An iterative expert survey approach for estimating parties' policy positions

Kostas Gemenis
University of Twente

Abstract. This article introduces the iterative expert survey approach in estimating parties' policy positions. Methodologically, the proposed approach is based on the tradition of 'judgemental' coding in the content analysis of political text, and incorporates the idea of anonymous iteration among a panel of expert coders taken from the method known as 'Delphi'. Anonymous iteration presents an effective way of reducing the random error, and potential bias arising from inter-expert/coder disagreement evident in other popular methods. I provide an empirical demonstration of the approach by estimating parties' policy positions in the context of a voting advice application in Germany, and argue that the method has considerable potential to generate valid and reliable data on party positions cross-nationally and retrospectively.

Keywords: Content analysis, Expert judgement, Delphi method, political parties

1. Introduction

Some of the most important questions in political science require information about the location of political parties in terms of their policy preferences. This is particularly true for theories that use ideological proximity to explain government coalition formation (De Swaan, 1970), government spending (Lau and Frey, 1971), as well as spatial models of voting (Downs, 1957) and party competition (Stokes, 1963). Although researchers have proposed various methods for estimating parties' policy positions, no approach, no matter how popular, has proved immune to methodological criticism. If indeed the inferences in models employing ideological proximity as the main independent variable indeed inferences are sensitive to artefacts of the methods employed to measure parties' policy positions (e.g. Lewis and King, 1999, pp. 26–31; Müller and Strøm, 2003, pp. 8–9; Benoit et al., 2009, pp. 507–510; Golder and Stramski, 2010, pp. 98–99), this raises questions about the validity of several studies that have influenced substantial research agendas in political science. More importantly, as political science forays into actively engaging with the public through tools that promise to 'advise' citizens as to which party lies closer to them in ideological terms (Garzia and Marschall, 2014; Katakis et al., 2013), estimating parties' policy positions ceases to be a matter of academic debate and attains an important ethical dimension.

Preprint of paper accepted for publication in: Quality and Quantity

Recognizing that estimating parties' policy position is an important endeavour in its own right, not least because of its ethical and methodological ramifications, this article introduces a novel approach for doing so. To this end, I introduce the assumptions of the approach known as 'judgemental coding' that lies halfway between content analysis and the traditional expert survey. Furthermore, I present the previous attempts to estimate parties' positions in this tradition, and mention their shortcomings. Using the problem of estimating parties' policy positions in voting advice application as a motivating example, I introduce the idea of implementing a classic forecasting technique within the judgemental coding tradition in order to overcome these shortcomings. The article outlines the theoretical and methodological aspects of this approach, references previous research that has evaluated its efficacy, and evaluates its central assumption using data on German political parties' policy positions. Finally, I conclude with some recommendations for applied researchers.

2. Existing approaches and their shortcomings

The goal of estimating parties' policy positions has resulted in an extensive and lively debate with respect to the advantages and drawbacks of the proposed methods (for critical overviews, see Krouwel and Van Elfrinkhof, 2014; Mair, 2001; Volkens, 2007). The most popular of these methods can be grouped into two distinct approaches: surveys-based methods, and the content analysis of political text. The most popular embodiment of the latter is the Manifesto Project (formally Manifesto Research for Political Representation, formerly Comparative Manifestos Project). The Manifesto Project works as a classic content analysis scheme (Krippendorff, 2004; Neuendorf, 2002) where the manifest content ('quasi-sentences') of party manifestos is assigned to policy categories of a coding scheme and the output data are scaled to ideological dimensions of interest. Beginning in 1979, the project has amassed an impressive wealth of data on the policy positions for more than 916 parties across 629 elections in over 50 countries. Despite its popularity that stems from the wealth of provided data, the Manifesto Project approach to estimating parties' positions has been criticized extensively in terms of its validity and reliability (for a review, see Gemenis, 2013; Zulianello, 2014). More specifically, it has been argued that the Manifesto Project uses a coding scheme based on unrealistic assumptions about the nature of party competition—the so-called 'saliency theory' (Dolezal et al., 2014), disregards the amount of uncertainty associated with the coded documents (Benoit et al., 2009), uses only one coder per document even though it has been shown that different coders will handle the same document differently (Mikhaylov et al., 2012), and scales the coded content to ideological dimensions (such as the left-right) by making controversial assumptions as to what is considered 'left' and 'right'.[1] Moreover, the various computerized methods that were proposed to automate the hand-coding process and improve coding reliability have not been able to match the cross-national longitudinal data of the Manifesto Project. This is largely because the validity of the output is often context-specific and highly dependent on unrealistic assumptions about the nature of political text (Grimmer and Stewart, 2013; Lowe, 2008).

Expert surveys have emerged as the most popular survey-based approach. The logic of the expert survey is remarkably different, as compared to the manifesto coding of the Manifesto Project (see Table I). In expert surveys there is no specific document source, but experts are asked to make judgements of party positions based on their personal knowledge. Consequently, party positions are estimated by aggregating expert judgements using a measure of central tendency such as the mean. Expert surveys have the advantage of estimating parties' positions on dimensions of interest directly, and as such they do not need to make additional assumptions about scaling data into dimensions. Nevertheless, expert surveys have several drawbacks in their own right. It has been argued that the estimation process leaves much to be desired as it is not clear what experts actually evaluate (Budge, 2000). As expert survey questions come with minimal instructions, experts are asked to make personal judgements without reference points, resulting to judgements that are interpersonally and/or cross-nationally incomparable. This incomparability manifests as inter-expert disagreement that has been shown to correlate with certain party characteristics such as their size and ideological background (Steenbergen and Marks, 2007). Moreover, expert surveys are prone to projection effects as the experts' sympathy towards the evaluated parties has been shown to affect their judgements (Curini, 2010). Proponents of expert surveys contend that expert disagreement can be resolved through statistical aggregation because the errors will 'cancel each other out' (Steenbergen and Marks, 2007, p. 359). Inasmuch the error component in personal judgement is not only a function of party characteristics (as Steenbergen and Marks contend) but also of personal characteristics such as the experts' knowledge or ideological background, this 'cancelling out'

[1] The scaling critiques of the Manifesto Project are too many to be referenced here. Most are reviewed in Gemenis (2013a), while Dinas and Gemenis (2010) compare some of the most popular scaling methods in terms of their validity and reliability.

will produce an estimate ‘centering upon the mean of the erroneous judgements rather than the true value’ (Rowe, 1992, p. 158, see also Booker and Meyer, 1988).[2] If ‘low reliability increases the likelihood of low validity’ (Krippendorff, 2004, p. 214), in the presence of expert disagreement, simple statistical aggregation is not sufficient to guarantee valid measurement of party positions.

Between the coding of manifest content in the Manifesto Project and the conventional expert survey lies a less-studied but, paradoxically, often used approach. Unlike classic content analysis that works by assigning the manifest content (words, sentences, paragraphs) of text into the categories of a coding scheme, this ‘judgemental coding’ (see Table I) requires from the coder to make judgements regarding positions by reading the whole of a text. Krippendorff (2004, pp. 139–141) calls this approach ‘simulation of interviewing’, while Hawkins (2009, p. 1049) likens it to the pedagogical assessment technique that assesses the quality of work by assigning grades without any intervening calculations. Moreover, since the distinction between manifest and latent content is not clear-cut, those engaging in judgemental coding could also look for underlying meanings by ‘reading between the lines’, in addition to coding parties’ positions articulated through manifest content (Neuendorf, 2002, pp. 23–24; Riffe et al., 2005, pp. 36–38).

Judgemental coding has been used extensively in sentiment analysis of media content (e.g. Dalton et al., 1998), where researchers have sought to systematize it through detailed coding schemes and inter-coder reliability tests (e.g. Walter et al., 2013). With respect to estimating parties’ policy positions judgemental coding bears resemblance to what Mair (2001, pp. 13–14) called ‘secondary reading’. The theoretical underpinnings of reading through a document rather than counting the frequency of terms used, lie on the position side of the salience versus position debate (Laver, 2001b). This side purports that the relative frequency that a subject is mentioned does not really matter. What matters is the general position that is borne out of the text. This also implies that the estimate will be directly given on a dimension of interest without the need for scaling. As seen in Table I, judgemental coding shares these two characteristics with expert surveys, while the remaining two, namely the coding of documents by a single person are attributes shared with the Manifesto Project approach to coding.

[2] This problem has also been identified in mass surveys where voters’ lack of knowledge leads to implausible centrist estimates of parties that are known to be extreme (Tilley and Wlezien, 2008).

Table I. A typology of approaches for estimating parties' positions.

	Manifesto Project	Expert survey	Judgemental coding	Kieskompas	Iterative expert survey
Theory	Position via salience	Position	Position	Position	Position
Source of evidence	Single document	Expert knowledge	Single document	Hierarchy of documents & elite survey	Documents & expert knowledge
Aggregation	None, $N = 1$	Statistical	None, $N = 1$	Unstructured behavioural	Structured behavioural
Scaling	Likert scaling or data reduction techniques	None needed	None needed or Likert scaling	Likert scaling	None needed or Likert scaling

The most known piece of scholarship in this tradition is the 'judgemental coding' approach of the International Comparative Political Parties (ICPP) project. Beginning in 1967, the ICCP estimated the positions of 158 parties across 53 countries between 1950–1963 on 13 issue dimensions using secondary sources (Janda, 1980). Over the course of years, this approach has been many times reinvented under different names. For instance, the Euromanifestos Project, in addition to the manifest coding that was carried with a coding scheme adapted from the Manifesto Project, also asked its coders to evaluate the content of each Euromanifesto *as a whole* on seven broad scales (Braun et al. 2010, p. 48, emphasis added). Other approaches that follow the same logic include the 'confrontational approach' (Pellikaan et al., 2003; Dolezal, 2008; de Lange, 2007; Gemenis and Dinas, 2010), 'holistic grading' (Hawkins, 2009), the 'check-list approach' (Gudbrandsen, 2010; Odmalm, 2012; Ruedin, 2013), and the coding scheme of the INTUNE project (Conti and Memoli, 2012). Most recently, the judgemental coding approach to content analysis has been picked up by the online tools that aim to provide voter information with regards to which party matches best their own policy preferences. In particular, Krouwel et al. (2012) proposed a hybrid variant that combines judgemental coding with an elite survey (see also Krouwel and Van Elfrinkhof, 2014). This approach is sometimes referred to as the '*Kieskompas*' method, after the popular voting advice application in the Netherlands, and has been most famously employed in 'EU Profiler', the 2009 pan-European voting advice application. In a nutshell, *Kieskompas* involves a questionnaire sent to party headquarters asking for a designated spokesperson to position the party and provide some factual evidence with regard the placements, while a small team of experts works concurrently to place parties on the on the basis of their manifestos or other public statements. The two placements are compared to one another and, in cases of disagreement, parties are asked to reconsider their initial placement. After several rounds of iteration, a consensus is usually reached between the team of experts and the parties, and these consensus positions are used as the final estimates of parties' positions (Krouwel et al., 2014).

Krouwel and Van Elfrinkhof (2014) argue that the iteration between the team of experts and party headquarters manages to overcome the weaknesses of the individual methods by combining their strongest aspects. To a certain extent this is true. The *Kieskompas* method seems to have found a workaround to the problem of adequacy and credibility of document sources. It can be the case that the document sources are too short and therefore inadequate for fully representing parties' policy positions, a point made effectively by Benoit et al. (2009)

with respect to the Manifesto Project. Moreover, even if the document source appears to be adequate, one's inferences could depend heavily on the type of coded document: manifesto, pamphlet, party leader speech and so on (Gemenis, 2012a). While the ICPP asked its coders to simply estimate the credibility and adequacy of sources per party using a 9-point scale (Janda, 1980, pp. 13–18), the EU Profiler dealt with document credibility and adequacy by using multiple sources in a document hierarchy (Krouwel et al., 2012). Moreover, the estimates of the elite survey (party headquarters) can be particularly useful when there is much uncertainty on the expert side, most likely because the issue in question is not salient and consequently does not appear in the party manifesto (see also Mair, 2001). Furthermore, experts can prevent parties from falsely presenting themselves as having the most popular, per electorate, positions.

Notwithstanding the improvements, the *Kieskompas* method does not address the problem of inter-coder agreement. As Krippendorff, (2004, p. 140) writes, judgemental coding 'becomes unreliable when the writing is voluminous and the informative passages are scarce and therefore easily overlooked.' Moreover, interpreting the found passages can be a highly subjective exercise (Riffe et al., 2005, pp. 38).Coders in judgemental coding are not just individuals trained to do a specific (and rather menial) task of assigning words or sentences to the categories of a coding scheme, but are treated more like experts as they can be asked anything from locating relevant bits of text in a document and interpret them in terms of party position to a specific issue (as in the INTUNE project), to providing an informed judgement after reading a variety of document sources (as in the ICCP project). In such a context, coders could disagree with one another due to a variety of reasons (Bolger and Wright, 1992; Einhorn, 1974; Meyer and Booker, 1991, pp. 19–20; Mumpower and Stewart, 1996). Coders could base their estimate on different documents or, if only one document is available, on different parts of the same document. Alternatively, coders may evaluate exactly the same parts of the same document, but understand and interpret the content differently. Furthermore, even if coders evaluate exactly the same parts in the same document and interpret them in the same way, they might be attaching different weights to the same pieces of information and therefore provide different estimates. Indeed, there is evidence of considerable disagreement during the coding of EU Profiler questions even when the coders were using the same documents (Gemenis, 2013b, pp. 278–279).

How does the *Kieskompas* method handle such disagreements? For instance, the EU Profiler, relied on 'discussions among team members' and consultations with experts and the EU Profiler leadership (Trechsel and Mair, 2011, p. 20). Despite claims that such group discussions would maximize inter-coder reliability (Trechsel and Mair, 2011, p. 13), methodologists in both content analysis and expert estimation literature think not. Achieving consensus by group discussions among team members, in a process formally known as 'unstructured behavioural aggregation' (Ferrell, 1985, p. 135), does not equal coding reliability. Armstrong (2006, p. 5) makes a strong point in support of this idea when he argues against estimating quantities based on face-to-face meetings. Some people are louder than others, some are more powerful or prestigious. Some people voice their opinions and do not listen to others, yet others listen to what others say but do not take time to think for themselves. In this respect, Meyer and Booker (1991, pp. 40–51) noted that elicitation in the context of unstructured behavioural aggregation is subject to 'group-think' effects due to social pressure. As Krippendorff (2004, p. 217) warned, 'in groups like these, observers are known to negotiate and to yield to each other in tit-for-tat exchanges, with prestigious group members dominating the outcome [...] and coding comes to reflect the social structure of the group'. Armstrong (2006, p. 6) therefore suggests that estimation can be improved when the elicitation method guarantees that the opinions of individual experts/coders can be stated independently from one another, and when they are aggregated using a 'pre-determined mechanical scheme'. Moreover, the practice of experts/coders working entirely independently to each other, enables the calculation of formal measures of disagreement that can be used to gauge reliability (Krippendorff 2004, p. 219). In the next section I propose a solution to the problem of expert disagreement by formalizing the coding process using an iterative expert survey.

3. Delphi: An iterative expert survey

The problem of expert disagreement is certainly not unique to political science. Psychologists, computer scientists, managers, professionals in the health sciences, and practitioners in many other fields encounter situations where expertise is called to estimate quantities of interest. Consider the following example. A group of physicians have just examined a patient and need to make a diagnosis. Each of the physicians has access to a set of medical tests that can be used for the diagnosis (x-rays, biopsies and so on). The physicians, however, are likely to interpret these tests differently and assign different weights to them when making their diagnoses. How should these disagreements be compromised? The quest for a dealing effectively with expert disagreement has resulted in an extensive literature across disciplines (e.g. Bolger and Wright, 1992; Einhorn, 1974; Mumpower and Stewart, 1996; Rowe, 1992; Shanteau and Hall, 2001; Tversky and Kahneman, 1974). When confronted with the aforementioned example, a political scientist would typically choose to aggregate the different physician judgements by taking the mean amongst them. This is exactly what conventional expert surveys conducted among political scientists do. However, because expert judgements in political science are not statistically independent from each other (i.e. they are correlated), the presence of random error and/or bias is likely to be shared my multiple experts (Booker and Meyer, 1988, p. 136). This means that averaging could easily lead to an erroneous judgement (Rowe, 1992, p. 158). If simple statistical aggregation will not 'cancel out' errors, how could expert disagreement be handled? One could propose for the physicians to convene in a room and try to agree on a diagnosis by taking to each other (i.e. in the 'unstructured behavioural aggregation' of the EU Profiler). As we saw, however, authors such as Armstrong (2006) and Krippendorff (2004) have warned that such discussion will most likely be dominated by those with high prestige or strong personalities, and not by those making the most persuasive arguments.

An alternative method to deal with expert disagreement that proved to be highly popular over the past 50 or so years is commonly known as 'Delphi'. The Delphi method was first used in the 1950s to forecast technological changes (Dalkey and Helmer, 1963) and has been since applied in many different contexts. It is based on three principles: a) anonymity, b) statistical aggregation, and c) feedback. Conventional expert surveys use the first two principles of anonymity and statistical aggregation, but do not employ any feedback mechanism. Conversely, the coding of party positions on the basis of group discussions as employed by *Kieskompas* and EU Profiler, uses feedback that results in some sort of behavioural aggregation, but is not anonymous. Therefore the Delphi method that typically employs all three principles, can be described as an 'iterative expert survey'.

The typical Delphi method involves a rather simple and intuitive process (see Figure 1). A 'moderator' selects a panel of experts (thereafter referred to as 'panellists') and asks them to provide estimates on quantities of interest using typical survey questions, but also to justify each of their estimates. This justification may come at the form of an argument, or could point to the source(s) that the panellist has consulted for his or her estimation. The panellists work independently of each other and without knowing the identities of the other panellists involved. The moderator then collects the responses and gives feedback to the panellists for a second round of estimation. The nature of the feedback in Delphi can vary. The moderator can give measures of

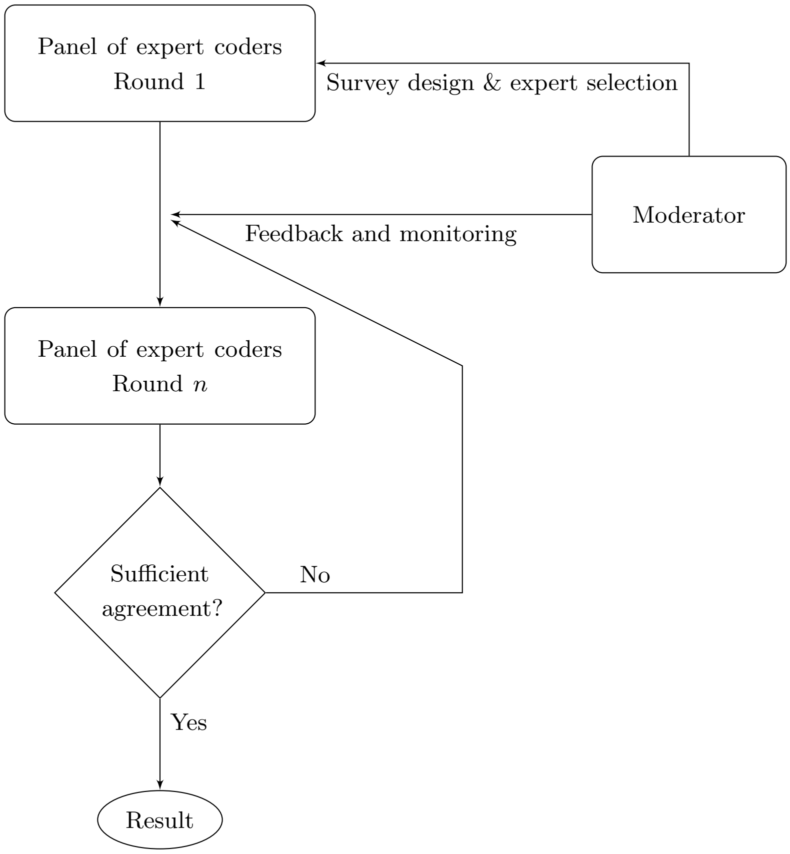

Figure 1. The 'Delphi' iterative expert survey.

central tendency of the responses in the first round (median, mean), the minimum and maximum scores with the associated justifications, a combination of both, or can even opt to feedback each and every estimate along with its justification. In the second round, the panellists are asked to consider the other anonymous estimates and the associated justifications and revise (if needed) their initial estimate. An important characteristic of this feedback is that it is anonymous. Experts are unaware of the identities of the other experts and cannot tell from whom each estimate/justification comes. In addition, it is possible to render the feedback process quasi-double blind to avoid moderator bias by using a system where the moderator knows the identities of the panellists but cannot tell who gave what estimate and justification. This iteration can be repeated over a number of rounds, that can be predefined or otherwise decided by the moderator.

The combination of statistical aggregation after each round with the highly structured way in which the feedback is provided, results in a final aggregation of individual judgements that can be best described as 'structured behavioural aggregation' (Ferrell, 1985, p. 140). A considerable body of evidence (for a comprehensive meta-analysis, see Rowe and Wright, 1999) has shown that such aggregation produces more accurate estimates compared to the mathematical aggregation of conventional expert surveys or the unstructured behavioural aggregation of group discussions. Anonymity plays a crucial role as it postulates that consensus is reached due to the quality of information associated with the estimates and is unaffected by group dynamics stemming from the personalities (and biases) of individual panellists.

The presence of feedback between rounds of estimation complements the desirable properties of anonymity. Although anonymity can ensure the absence of personality effects in the context of a feedback situation, it cannot eliminate the additive effect of individual personality traits on the final outcome. Put otherwise, a hard-headed panellist will not be more likely to influence the outcome, but will be personally less likely to update his or her own estimate. In this respect, Mulgrave and Ducanis (1975) found that dogmatic panellists were more likely to change their estimate in the second round towards the mean response of the first round. They therefore hypothesized that, in the absence of an eponymous authority that the dogmatic individual is likely to follow, the average estimate was considered to be such an authority. In the presence of individual personality effects, the revision of estimates towards the average as postulated in the Delphi mechanism may work in favour of consensus but to the detriment of accuracy. Consider a situation where a highly informed but dogmatic panellist changes his or her estimate toward the average of the remaining panellists who provided inaccurate estimates because they lacked information. In the context of the Delphi method, this problem can be solved via increasing the amount of feedback information between rounds. Rather than providing the mean estimate of the first round, many applications of the Delphi method choose to provide the full distribution of estimates, along with the provided justifications of the minimum and maximum estimates given. This idea was supported in the meta-analysis of Rowe and Wright (1999) who found that when feedback consisted of the justifications the panellists were asked to provide during the first round (as opposed to a measure of central tendency), change from the first to the second round became less likely, but when it occurred the estimates became more accurate.

The idea that panellists will adjust their estimates based on the feedback information provides a Bayesian insight within the framework of classical statistics. The problem is, however, that experts should not be regarded as being naturally Bayesian because, as many studies have shown, they often fail to update their estimates in light of the new information (see Meyer and Booker, 1991, p. 23). Rather than Bayesianism, the theory that best seems to fit the empirical manifestations of the Delphi-type iterative expert survey is rather informal Dalkey (1975, pp. 243–246), and known as the 'theory of errors': knowledgeable panellists will stick to their original estimates, whereas those with little information will revise their estimates towards the group average (Parenté and Anderson-Parenté, 1987). The psychological mechanism behind this 'theory of errors' can be postulated as follows. As Scheibe et al. (1975, p. 270) noted, when Delphi panellists are confronted with the estimates of the first round, they can ignore the feedback information and continue providing the same estimates, or move towards the group average as hypothesized above.[3] As argued in the 'theory of errors', panellists' willingness to ignore the feedback information will be a function of the certainty they attached to their first round estimates. The idea here is that highly confident persons are less likely to move to majority opinion (Scheibe et al., 1975, p. 274).

Despite the proliferation of the Delphi method in many disciplines, there have been very few applications in political science, almost exclusively in the election forecasting literature (e.g. Jones, 2008). This lies in stark contrast with the desirable properties of the iterative expert survey in the context of parties' policy positions outlined in the previous sections. The only application of the Delphi method in estimating parties' policy positions has been in the context of the emerging field of voting advice applications. Elsewhere, I provided anecdotal evidence from Greece, where iteration between rounds helped achieve consensus among experts in estimating parties' positions in policy areas where most experts were uncertain (Gemenis 2012b). In addition, Gemenis and Van Ham (2014) compared the Delphi method to a conventional expert survey in estimating Dutch parties' policy positions. The average panellist agreement across eight parties and six policy issues in the first round was close to experts' agreement for parties' positions on the left-right scale in a conventional expert survey. Expert agreement on the six specific policy questions in the conventional expert survey, however, varied a lot in line with issue salience expectations, with experts disagreeing about the placements on issues that parties do not 'own'. Even in such issues, however, agreement proved to be higher in the iterative expert survey, especially after the second round. In the

[3] A third possibility occurs when the panellist is convinced that the estimate he or she has given in the first round was correct. In this case, the panellist can provide an estimate towards the extremes of the scale in an attempt to influence the group average.

following section I provide an extension to these arguments using data from German political parties.

4. Empirical application: Estimating parties' policy positions in Germany

To investigate empirically the assumptions behind the 'theory of errors' of the iterative expert survey I use data collected during the implementation of *Parteienavi*, a voting advice application (VAA) designed by the Preference Matcher consortium in collaboration with a team at the University of Konstanz (Shikano et al. 2014). The application was launched a few weeks prior to the 2013 German federal election and its design included estimating German parties' positions on 35 policy questions using the iterative expert survey method outlined in the previous section. A total of 12 panellists agreed to take part in the coding for a modest remuneration.[4] Following the spirit of the method that is intended to be used with a diverse group of experts (Rowe and Wright, 1999, p. 371) the panel included a mix of faculty members and graduate students specializing in German elections and party politics. The original idea was to assign panellists to parties in a way that there would be at least five panellists per party. When this proved to be impossible due to time constraints, five panellists were assigned to each of the five major parties (CDU/CSU, SPD, *Bündnis 90/Die Grünen*, *Die Linke*, FDP) and three to each of the remaining two (AfD and *Piratenpartei*). Because the standard practice of the Delphi method is to use at least five panellists, I excluded from the analysis the estimates of the two smaller parties.[5] The dataset therefore contains 1750 observations (5 parties*5 panellists*35 issues*2 rounds) that were arranged in wide format: 875 observations over two rounds.

In addition to the estimate for each party on each policy issue of a 5-point symmetric scale ranging from 'completely agree' to 'completely disagree', the panellists were asked to provide a measure of their confidence in estimating each party position using a 3-point ordinal scale

[4] There is disagreement whether panellists should be remunerated. Meyer and Booker (1991, pp. 89–90) generally advise against remuneration. Future studies may use remuneration for graduate students, but not for panellists who are considered to have substantial expertise and could otherwise be motivated in participating because of their interest in the intrinsic aspects of the project.

[5] In their meta-analysis of evaluative studies of the Delphi method, Rowe and Wright (1999) reported that only three out of 27 studies reviewed used panels with fewer than five panellists.

(very confident, somewhat confident, not confident at all),[6] as well as a written justification for each estimated position. In particular, the panellists were advised to refer to 2013 election party manifestos or other verifiable sources as much as possible, although they could always provide a personal justification in case this was not feasible. All this information was entered in an online platform that was designed for the application of the iterative expert survey method to estimating parties' policy positions (see Djouvas et al., 2014). The online platform was designed in a way where upon completion of the first round of estimation, panellists could be invited in subsequent rounds and provided with certain feedback information from previous rounds. For the second round, the feedback information consisted of the individual estimates along with the respective confidence estimates and provided justifications and sources. After the first round, the online platform presented the relevant information per question in an informative left-to-right structure. Panellists could see the distribution of responses in a graph on the left of the page, the (anonymized) feedback information per panellist in the middle of the page, whereas they could enter they could update their initial estimates (position, confidence, justification, and source) using a menu on the right of the page.

If the feedback mechanism in the iterative expert survey method is capable of mitigating expert uncertainty and bias, we would expect to see some evidence in support of the 'theory of errors' where confident panellists would hold on to their estimates and non-confident ones would move towards the panel average. Previous research has showed that confidence associated with individual responses was significantly, negatively, but not very highly correlated with change in round 2 in some studies (Scheibe et al., 1975, p. 275) but insignificant in others (see Rowe and Wright, 1999, p. 372). This implies a conditional relationship. In particular, I hypothesize that the effect of distance from the median in the first round, on change in the second round will be conditional to the level of confidence the panellist attached in the first round estimate. Unsure panellists that have given estimates far from the group average are expected to revise them towards the median, whereas confident ones are expected to hold on to their estimates. If such conditional relationship is found, then we would have empirical support of the 'theory of errors'. If not, then change in the second round could be simply reflect the artefact of 'regression towards the mean' rather than

[6] Other studies opted to operationalize confidence in estimation using self-rated measures of expertise. These measures produced inconclusive results in Delphi studies (see Rowe and Wright, 1999, pp. 371–372), or have failed to establish a link between self-rated expertise and distance from the group average in conventional expert surveys (Albright and Mair, 2011).

increase in certainty about the given estimate (Tversky and Kahneman, 1974, pp. 1126–1127). I therefore investigated whether panellists update their initial estimates according to their estimates' distance from the panel average, conditional to the degree of confidence they attach to them.

To do so, I created a binary dependent variable, where 1 indicates change in the 5-point response scale estimate from the first round to the second, and 0 indicates no such change. The main independent variable is distance of the panellists' estimates from the group median, the measure of central tendency typically used in Delphi surveys (Dalkey, 1975, p. 249; see also Meyer and Booker, 1991, pp. 316–317), in each of the 35 policy questions. The modifying variable which is expected to moderate the effect of the independent (distance) on the dependent (change) variable, assuming the conditional effect of the 'theory of errors', is the confidence panellists attached in each of their estimates in the first round.

Conditional effects in regression analyses are typically assessed with multiplicative interaction terms. Nonetheless, because multiplicative terms in models with binary dependent variables, such as the one used in our analysis, cannot be interpreted directly, the results need to be presented in graphs plotting predicted changes in logged odds (the marginal effect) across the values of the modifying variable (Brambor et al., 2006). Figure 2 presents the results of our logistic regression with interaction terms analyses.[7] The left panel of the figure presents clear evidence supporting the conditional effect in the hypothesis behind the 'theory of errors'. The positive values of the marginal effect on the Y axis indicate that the further the estimate from the median in round 1, the more likely panellists would change this estimate in round 2. This relationship however, is conditional to the confidence the panellists attached to these first round estimates. The marginal effect of distance from the median in round 1 on change in round 2 increases when the confidence panellists attach on their estimates decreases. Note, however, that the difference between estimates assessed as 'very confident' and 'somewhat confident', and 'somewhat confident' and 'not confident' is not statistically significant. The conditional effect is present when one compares the 'highly confident' to 'not confident' estimates. 'Highly confident' panellists are much less likely to change their estimates as a function of distance from the median estimate compared to panellists that are 'not confident'.

---

[7] To account for multiple observations from each panellist, the logistic regressions were specified with standard errors clustered at the panellist level (see Arcenaux and Nickerson, 2006).

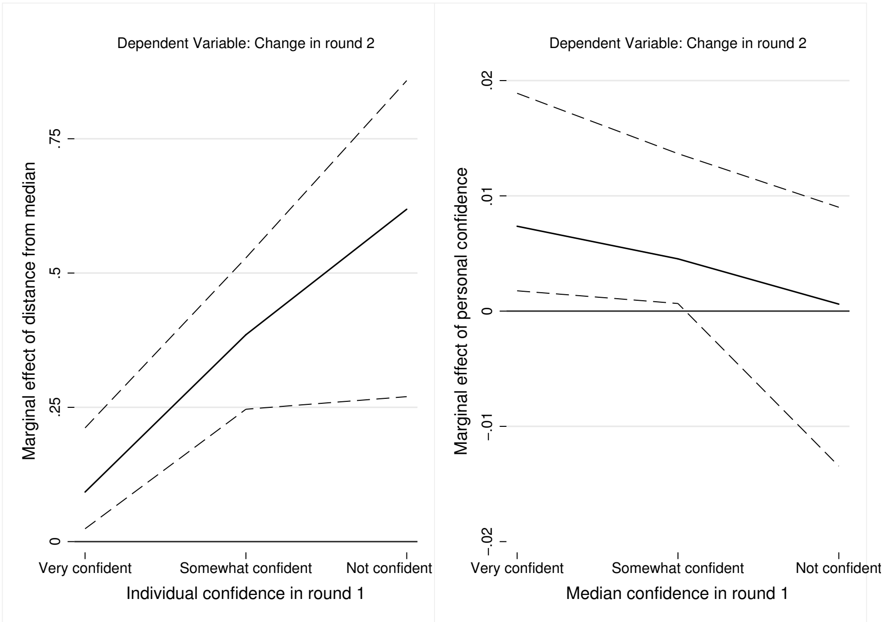

Figure 2. Conditional effects on estimate updating in a two-round iterative expert survey.

Furthermore, the panel on the right provides additional insight on the 'theory of errors'. In this case the main independent variable is confidence to the first round estimates whereas the modifying variable is the median confidence in the first round. The idea here is that the effect of individual confidence on change would be conditional on the average group confidence. If a panellist is not confident but others are, we should expect that the panellist should change his or her estimate. If, however, the average confidence in the group is low, we should not expect that the non-confidence panellist would change his or her estimate. This hypothesis is supported by the analysis presented in the right panel of Figure 2. The positive values of the marginal effect on the Y axis indicate that, when confidence decreases, change in round 2 is more likely, but this effect is conditional on the average confidence. When the median panellist is 'not confident', the marginal effect of confidence attached to individual estimates on change becomes indistinguishable from zero.

The analysis of the German data therefore shows considerable evidence in support of the 'theory of errors'. While panellists who are confident about the validity of the estimates they provided will insist, those who are not confident will be likely to updated them and converge towards the panel median. What are the implications of this panellist-level analysis on the aggregated estimates of parties' positions? The empirical analysis showed changes in 25 out of the 175 (5 parties*35 questions) estimates (14.3%). This is a considerable rate of change, considering that the aggregation of panellists' judgements is based on the medians which are robust to outliers. This can be attributed to the theory of errors, since change at the individual level is not only a function of the distance from the median but also a function of confidence in the estimation. The anonymous iteration with feedback can therefore reduce uncertainty in expert judgement and increase the validity of estimation.

5. Conclusions

In this article, I presented the approach of the iterative expert survey as applied to estimating parties' policy positions. The initial motivation in doing so, was the difficulty of established approaches to estimate parties' positions in a valid and reliable way. Like other judgemental coding approaches to political text, and similar to conventional expert surveys, the iterative expert survey has the advantage of producing estimates that fare better in terms of validity compared to 'manifest' coding approaches (Gemenis and Dinas, 2010; Steenbergen and Marks, 2007). The proposed approach also deals effectively with the problem of inter-coder/expert disagreement via the structured behavioural aggregation, compared to the problematic way in which such disagreement is dealt with in conventional expert surveys (statistical aggregation), or other judgemental coding approaches (unstructured behavioural aggregation). In this respect, the presence of iteration and anonymity among panellists ensures higher inter-coder/expert agreement compared to both conventional expert surveys and content analysis approaches (Gemenis and Van Ham, 2014; Mikhaylov et al., 2012). The analysis on German parties tested the efficacy of the Delphi method, by empirically verifying the 'theory of errors' which purports that anonymous iteration can mitigate uncertainty and error in the context of expert judgement.

Future research should further explore the methodological premises of the iterative expert survey especially in the framework of randomized experiments that are becoming increasingly relevant in the context of estimating parties' positions (e.g. Albright and Mair, 2011; Lacewell and Werner, 2013; Tilley and Wlezien, 2008). Such research could increase our understanding of the iterative expert survey. In the meanwhile, the drawbacks of the proposed approach appear to be largely practical and refer to its costs and complexity. While the iterative expert survey is taxing on panellists' time and therefore can be costly if panellists are remunerated for their effort, the proposed approach is generally less costly and time-consuming compared to the hand-coding of party manifestos as exemplified by the Manifesto Project (Volkens, 2007). Moreover, the possibility to conduct an iterative expert survey using a specialized online platform could offset the additional effort needed to conduct estimation over multiple rounds.

What could then be the practical uses of the iterative expert survey method presented in this paper? I contend that the iterative expert survey method proposed here is unlikely to be adopted by either the proponents of the conventional expert surveys, or content analysis, since the research teams of both the Chapel Hill Expert Survey and Manifesto Project have invested considerable resources over the years to keep their datasets comparable over time. There are, nevertheless, two areas where the proposed approach could make an impact.

Firstly, the iterative expert survey can be used in judgemental coding approaches in the context of voting advice applications. Currently used approaches, such as the *Kieskompas* stress the advantage of factoring in estimation the self-placement of parties, although their cooperation cannot be guaranteed. The proposed method does not rely on the cooperation of parties, although, if available, this information can be included in the iterative expert survey after the first round. Moreover, the iteration proposed here, will increase the confidence and reliability of estimated positions, regardless of whether a party has provided self-placement information. I therefore propose that voting advice applications designers should seriously consider using the iterative expert survey approach.

Secondly, the judgemental coding approach within the iterative framework is not generally prone to the uncertainty and telescoping effects associated with retrospective application of conventional expert surveys (Budge, 2000; Ruedin, 2013, p. 96; Steenbergen and Marks, 2007, p. 347). This opens an enormous potential for applying the proposed approach to estimate parties' policy positions *retrospectively.* In this respect, the iterative expert survey can be used to measure parties' positions on broad ideological dimensions that are more often the focus of interest among political scientists. This can be done either by eliciting panellists' responses on broad ideological statements such as those employed by conventional expert surveys, or scale the narrow policy statements into more general dimensions using the full multi-round information on uncertainty, such as the information on panellists' confidence and agreement (see Shikano, 2013). Ultimately, the iterative expert survey would be used to generate a new longitudinal and cross-national dataset on parties' policy and ideological positions.

Acknowledgements

Earlier versions of this article were presented in a seminar at the Institute for Advanced Studies, Vienna (April 2014) and a VAA workshop at the University of Twente (November 2013). I would like to thank the participants of these events, and particularly Fernando Mendez, Emmanuel Sigalas, Markus Wagner, and Jonathan Wheatley, for their helpful comments. The usual disclaimer applies.

References

Albright, J. J., Mair, P.: Does the number of parties to place affect the placement of parties? Results from an expert survey experiment. Elect. Stud. 30, 858–864 (2011)

Arceneaux, K., Nickerson, D. W.: Modeling certainty with clustered data: A comparison of methods. Polit. Anal. 17, 177–190 (2006)

Armstrong, J. S.: How to make better forecasts and decisions: Avoid face-to-face meetings. Foresight 5, 3–8 (2006)

Benoit, K., Laver, M., Mikhaylov, S.: Treating words as data with error: Uncertainty in text statements of policy positions. Am. J. Polit. Sci. 53, 495–513 (2009)

Bolger, F., Wright, G.: Reliability and validity in expert judgment. In: Wright, G., Bolger, F. (Eds.) Expertise and Decision Support, pp. 47–76. Plenum, London (1992)

Booker, J. M., Meyer, M.A.: Sources and effects of interexpert correlation: An empirical study. IEEE T. Syst. Man Cyb. 18, 135–142 (1988)

Brambor, T., Clark, W. R., Golder, M.: Understanding interaction models: Improving empirical analyses. Polit. Anal. 14, 63–82 (2006)

Braun, D., Mikhaylov, S., Schmitt, H.: EES (2009) Manifesto Study Documentation Advance Release 22/07/2010, www.piredeu.eu (2010)

Budge, I.: Expert judgements of party policy positions: Uses and limitations in political research. Eur. J. Polit. Res. 37, 103–113 (2000)

Conti, N., Memoli, V.: The multi-faceted nature of party-based Euroscepticism. Acta Polit. 47, 91–112 (2012)

Curini, L.: Experts' political preferences and their impact on ideological bias: An unfolding analysis based on a Benoit-Laver expert survey. Party Polit. 16, 299–321 (2010)

Dalkey, N. C.: Toward a theory of group estimation. In: Linstone, H.A, Turoff, M. (Eds.) The Delphi Method: Techniques and Applications, pp. 236–261. Addison Wesley, Reading, MA (1975)

Dalkey, N. C., Helmer, O.: An experimental application of the Delphi method to the use of experts. Manage. Sci. 9, 458–467 (1963)

Dalton, R. J., Beck, P. A., Huckfeldt, R.: Partisan cues and the media: Information flows in the 1992 presidential election. Am. Polit. Sci. Rev. 92, 111–126 (1998)

Dinas, E., Gemenis, K.: Measuring parties' ideological positions with manifesto data: A critical evaluation of the competing methods. Party Polit. 16, 427–450 (2010)

Djouvas, C., Gemenis, K., Mendez, F., Wheatley, J.: SmartCoding: An online platform for estimating parties' policy positions. (2014)

Dolezal, M.: The design of the study: The distinguishing characteristics of our approach. In: Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S., Frey, T. (Eds.) West European Politics in the Age of Globalization, pp. 53–74. Cambridge University Press, Cambridge (2008)

Dolezal, M., Ennser-Jedenastik, L., Müller, W. C., Winkler, A. K.: How parties compete for votes: A test of saliency theory. Eur. J. Polit. Res. 53, 57–76 (2014)

Downs, A.: An Economic Theory of Democracy. Harper and Brothers, New York (1957)

Einhorn, H. J.: Expert judgment: Some necessary conditions and an example. J. Appl. Psychol. 59, 562–571 (1974)

Ferrell, W. R.: Combining individual judgments. In: Wright, G. (Ed.) Behavioral Decision Making, pp. 111–145. Plenum Press, London (1985)

Garzia, D., Marschall, S. (eds.): Matching Voters with Parties and Candidates. ECPR Press, Colchester, England (2014)

Gemenis, K.: Proxy documents as a source of measurement error in the Comparative Manifestos Project. Elect. Stud. 31, 594–604 (2012a)

Gemenis, K.: A new approach for estimating parties' positions in voting advice applications. Paper presented at the XXVI Congress of the Italian Political Science Association, Rome, September 13–15 (2012b)

Gemenis, K.: What to do (and not to do) with the Comparative Manifestos Project data. Polit. Stud.-London, 61(S1), 23–43 (2013a)

Gemenis, K.: Estimating parties' policy positions through voting advice applications: Some methodological considerations. Acta Polit. 48, 268–295 (2013b)

Gemenis, K., Dinas, E.: Confrontation still? Examining parties' policy positions in Greece. Comparative Eur. Polit. 8, 179–201 (2010)

Gemenis, K., Van Ham, C.: Comparing methods for estimating parties' positions in voting advice applications. In Garzia, D., Marschall, S. (Eds.) Matching Voters with Parties and Candidates, pp. 33–47. ECPR Press, Colchester, England (2014)

Golder, M., Stramski, J.: Ideological congruence and electoral institutions. Am. J. Polit. Sci., 54, 90–106 (2010)

Grimmer, J., Stewart, B. M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Pol. Anal. 21, 267–297 (2013)

Gudbrandsen, F.: Partisan influence on immigration: The case of Norway. Scand. Polit. Stud. 33, 248–270 (2010)

Hawkins, K. A.: Is Chávez populist? Measuring populist discourse in comparative perspective. Comparative Polit. Stud. 42, 1040–1067 (2009)

Janda, K.: Political Parties: A Cross-National Survey. Free Press, New York (1980)

Jones, R. J. Jr.: The state of presidential election forecasting: The 2004 experience. Int. J. Forecasting 24, 310–321 (2008)

Katakis, I., Tsapatsoulis, N., Mendez, F., Triga, V., Djouvas, C.: Social Voting Advice Applications: Definitions, challenges, datasets and evaluation. IEEE T. Syst. Man Cyb. (2013) doi: 10.1109/TCYB.2013.2279019

Krippendorff, K.: Content Analysis: An Introduction to its Methodology, second edition. Sage, Thousand Oaks, CA (2004)

Krouwel, A., van Elfrinkhof, A.: Combining strengths of methods of party positioning to counter their weaknesses: The development of a new methodology to calibrate parties on issues and ideological dimensions. Qual. Quant. 48, 1455–1472 (2014)

Krouwel, A., Vitiello, T., Wall, M.: The practicalities of issuing vote advice: A new methodology for profiling and matching. Int. J. Electronic Governance 5, 223–243 (2012)

Krouwel, A., Vitiello, T., Wall, M.: Helping voters to vote? Dynamics created by VAAs in election campaigns between parties, voters and media. In Garzia, D., Marschall, S. (Eds.) Matching Voters with Parties and Candidates, pp. 67–78. ECPR Press, Colchester, England (2014)

Lacewell, O. P., Werner, A.: Coder training: Key to enhancing reliability and validity. In Volkens, A., Bara, J., Budge, I., McDonald, I., Klingemann, H.-D. (Eds.) Mapping Policy Preferences from Texts III, pp. 169–194. Oxford University Press, Oxford (2013)

de Lange, S.L.: A new winning formula? The programmatic appeal of the radical right. Party Polit. 13, 411–435 (2007)

Lau, L. J., Frey, B.: Ideology, public approval, and government behavior. Public Choice, 10, 21–40 (1971)

Laver, M.: Position and salience in the policies of political actors. In: Laver, M. (ed.) Estimating the Policy Position of Political Actors, pp. 66–75. Routledge, London (2001)

Lewis, J. B., King, G.: No evidence on directional vs. proximity voting, Polit. Anal., 8, 21–33 (1999)

Lowe, W.: Understanding Wordscores. Polit. Anal. 16, 356–371 (2008)

Mair, P.: Searching for the positions of political actors: A review of approaches and a critical evaluation of expert surveys. In: Laver, M. (ed.) Estimating the Policy Position of Political Actors, pp. 10–30, Routledge, London (2001)

Meyer, M. A., Booker,J. M.: Eliciting and Analyzing Expert Judgment: A Practical Guide. Academic Press, London (1991)

Mikhaylov, S., Laver, M., Benoit, K.: Coder reliability and misclassification in the human coding of party manifestos. Polit. Anal. 20, 78–91 (2012)

Mulgrave, N. W., Ducanis, A. J.: Propensity to change responses in a Delphi round as a function of dogmatism. In: Linstone, H.A, Turoff, M. (Eds.) The Delphi Method: Techniques and Applications, pp. 288–290. Addison Wesley, Reading, MA (1975)

Müller, W. C., Strøm, K.: Coalition governance in Western Europe: An introduction. In: Müller, W. C., Strøm, K. (eds.) Coalition Governments in Western Europe, pp. 1–31. Oxford University Press, Oxford (2003)

Mumpower, J. L., Stewart, T. R.: Expert judgement and expert disagreement. Think. Reasoning 2, 191–212 (1996)

Neuendorf, K. A.: The Content Analysis Guidebook. Sage, Thousand Oaks, CA (2002)

Odmalm, P.: Party competition and positions on immigration: Strategic advantages and spatial locations. Comparative Eur. Polit. 10, 1–22 (2012)

Parenté, F. J.,Anderson-Parenté, J. K.: Delphi inquiry systems. In: Wright, G., Ayton, P. (Eds.) Judgmental Forecasting, pp. 129–156. John Wiley & Sons, New York (1987)

Pellikaan, H., Van der Meer, T., de Lange, S. L.: The road from a depoliticized to a centrifugal democracy. Acta Polit. 38, 23–49 (2003)

Riffe, D., Lacy, S., Fico, F. G.: Analyzing Media Messages: Using Quantitative Content Analysis in Research. Lawrence Erlbaum Associates, Mahwah, NJ (2005)

Rowe, G.: Perspectives on expertise in the aggregation of judgments. In: Wright, G., Bolger, F. (eds.) Expertise and Decision Support, pp. 155–180. Plenum, New York (1992)

Rowe, G., Wright, G.: The Delphi technique as a forecasting tool: Issues and analysis. Int. J. Forecasting 15, 353–375 (1999)

Ruedin, D.: Obtaining party positions on immigration in Switzerland: Comparing different methods. Swiss Polit. Sci. Rev. 19, 84–105 (2013)

Scheibe, M., Skutsch, M., Schofer, J.: Experiments in Delphi methodology. In: Linstone, H.A, Turoff, M. (Eds.) The Delphi Method: Techniques and Applications, pp. 262–287. Addison Wesley, Reading, MA (1975)

Shanteau, J., Hall, B.: What does it mean when experts disagree. In: Salas, E., Klein, G. (Eds.) Linking Expertise and Naturalistic Decision Making, pp. 229–244. Lawrence Erlbaum Associates, Mahwah, NJ (2001)

Shikano, S.: Estimating ideological positions of political parties using a deliberative expert survey. Paper presented at the Voting Advice Applications Workshop, University of Twente, The Netherlands, November 15 (2013)

Shikano, S., Busemeyer, M., Djouvas, C., Elff, M., Gemenis, K., Mendez, F., Wheatley, J.: ParteieNavi: Voting Advice Application for the German Federal Election 2013 [electronic file]. GESIS-Leibniz Institute for the Social Sciences [distributor] (2014)

Steenbergen, M. R., Marks, G.: Evaluating expert judgments. Eur. J. Polit. Res. 46, 347–366 (2007)

Stokes, D. E.: Spatial models of party competition, Am. Polit. Sci. Rev., 57, 368–377 (1963)

De Swaan, A.: An empirical model of coalition formation as an n-person game of policy distance minimization. In: Groenings, S., Kelley, E. W., Leiserson, M. (eds.) The Study of Coalition Behavior, pp. 424–444. Holt, Rinehart and Winston. New York (1970)

Tilley, J., Wlezien, C.: Does political information matter? An experimental test relating to party positions on Europe. Polit. Stud.-London 56, 192–214 (2008)

Trechsel, A. H., Mair, P.: When parties (also) position themselves: An introduction to the EU Profiler. J. Inf. Technol. Polit. 8, 1–20 (2011)

Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science 185, 1124–1131 (1974)

Volkens, A.: Strengths and weaknesses of approaches to measuring policy positions of parties. Elect. Stud. 26, 108–120 (2007)

Walter, A. S., van der Brug, W., van Praag, P.: When the stakes are high: Party competition and negative campaigning. Comparative Polit. Stud. (2013) doi: 10.1177/0010414013488543

Zulianello, M.: Analyzing party competition through the comparative manifesto data: Some theoretical and methodological considerations. Qual. Quant. 48, 1723–1737 (2014)