma_\zeta)$$

$$\lambda_{dj} \sim N(\mu_{\lambda_d}, \sigma_{\lambda_d})$$

$$\theta_{di} \sim N(0, \sigma_{\theta_d})$$

$$\mu_\alpha \sim N(0, 5)$$

$$\mu_\zeta \sim N(0, 5)$$

$$\mu_{\lambda_d} \sim N(0, 5)$$

$$\sigma_\alpha \sim \text{Half-Cauchy}(0, 5)$$

$$\sigma_\zeta \sim \text{Half-Cauchy}(0, 5)$$

$$\sigma_{\lambda_d} \sim \text{Half-Cauchy}(0, 5)$$

$$\sigma_{\theta_d} \sim \text{Half-Cauchy}(0, 5)$$

with (hard) constraints applied to some of the λ_{dj}, as just described.

We post-process the HMC draws so that the variances of the positions (in each dimension) equal one in each draw, and we mean-deviate the ζ_j.

E Assessing convergence

To assess convergence, we rely on the potential scale reduction factor $\hat{R}$ as suggested by Gelman and Rubin (1992) and implemented in the R interface for STAN (Stan Development Team, 2015). Values near one indicate that inferences about the target distribution are unlikely to improve when taking additional draws. Gelman and Rubin (1992, : 297) suggest a threshold of 1.1 as rule-of-thumb for acceptable values. Table 4 shows the distribution of $\hat{R}$ values for the respective parameter across its subscripts (i.e. documents i or categories j).

For the one-dimensional models, all $\hat{R}$ values are very close to one. Examples for the positions of the three UK parties from 1987 that were shown in Table 1 are displayed in Figure 8.

The parameters of key interest from the two-dimensional model also seem to converge very well. Only some of the diagnostics for ζ_j are somewhat above the standard threshold. It appears to be mainly an autocorrelation issue. In a longer run with a stronger thinning factor (1,000 iterations warmup plus 15,000 iterations, thinned by a factor of 15), the mean and the maximum of $\hat{R}$ for ζ_j are smaller (1.015 and 1.048, respectively).

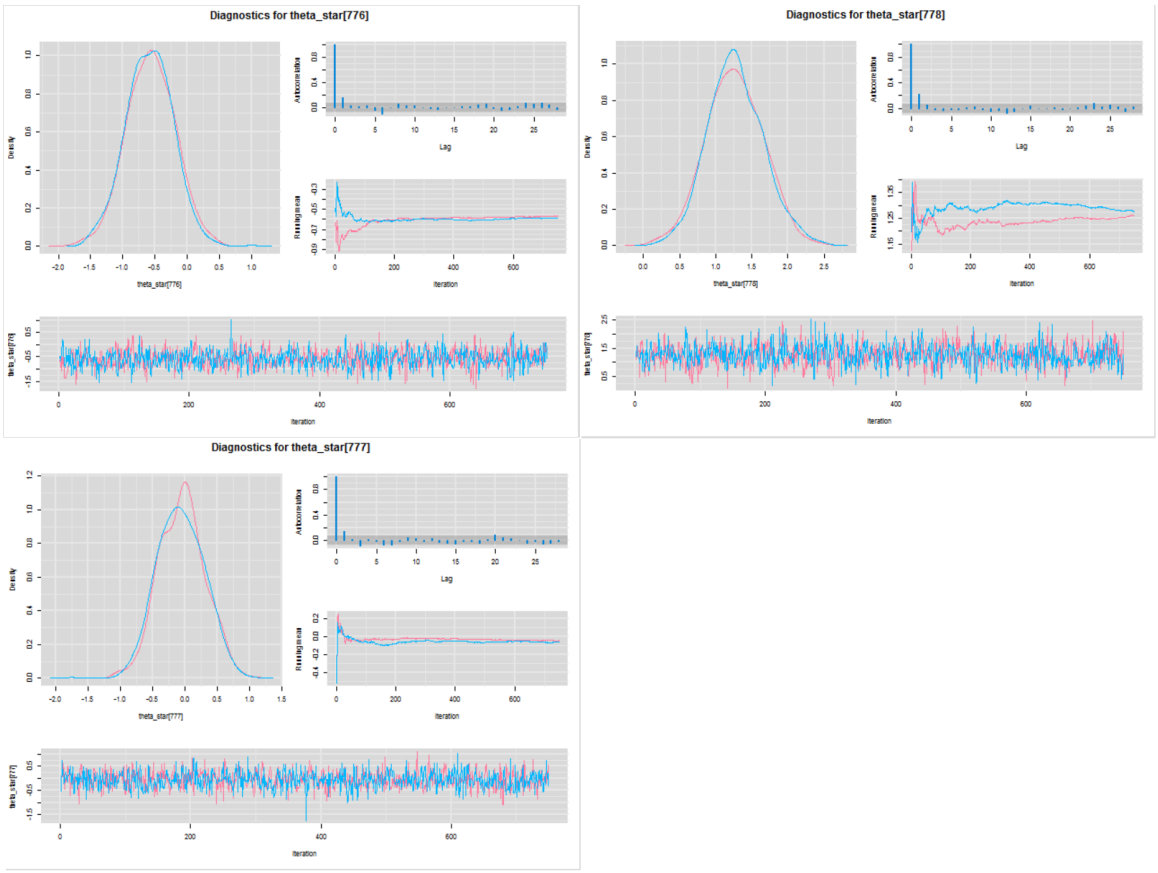


Figure 8: HMC draws for positions of the three UK parties in 1987 from Table 1.

Parameter	Min	Median	Mean	95%	99%	Max
1d-Poisson						
θ_i	0.999	1	1	1.003	1.005	1.009
λ_j	0.999	1	1	1.002	1.002	1.002
ζ_j	0.999	1	1.001	1.004	1.006	1.007
1d-Neg. bin.						
θ_i	0.999	1	1	1.003	1.005	1.01
λ_j	0.999	1	1	1.002	1.006	1.009
ζ_j	0.999	1	1	1.004	1.006	1.008
ϕ_j^{-1}	0.999	1	1	1.003	1.004	1.005
2d-Poisson						
θ_{1i}	0.999	1.002	1.004	1.014	1.023	1.042
λ_{1j}	0.999	1.001	1.002	1.003	1.024	1.037
θ_{2i}	0.999	1.002	1.004	1.015	1.023	1.038
λ_{2j}	0.999	1	1.001	1.005	1.007	1.008
ζ_j	1	1.037	1.047	1.111	1.137	1.145

Table 4: Potential scale reduction factors

F Full results of scaling model for CAP data

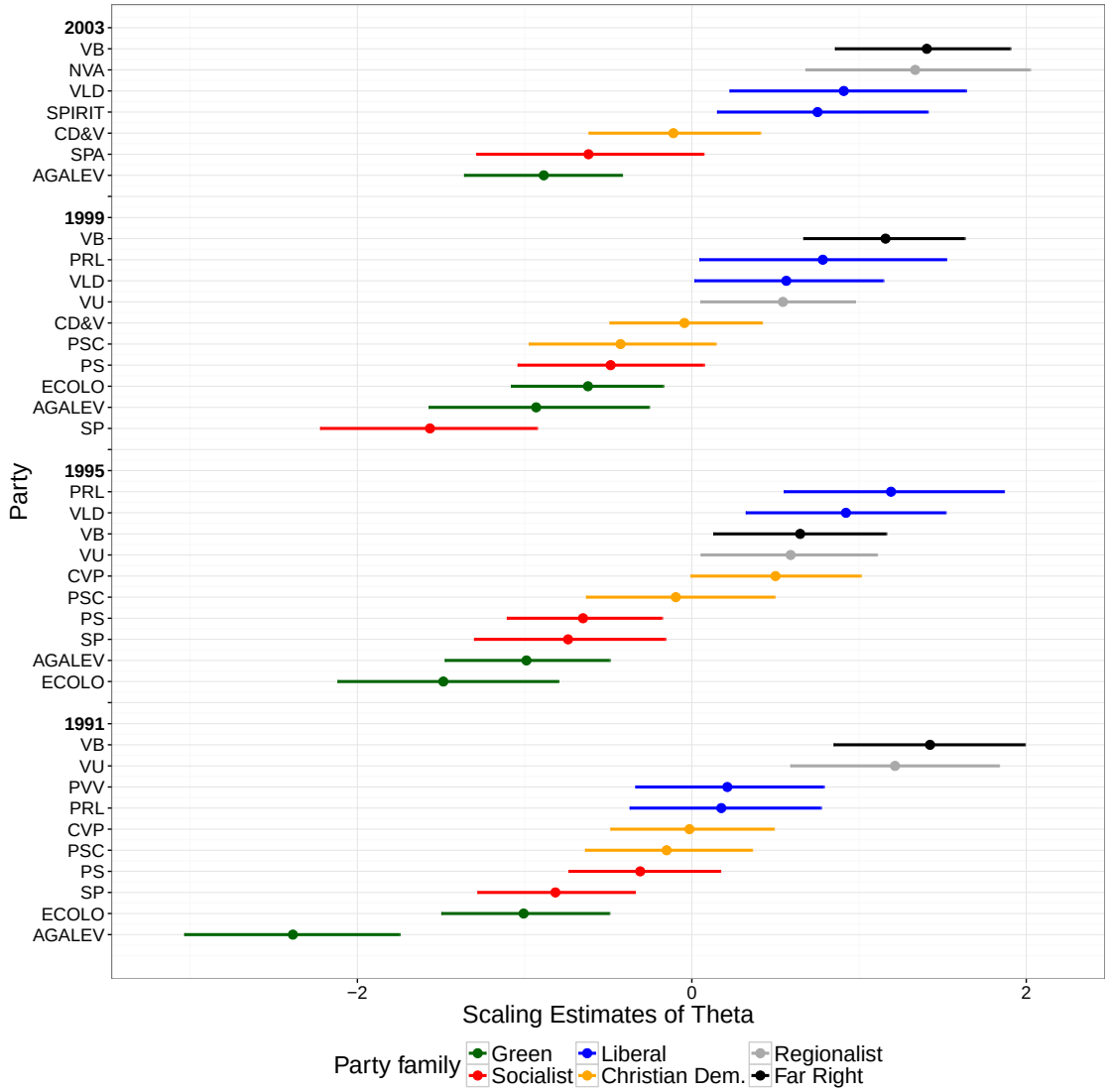


Figure 9: Left-right estimates for Belgian Parties from the Comparative Agendas Project dataset (1991–2003).