Original Article

Estimating parties' policy positions through voting advice applications: Some methodological considerations

Kostas Gemenis
Department of Public Administration, University of Twente, PO Box 217, Enschede 7500 AE, The Netherlands.
E-mail: k.gemenis@utwente.nl

Abstract The past few years have seen the advent and proliferation of Voting Advice (or Aid) Applications (VAAs), which offer voting advice on the basis of calculating the ideological congruence between citizens and political actors. Although VAA data have often been used to test many empirical questions regarding voting behaviour and political participation, we know little about the approaches used by VAAs to estimate the positions of political parties. This article presents the most common aspects of the VAA approach and examines some methodological issues regarding the phrasing of statements, the format of response scales, the reliability of coding statements into response scales and the reliability and validity of scaling items into dimensions. The article argues that VAAs have a lot of potential but there is also much space for methodological improvements, and therefore concludes with some recommendations for designing VAAs.

Acta Politica (2013) **48,** 268–295. doi:10.1057/ap.2012.36;
published online 28 December 2012

Keywords: parties; policy positions; Likert scales; validity; reliability

Introduction

The advent and proliferation of Voting Advice (or Aid) Applications (VAAs), which offer voting advice on the basis of calculating the ideological congruence between citizens and political actors, have resulted to the proliferation of numerous estimates of the positions of political parties in several different electoral contexts (local, regional, national and supranational) across Europe. The proliferation of these data has effectively enabled political scientists to put many empirical questions under rigorous tests. In particular, researchers investigated whether VAAs influence voting behaviour (Marschall and

Schmidt, 2008; Walgrave *et al*, 2008; Garzia, 2010), participation and voter turnout (Fivaz and Nadig, 2010; Hartwig, 2010; Ladner and Pianzola, 2010), and have used VAA data to test questions relating to voting behaviour, pledge fulfilment, policy congruence and the dimensionality of political space (Schwarz *et al*, 2011; Talonen and Sulkava, 2011; Katsanidou and Lefkofridi, 2012; Wheatley *et al*, 2012).

Although much has been written about the effects of VAAs, little is known about how VAAs estimate the position of political parties and political actors, in general. Many of the empirical applications assume that the VAA data on parties' positions are uncontroversial and do not examine their validity and reliability. In this sense, many third-party users consider the VAA data as 'given' or at least collected through some objective process. In particular, researchers have devised numerous methods of doing so, which include survey-based approaches, the analysis or roll-call voting and the analysis of political text. Each of these approaches uses different data sources and makes different implicit or explicit assumptions regarding the dimensionality of political space, whereas some methods are considered to be more suitable than others depending on the political context and the particular research question. These differences notwithstanding, there is no apparent consensus regarding the validity and reliability of party position estimates when alternative methods are compared with one another (Volkens, 2007; Dinas and Gemenis, 2010).

Most VAAs follow an alternative to the aforementioned approaches that is based on the construction of Likert scales and, with respect to data collection, includes elements from content analysis. This is, in many respects, an innovative approach that has a lot of potential; however, as this article argues, we need to consider many methodologically controversial aspects of the estimation process in the context of VAAs. Nevertheless, failure to do so can result in a proliferation of unreliable and invalid data on party positions with potential consequences, not only to the accuracy of our inferences about how democracy works but also to the democratic process itself (see Walgrave *et al*, 2009; Ramonaitė, 2010). In the following section, I outline the most common features of the VAA approach to estimating parties' positions, while specific aspects of this approach are examined critically in the respective sections. To reinforce the constructive aspect of this article, the concluding section offers some suggestions to VAA designers regarding improving the validity and reliability of party position estimates in the context of VAAs.

How VAAs Position Political Parties

With the exception of a few instances where the positions of political actors are based on estimates from previously conducted expert surveys (see Wall *et al*, 2009) or roll-call voting records (see, Škop, 2010), most VAAs have adopted the summated rating scale approach to measurement, even though this is not always mentioned explicitly in the VAA websites. In essence, most VAAs use the approach developed by Rensis Likert for psychological attitude measurement in the early 1930s. According to Likert (1932), the measurement of attitudes can be achieved through measuring responses on individual items that are representative of the attitude in question. Each item, as proposed by Likert, consists of a stem statement and a response key, which indicates the respondent's attitude towards the stem statement. When these items are combined, the resulting measure is a Likert scale. The idea behind Likert scales is that each individual item by itself is an imperfect indicator of the latent concept (in this case, party policy or ideology). Likert scales aim to fully capture the latent concept by combining responses from multiple items.

The advantage of simplicity and ease of use has established Likert scales as the *prima facie* technique for psychological attitude measurement. Nevertheless, with some exceptions (Pellikaan *et al*, 2003; Kriesi *et al*, 2006; Gemenis and Dinas, 2010), until the proliferation of VAAs, the use of Likert scales in measuring parties' positions was limited. Most expert, elite and mass surveys ask respondents to place parties directly on interval scales, which represent the latent dimensions of interest. Moreover, the assumption that political actors are able to take explicit in favour, against or neutral positions on political issues sets the Likert scaling approach apart from the approaches such as the Comparative Manifestos Project (Budge, 2001) and *Wordscores* (Laver *et al*, 2003), which attempt to estimate positions on the basis of the relative occurrence of certain type of words or quasi-sentences in political text. It should be rather evident that the relative emphasis political parties place on certain words or quasi-sentences associated with particular issues may or may not reveal their positions on these issues (Laver, 2001). Words or quasi-sentence frequencies can be a useful way to measure issue salience but can be problematic in the context of measuring positions (Lowe, 2008; Dinas and Gemenis, 2010). Conversely, Likert scales are concerned with measuring positions, but cannot measure the salience or the importance that parties attach to the issues in question.

However, as Walgrave *et al* (2009, pp. 1161–1162) noted, VAAs differ with regard to their assumptions on the dimensionality of political space and on how issue statements and policy positions are determined, weighted and presented. Nevertheless, in contrast to the approach of estimating the positions of political actors on the latent (and unobservable) ideological dimension of interest directly, most VAAs use the logic of Likert scaling. VAAs position political actors on a number of issues (Likert items) that can be considered to be partial and imperfect indicators of the latent dimension(s) and subsequently average positions across issues in order to achieve a positional estimate on the latent ideological dimension(s) (Likert scales). In addition, VAAs use the same items and scales to position political actors and voters, and therefore are in a position to claim that they are able to 'advice' voters as to which party or candidate lies closer to their views.

The advantages of the Likert scale compared with survey approaches and the salience-based content analysis notwithstanding, it should be stressed that no method of estimating the positions of political actors should be viewed as uncontroversial. Likert scales are quite straightforward to design but are often misused. To this view, this article critically reviews three aspects in the use of Likert scales by VAAs and examines whether the party positions given by VAAs can be considered to be valid and reliable. First, the article considers issues regarding the phrasing of stem statements and the selection of response keys. Second, as it is not always feasible (or advisable) for researchers to ask political parties to position themselves, researchers need to code the actors' policy statements into the response keys. The article therefore explores the issue of intercoder reliability. Finally, the article explores the practice of constructing Likert scales on the basis of Likert items in the context of VAAs. The following sections explore these three aspects by using the data primarily from 'EU Profiler', a cross-national VAA used for the 2009 elections to the European Parliament (EP).

Stem Statement and Response Key Issues

One of the first things to consider when designing a VAA (or any type of survey for that matter) is making sure that the question wording and answer format have been used appropriately and in a consistent manner. Likert scales are no exception as Likert items feature stem statements akin to survey questions and response keys in varying formats (Johns, 2010, p. 3). Textbooks on asking survey questions have long advised against the use of the so-called double-barrelled questions (Oppenheim, 1992, pp. 126–128; Spector, 1992, pp. 23–24; Bradburn *et al*, 2004, pp.142–144; DeVellis, 2011, pp.82–83), that is, questions that ask for opinion on two different things. Despite this advice, VAAs often use double-barrelled questions in their stem statements. Consider the following statement found in 'Help Me Vote' VAA in Greece: 'The electoral system must be changed and single-member constituencies should be established in order to strengthen the relationship between citizens and MPs and eliminate the need for "campaign spending" '. This statement contains at least two questions: (a) whether the respondent feels that the electoral system should change, and (b) whether single-member constituencies should be established. It is rather obvious that a candidate, party or voter may agree with (a) but disagree with (b), perhaps because the latter is most often than not associated with non-proportional electoral systems. Moreover, it is not clear whether (c) strengthening the relationship between citizens and MPs, and (d) eliminating the need for 'campaign spending' are always a consequence of (a) and (b). A candidate, party or voter might feel that there might be a need for (c) but not for (d); think that (a) is good for (c) but not for (d); think that (b) is good for (d) but not for (a) and so on.

Although the mentioned example from 'Help Me Vote' is rather unfortunate, it is quite easy for double-barrelled questions to slip in VAA stem statements. Consider the statements in the most prominent VAA, the EU Profiler. As shown in Table 1, EU Profiler includes many double-barrelled statements, even though this VAA has been designed by using the expertise of a large team of experienced researchers. The problem of double-barrelled questions seems to be stemming from two different problems. First, when formulating questions in VAAs, many designers feel the need to present policy alternatives in terms of trade-offs. This is a legitimate concern when asking questions about spending preferences, because asking questions that are not framed in terms of trade-offs has been shown to elicit inconsistent responses where 'majorities wish to spend more money on just about everything' (Hansen, 1998, p. 514). This implies that although the EU Profiler double-barrelled statements 1 and 11 may be accepted on the basis of the need to ask a single question about two different things in order to outline the trade-off that would lead to consistent responses about people's (and parties) fiscal attitudes, the inclusion of other double-barrelled questions (such as statement 4 on immigration/economic growth) cannot be sustained on similar methodological grounds.

Double-barrelled questions may also arise from the designers' need of researchers to adapt the VAA statements to very specific electoral contexts. This is shown in Appendix A that contains a non-exhaustive list of recent

Table 1: Double-barrelled stem statements in the EU Profiler

Item	*Stem statement*
1	*Social programmes* should be maintained even at the cost of *higher taxes*
4	*Immigration policies* oriented towards skilled workers should be encouraged as a means of fostering *economic growth*
11	Government spending should be reduced in order to lower taxes
14	Governments should reduce *workers' protection regulations* in order to fight *unemployment*
16	*Renewable sources of energy* (eg., solar or wind energy) should be supported even if this means higher *energy costs*
17	The promotion of *public transport* should be fostered through *green taxes* (eg., road taxing)
18	Policies to fight *global warming* should be encouraged even if it hampers *economic growth* or *employment*

Note: The elements of double-barrelled statements are in italics.

examples from VAAs in countries such as Austria, The Netherlands, Switzerland and the United Kingdom. Apparently, the debate in the United Kingdom regarding the spending cuts evolved around different questions: whether the budget deficit should be halved (or reduced by another fraction); whether this should be done by means of cuts in public spending (or increasing taxes); whether this should be done within 4 years (or a longer period); and whether the cuts should be equally distributed among the different government departments (or distributed unequally). Researchers may feel the urge to compress as much information within a single VAA statement in order to present the 'whole debate', but this urge should be resisted. As advised in all the classic survey design textbooks, researchers can opt for transforming the double-barrelled questions in two (or more) separate questions. In the Vote Match example, researchers could have separated the two double-barrelled questions into four different statements.

Similar to double-barrelled questions and statements that introduce a certain qualification (Bradburn *et al*, 2004, pp. 144–145) can stem from the researchers' need to define the context of the debate. Consider the following statement in 2008 *Wahlkabine*: 'should important decisions taken at the EU level (such as Turkey's accession, reform treaty …) be subject to a referendum in Austria, even if such a referendum is not held across the EU?' Apparently, the VAA designers entered 'Turkey's accession' as an example to aid respondents (and expert coders of party positions) in defining 'important decisions'. By doing so, however, they introduced a qualification that transformed the statement into a which could lead respondents (and coders) towards a specific response. In the particular example, this implies that responses to the particular question (arguably about attitudes towards European integration) could be contaminated by attitudes regarding Turkey's accession.

The presence of 'quantitative statements' is also something that could be avoided in Likert item stem statements (Johns, 2010, p. 4), but several such statements can be found in VAAs. Appendix A gives some examples from some recent VAAs. For instance, 'criminals should be punished more severely' (statement 20 in EU Profiler) is a good example of a quantitative statement. The 'more' qualification reveals that the statement is implicitly trying to measure some sort of quantity. Does the respondent who disagrees with the statements about criminals being punished more severely think that criminals should be punished less severely or as severely as they are currently punished? As Johns (2010, p. 4) put it, 'the key point is that Likert items are intended to capture the extent of agreement or disagreement with an idea, and not to measure some sort of quantity or "hidden variable"'. The problem is that, by making a quantitative statement, researchers superimpose a dimension suggested by the stem ('more'/'less') over the dimension suggested by the response key ('agree'/'disagree'). This violates the equivalence between the response dimension and the underlying dimension (Krosnick and Presser, 2010, p. 278) and 'measurement is contaminated by the existence of multiple dimensions' (Hodge and Gillespie, 2004, p. 56). Moreover, the use of quantitative statements requires increased cognitive effort from the part of the respondent or the researcher who will be engaged in coding actor's positions (Krosnick and Presser, 2010, pp. 265–266).

Stem statements that are too general do not make very good Likert scale items either. As outlined in the previous section, the measurement theory behind Likert scales assumes that the statements used for constructing the scale are imperfect indicators of a latent unobservable dimension. Therefore, stem statements should be selected within this context. The items within a summated rating scale should measure attitudes on different aspects of the latent dimension and not measure attitudes directly on the latent dimension by using general encompassing statements. If the goal was to capture attitudes using a single question, researchers would use interval level scales, such as the ones used in expert, elite and mass surveys and not binary or ordinal response keys. This means that although 'European integration is a good thing' (statement 23 in EU Profiler stem) is appropriate in the context of expert, elite and mass surveys as it purports to measure the latent dimension directly, it is not appropriate in the context of a Likert scale. The fact that such general statements can capture a great deal of the latent dimension is shown in Table 2, where I correlate each of the items of the EU Profiler European integration scale against an encompassing measure that can be considered to be indicative of the latent dimension (expert survey estimates from Hooghe *et al*, 2010). It turns out that the item associated with the most specific stem statement (12) has the smallest correlation coefficient, whereas the most general (23) has the largest, almost as large as the encompassing scale, which includes all items in Table 2.

Table 2: Correlations between expert survey estimates and EU Profiler items

Item	*Statement*	*r*	*n*
12	The EU should acquire its own tax-raising powers	0.361	111
21	On foreign policy issues, such as the relationship with Russia, the EU should speak with one voice	0.703	132
22	The European Union should strengthen its security and defence policy	0.743	131
23	European integration is a good thing	0.809	140
24	[Country] is much better off in the EU than outside it	0.774	139
26	The European Parliament should be given more powers	0.552	123
27	Individual member states of the EU should have less veto power	0.762	113
	Scale	0.833	143

Note: Pearson's product–moment correlation coefficients with $P<0.001$.

In effect, using statements that are too specific in conjunction with statements that are too general works against the idea of a Likert scale in which the same weight is attached to each and every item in the scale.

Considering the response keys of Likert items, I examine three issues. The number of options in the response key, whether neutral midpoints are offered and whether reference points are available. The response keys in EU Profiler and other VAAs in the *Kieskompas* family are the classic 5-point Likert items with a neutral midpoint, whereas the *StemWijzer* family of VAAs features response keys with agree/disagree and 'no opinion' options. Likert originally suggested using 5-point response scales as it would be difficult to position an actor who might only be moderately against or in favour of a particular statement (Krosnick and Presser, 2010, p. 269). As the task of the respondents becomes easier when labels are used on the response key options (Krosnick and Presser, 2010, p. 271), the classic labelling in 5-point response keys follows variations along the 'completely disagree/tend to agree/neutral/tend to agree/completely agree' pattern of responses. However, such labelling confounds direction (agree/disagree) with intensity (completely/tend to) by asking respondents (and coders) to think along multiple dimensions, which can lead to measurement contamination (Hodge and Gillespie, 2004, pp. 56–57), especially when the response key is associated with a 'quantitative statement' as argued above. Although 5-point scales may outperform 3-point response keys on some aspects (Kleinnijenhuis and Krouwel, 2008), research has shown that the direction/intensity confounding occurs fairly often in Likert items (Duncan and Stenbeck, 1987). In addition, if 5-point response keys offer more choices, 3-point response keys might be preferable in terms of intercoder reliability (see below). This means that the choice between few or more response choices is essentially a trade-off between validity and reliability (Krosnick and Presser, 2010, pp. 272–273), although it is not entirely clear whether the response keys in VAAs are selected on the basis of such considerations or simply conform to the format of the 'parent' VAA.

Turning to whether a response scale midpoint is offered, there seems to be a near consensus, as nearly all VAAs feature a neutral midpoint ('neither agree/disagree', 'neutral' or 'open-minded'). Much research has shown that offering midpoints can be advantageous in terms of validity and reliability (Krosnick and Presser, 2010, p. 274). Nevertheless, research has also shown that the midpoint is often used by respondents as a 'don't know' option or 'as a safe haven by a "silent minority", taking refuge in that option rather than confessing to an unpopular viewpoint' (Johns, 2005, p. 237). This problem is particularly acute when parties attempt to manipulate their placement in order to gain more voting recommendations (see below). For such reasons, it is advisable to offer a neutral midpoint only in statements associated with less prominent issues where parties are expected to take a neutral position (Johns, 2005).

A related problem is the averaging of items in Likert scales that makes non-response indistinguishable from truly neutral positions. The EU Profiler follows the practice of combining the Likert items to scales by averaging, but research has highlighted the problems associated with this approach (Gemenis and Dinas, 2010, pp. 184–185). Monte Carlo simulations showed that replacing missing values by the response key midpoint should be used only when 'there is consistency in scores for an item across observations' (McDonald *et al*, 2000, p. 89). Such consistency should not be assumed when units are political parties, which represent varied response combinations, and therefore it is advisable to calculate parties' positions using scales where missing values are excluded rather than imputed (cf. Gemenis and Dinas, 2010, p. 194).

Finally, we need to consider whether response scales offer a reference point for the *status quo*. Pellikaan *et al* (2003, p. 34) have argued that the placement of parties by using Likert scales should be made in relationship to a reference point indicating the *status quo*. None of the most prominent VAAs, however, has used such reference points explicitly, although sometimes stem statements include some contextual information. Respondents in surveys have been shown to rely on informational cues (Krosnick, 1991) and Gschwend and Proksch (2010) found that indicating the *status quo* in the response key significantly reduced item non-response, changed respondents' self-placement, changed respondents' placement of political parties and changed the perception of ideological distance between voters and their preferred parties. To the extent that the measurement of parties' positions depends on the coding of various sources by a group of experts (see below), the inclusion of the *status quo* can serve as an additional piece of information, which may generate more variation in party positions that is otherwise possible (Gschwend and Proksch, 2010, p. 2).

Coding Reliability Issues

Notwithstanding the few VAAs that use expert surveys or roll-call voting records, most VAAs estimate the policy positions of parties on the basis of expert coding of party manifestos and other publicly available statements, on the basis of elite surveys using responses provided by the parties or through a combination of the latter two approaches. The advantage of elite surveys (that is, asking parties to self-placements) is that parties are positioned by using the same procedure that applies to voters. Simple and desirable as this might sound, there are several problems with this approach. First, there are many parties and candidates who refuse to position themselves either because they are uninterested in VAAs or because they do not agree with the way the questions are phrased (Trechsel and Mair, 2011, pp. 13–16; Nezi *et al*, 2010, p. 81). This is not new, of course, as elite surveys have been traditionally plagued by low response rates.

In such cases, researchers have no other solution than place the parties on the basis of political text (manifestos, press releases, green/white papers, statements in websites, speeches, parliamentary debates, interviews and so on).

Second, as VAAs are becoming increasingly popular, political parties have incentives to manipulate their stated positions in order to present an image that will be appealing to voters (Wagner and Ruusuvirta, 2012, p. 406). Politicians may wish to duck some controversial issues, or appear as moderate and centrist, even if they are generally considered to have extreme views. Evidence from the Lithuanian *Mano Balsas* revealed that the VAA was strategically manipulated and the voting recommendations gave an advantage to non-programmatic parties (Ramonaitė, 2010). A new party that was not constrained by a detailed electoral programme seems to have instructed its candidates to answer the VAA questions in a specific way in order to attain the inconsistent, but yet very popular among voters' placement indicating support towards the increase of social services and reduction of taxes (Ramonaitė, 2010, pp. 134–135). Of course, researchers are well aware of this possibility. When the EU Profiler country teams asked national parties to complete the VAA questionnaire, many parties adopted positions which came in conict with what VAA researchers thought to be their 'true' positions, on the basis of their public statements. Political parties that expected to earn votes through VAAs cooperated with the researchers, whereas those that did not expect any gains threatened with legal action (Trechsel and Mair, 2011, pp. 13–15). These anecdotal findings are rather illustrative of what could happen if we let political parties position themselves at will.

The alternative approach of positioning parties on the basis of political text is bound with problems of coding reliability (Groot, 2003, pp. 23–24). Essentially, for each item in a VAA, expert coders are asked to do two different things: (a) locate the party manifesto, or some statement that presents the party's position, and (b) use it to code each stem statement on a 3- or 5-point response key. In turn, both steps can introduce unreliability in the coding of parties' positions. It is well known that many parties do not publish manifestos (Gemenis, 2012), whereas some parties do not have very clearly stated positions on all issues that are included in VAAs. Therefore, party position estimates in VAAs are partially dependent on how 'deep' researchers are willing to search in order to find some evidence, which can be used for coding. EU Profiler has issued a coding manual that outlined the 'hierarchy of documents' that can be used for coding. However, in the absence of documents at the top, expert coders are left to decide themselves about how much further in the hierarchy of documents to go. Moreover, interviews and statements made during parliamentary debate appear at the bottom of the hierarchy, but it is still questionable whether such evidence can be considered as representative under the party as a unitary actor assumption (Gemenis, 2012). Equally controversial is using evidence from legislative voting in order to gauge parties' positions, as it is well established that voting patterns often reflect party discipline and government–opposition dynamics, which implies that roll-calls cannot always be considered as parties' ideal point estimates (Spirling and McLean, 2007).

Apart from variability in the coding sources, the double-barrelled or vague wording of some of the stem statements in VAAs leaves open the possibility for alternative interpretations regarding the coding of text into the categories of the response key. Again, consider item 23 in EU Profiler: 'European integration is a good thing'. It is rather obvious that any positive statement towards the EU can be used to code a party as agreeing or completely agreeing. Another example from EU Profiler is item 26: 'The European Parliament should be given more powers'. What kind of powers this statement refers to? Are they policy-making powers, or just some 'oversight of the work of the EU', the statement that was used to code the Green Party of England and Wales as 'completely agree'. Does arguing that policy making should be transferred from the national parliaments to the European Parliament be equated with the aforementioned statement of the Green Party? And how do we deal with parties that are for transferring some policy-making powers (say environmental policy) but against transferring others?

Experts coders involved in VAAs are essentially asked to content analyse statements on the basis of their expertise and (sometimes) the specific instructions of a codebook. In the context of content analysis, coding reliability is usually measured by the degree of agreement among coders working independent of one another. Put otherwise, we can infer coding reliability by measuring coder agreement (Krippendorff, 2004, p. 414). Considering the problems outlined above, it is easy to expect that different coders will come up with different positions for the same political actors, therefore rendering the VAA approach to position political parties as unreliable. To date, however, most VAAs have not addressed the possibility of coding unreliability stemming from variability in document sources or the ambiguity of stem statements in a systematic way. As far as I know, VAAs do not routinely report intercoder reliability. EU Profiler, for instance, relies on 'discussions among team members' and consultations with experts and the VAA leadership (Trechsel and Mair, 2011). However, making decisions by consensus among the VAA team members does not guarantee coding reliability. 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 reect the social structure of the group'.

In this article, I present a first attempt to examine the intercoder reliability in the Likert items used by VAAs to position political parties. To do so, I use data from a coding exercise conducted at the University of Twente where 80 European Studies undergraduate students attending a course on conceptualisation and measurement agreed to participate after being offered partial course credit. The students were assigned to German ($n = 41$) and Dutch ($n = 39$) language groups on the basis of their native or uency of language, and were asked to code the Dutch and German parties on 16 selected EU Profiler statements using their 2009 EP election manifestos. The details of the exercise are presented in the online supplementary material of this article. In summary, the results in terms of Krippendorff's alpha, the statistic that is generally agreed to be the 'measure with appropriate reliability interpretations in content analysis' (Krippendorff, 2004, p. 221), were found to be unacceptably low for each of the 16 statements for the coding to be considered reliable even for drawing tentative conclusions. Nevertheless, because the placement of parties was not the result of a content analysis according to some specified coding unit (which is the focus of Krippendorff's alpha) but rather latent coding of party positions, I also assessed coding reliability using van der Eijk's (2001) coefficient of agreement, which has been devised to measure respondent agreement on ordinal rating scales. The latter assessment showed a more mixed picture of some of the statements scoring $A > 0.7$ on the agreement coefficient across parties. As expected, ambiguous statements such as 'the European Parliament should be given more powers', in which conflicting evidence can be taken into account, scored much lower in terms of coding reliability than specific statements such as the one referring to possibility of Turkey becoming a full member-state. In addition, statements referring to environmental policies fared generally worse than statements referring to European integration. This implies that coding reliability is also constrained by the type of documents used (cf. Gemenis, 2012). When lacking credible documents that explicitly refer to the statements used by VAAs, coders are in a much worse position in estimating the positions of parties.

The results presented here do not imply that the data of the EU Profiler are unreliable. The coding exercise did not attempt to replicate the precise conditions in which the EU Profiler coding took place. As mentioned above, the EU Profiler researchers do not work independently to one another, which means that the reliability of their coding work cannot be estimated; thus, the presented results should be seen as measures of the potential reliability for coding party positions by using Likert items. Designers of VAAs should be vigilant about the possibility that the estimation procedure might have low intercoder reliability, which in turn may compromise the validity (Krippendorff, 2004, p. 214) of party placements.

Scaling Issues

One of the features that arguably makes VAAs so popular is their ability to give voting recommendations on the basis of graphical representations of the congruence between voters and parties (or candidates). The *Kieskompas* family of VAAs, which includes the EU Profiler, routinely uses a two-dimensional space to present the positions of candidates and parties, whereas VAAs in the *Smartvote* family use radar (spider) charts that are able to represent positions in multiple policy dimensions. Even VAAs that do not provide graphical representations of the agreement between voters and parties often adopt spatial models for their voting recommendation calculations (Louwerse and Rosema, 2011, p. 2).

A crucial aspect of placing political actors on political space relates to the dimensionality of political space. Broadly speaking, political scientists have a choice between adopting a deductive and an inductive approach in defining dimensionality (Benoit and Laver, 2012). The purely deductive approach involves defining *a priori* the most important dimensions and then placing the actors on these dimensions. Expert surveys follow this deductive approach. The inductive approach involves selecting a number of items and using a scaling method to arrange the items in policy or ideological dimensions. The problem with the inductive approach, however, is that it is only quasi-inductive in the sense that the results depend largely on the quantity and type of issues that are used in the calculations (Benoit and Laver, 2012, pp. 207–208). This brings us back to the problem of item selection, a deductive exercise with important consequences for party placement. In a seminal methodological article, Walgrave *et al* (2009) demonstrated how the selection of different sets of statements (that is, Likert items) in VAAs can lead not only to different estimates of the positions of parties, but also to different levels of best matches among voters. The latter finding implies that, in some instances, some parties may find themselves in the beneficial position of receiving more voting recommendations because of artefacts in the VAA construction. Taking the EU Profiler as an illustrative case, the remainder of this section examines the reliability and validity of the scales used by VAAs to position parties.

EU Profiler has adopted the deductive approach of the *Kieskompas*. In particular, EU Profiler places parties on a two-dimensional political space, on the basis of previous literature that defined the EU polity as a two-dimensional political space based on socio-economic left–right (L–R) dimension and a pro/anti-EU integration (EU) dimension, which are orthogonal to one another (Hooghe *et al*, 2002; Hix *et al*, 2006). Each dimension includes a number of issues that have been identified (through consultations with experts and by looking at party manifestos) as being salient during the 2009 EP election campaign. Given that different items are 'fitted' into scales based on prior expectations about issue salience and their *a priori* fit, it would be interesting to examine the reliability and validity of the constructed EU Profiler scales.

Unfortunately, many VAA websites do not give precise details regarding which items are included in each scale. The EU Profiler is no exception as the detailed 'General Description and Methodology' document does not look at scale construction, although I was able to obtain the relevant information by contacting a member of the EU Profiler team. By using the provided information, I was able to fully replicate parties' scores on the EU scale, that is, the scores given in the EU Profiler data set (Trechsel, 2010), but I was unable to do so for the L–R scale. As party scores in my reconstruction of the L–R scale correlate highly with the scores in the original data set (0.961), save a single outlier and no bias is detected, I proceed into analysis with this small caveat in mind.

As researchers are often confronted with stationary measurement, assessing scale reliability is often associated with the split-half method, where a scale consisting of half of the original scale items is compared with the scale consisting of the other half. In this sense, the typical statistic used to assess the reliability of a Likert scale is Cronbach's alpha, which, as the original article showcased (Cronbach, 1951, p. 300), provides the 'average of all possible split-half coefficients' in a scale. Cronbach's alpha for the EU and L–R scales is 0.925 and 0.878, respectively, two values that indicate that the scales are reliable in terms of split-half reliability. Although the high Cronbach's alpha indicates a high split-half reliability, it is not a measure of internal consistency, homogeneity or unidimensionality of a scale (Sijtsma, 2009). In other words, a high alpha does not necessarily indicate that the items of the scale belong in the same dimension or that they are measuring attributes of a common latent variable (the latter refers to measurement validity). Assessing whether the items fit within a given scale has seen the development in many indices based on answer patterns, factor analysis, principal components or latent trait models (Hattie, 1985, pp. 141–142). Recognising that different methods might lead to different (and sometimes conicting) results, I first assess scale validity using an informal visualisation approach and then switch to inductive approaches (factor analysis and Mokken scaling). In addition, I validate the EU Profiler party scores against expert survey data.

The first approach looks at internal association by fitting a locally adjusted regression curve on scatter plots between each item and a scale consisting of all other items (Dinas and Gemenis, 2010; Gemenis and Dinas, 2010). I follow this approach because a central assumption of the Likert scale is that, every item in the scale is monotonically related to the underlying latent dimension (Jacoby, 1991, p. 40). This implies that the scores of an item that fit well into a scale should increase as the scores in the latent dimension increase and vice versa. In this case, the scores in the latent dimension are proxied by the values of the scale consisting of all other items. Nevertheless, because item-rest correlations measure linear relationships, while we are merely interested in monotonic relationships, I evaluate monotone homogeneity by using locally adjusted regression (loess) curves instead of item-rest correlations.

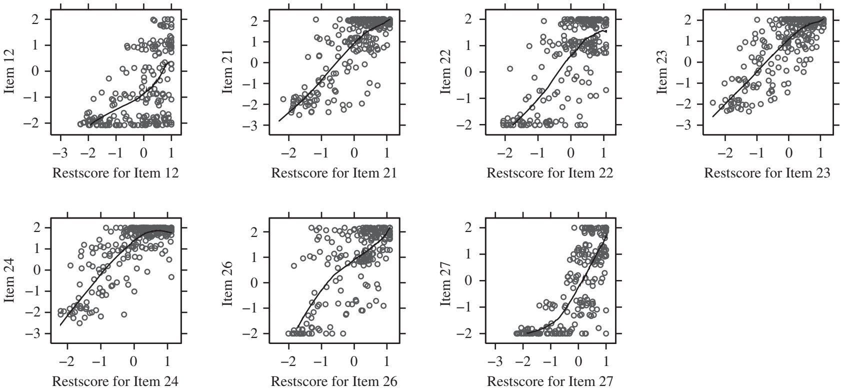

Figure 1: Examining the scalability of the items in the EU Profiler EU scale (loess bandwidth = 0.7; points jittered 5 per cent).

The scatter plots in Figure 1 show that every item in the EU scale is monotonically related to the respective rest scores, which imply a good item-scale fit. The situation is quite different for the L–R scale, however Figure 2 of the 18 items included in the scale, there are at least two items that do not seem to fit in the scale. Item 4 ('Immigration policies oriented towards skilled workers should be encouraged as a means of fostering economic growth') appears to have a relationship with the latent L–R dimension, which is clearly not monotonic. This possibly stems from the double-barrelled nature of the stem statement, which, as argued in Section 'Stem statement and response key issues', is particularly problematic as the two different issues introduced into the statement (immigration and economic growth) do not seem to have a clear trade-off relation to each other. The same goes for item 15, which refers to reducing the EU subsidies to Europe's farmers. This item refers to EU policy and does not sit comfortably in the L–R scale. In addition, items 8, 17 and 19 seem somewhat problematic, although the change in direction of the loess curves might be because of observations in specific countries, which would imply problems in cross-country comparability of the content and meaning of the L–R scale.

Seeing some problems in the construction of the L–R scale, I switch to a quasi-inductive examination of the EU Profiler items. I first do so by using exploratory factor analysis. As the individual items of a Likert scale 'are ordinal-level functions of the latent dimension' (Jacoby, 1991, p. 40), our factor analysis is based on a polychoric correlations matrix to account for the ordinal level measurement of the items. To avoid the problem of reducing the number of observations because of list wise deletion, I replace missing values by the neutral zero category. As argued in Section 'Stem statement and response key

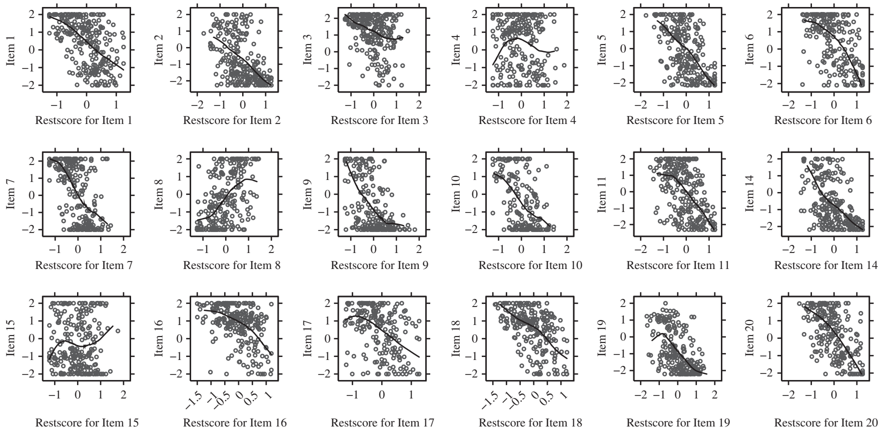

Figure 2: Examining the scalability of the items in the EU Profiler L–R scale (loess bandwidth = 0.7; points jittered 5 per cent).

issues', this is a controversial practice, but is in line with the EU Profiler practice where missing values are treated as neutral categories for computational reasons.

The results in Table 3 suggest that the political space for the 2009 EP elections can be best represented by four dimensions (corresponding to four factors with eigenvalues >1). All the items loading strongly on the first factor refer to the pro/anti-EU dimension. In fact, all the items of the EU Profiler EU scale load highly on the first factor, in addition to item 28 referring to the need for referendums in approving EU treaties. Turning to the remaining factors, it appears that the EU Profiler L–R scale is not unidimensional as previously assumed. Issues pertaining to 'authoritarian versus libertarian' and 'taxes versus spending' sub-dimensions of the L–R appear to have high loadings on the second and the third factors, respectively. The three environmental items together with one item about immigration load mostly on the fourth factor. Nonetheless, even one disregards the evidence in favour of a multidimensional measurement of the L–R, the factor analysis results confirmed the previous assessment that issues 4 and 15 should not have been included in the L–R scale. Finally, the item with the largest uniqueness in the data set is item 13, which refers to bank bailouts by governments. This reveals that it does not fit comfortably within any of the extracted factors, which makes sense given that the issue of bank bailouts entered the political agenda only very recently.

Furthermore, I considered that different items in the scale many represent different difficulty factors. For instance, a party may generally agree that 'European integration is a good thing' (item 23), but disagree that 'the EU

Table 3: Factor analysis of the EU Profiler items

Item	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*	*Uniqueness*
1	−0.142	0.218	0.755	0.184	0.329
2	0.128	−0.175	−0.775	−0.041	0.351
3	0.076	0.134	0.540	0.112	0.673
4	0.414	0.100	−0.357	0.384	0.544
5	−0.113	−0.466	−0.121	−0.668	0.310
6	0.058	−0.458	−0.241	−0.477	0.501
7	0.135	0.848	0.102	0.298	0.163
8	0.071	−0.706	−0.136	−0.119	0.463
9	0.057	0.749	0.226	0.029	0.385
10	0.147	0.791	0.123	−0.017	0.338
11	0.105	−0.278	−0.684	−0.158	0.418
12	0.577	0.240	0.286	0.076	0.522
13	0.455	−0.077	−0.066	−0.079	0.776
14	0.138	−0.154	−0.742	−0.239	0.349
15	−0.185	0.325	0.436	0.176	0.639
16	0.164	0.137	0.427	0.511	0.512
17	0.170	0.136	0.297	0.585	0.523
18	0.075	0.132	0.454	0.569	0.447
19	0.026	−0.515	−0.268	−0.427	0.480
20	−0.078	−0.549	−0.306	−0.426	0.418
21	0.902	0.064	−0.038	−0.009	0.181
22	0.834	−0.207	−0.140	−0.085	0.235
23	0.906	0.048	−0.049	0.080	0.168
24	0.888	0.019	−0.139	0.091	0.183
25	0.392	0.273	0.046	0.433	0.583
26	0.729	0.263	0.209	0.058	0.352
27	0.812	0.227	−0.016	0.119	0.275
28	−0.588	0.201	0.277	−0.225	0.487
% of variance explained	30.4	21.7	20.4	14.1	Total 86.6

Note: Analysis based on polychoric correlations matrix; varimax rotation; highlighted cells indicate loadings greater than 0.5.

should acquire its own tax-raising powers' (item 12). However, the reverse is rather unlikely. If a party would agree on the most difficult item ('tax raising powers'), it would be unlikely to disagree on the easy item (that European integration is generally desirable). To examine unidimensionality within this context, I use Mokken scale analysis (van Schuur, 2003), which is based on the idea of the Guttman scale but looks at the difficulty ordering of items in probabilistic terms. The selection of items into a scale in Mokken scale analysis is determined by Loevinger's coefficient of homogeneity *h* under a null hypothesis test 'that all item pairs are stochastically independent of one

Table 4: Mokken scale analysis of the EU Profiler items

Item	*Item selection with*												
	h>0.5					*h>0.4*					*h>0.3*		
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*
1*	—	0.64	—	—	—	—	0.59	—	—	—	—	0.50	—
2	—	—	—	—	—	—	0.48	—	—	—	—	0.40	—
3*	—	—	—	—	—	—	—	—	—	—	—	—	—
4*	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	0.61	—	—	—	—	0.54	—	—	—	0.44
6	—	—	—	0.56	—	—	—	—	0.57	—	—	—	0.41
7*	—	—	0.57	—	—	—	—	0.54	—	—	—	—	0.48
8	—	—	0.63	—	—	—	—	0.54	—	—	—	—	0.43
9*	—	—	0.58	—	—	—	—	0.54	—	—	—	—	0.45
10*	—	—	0.58	—	—	—	—	0.52	—	—	—	—	0.42
11	—	0.59	—	—	—	—	0.56	—	—	—	—	0.50	—
12	0.61	—	—	—	—	0.52	—	—	—	—	0.48	—	—
13	—	0.51	—	—	—	—	0.49	—	—	—	0.32	—	—
14	—	—	—	—	—	—	—	—	—	—	—	0.43	—
15*	—	—	—	—	—	—	—	—	—	—	—	—	—
16*	—	—	—	—	0.55	—	—	—	0.41	—	—	0.35	—
17*	—	—	—	—	—	—	—	—	—	0.51	—	—	—
18*	—	—	—	—	0.55	—	—	—	—	0.51	—	0.34	—
19	—	—	—	—	—	—	—	0.42	—	—	—	—	0.38
20	—	—	—	0.54	—	—	—	0.43	—	—	—	—	0.46
21	0.70	—	—	—	—	0.68	—	—	—	—	0.64	—	—
22	0.59	—	—	—	—	0.58	—	—	—	—	0.55	—	—
23	0.69	—	—	—	—	0.68	—	—	—	—	0.64	—	—
24	0.67	—	—	—	—	0.68	—	—	—	—	0.65	—	—
25*	—	—	—	—	—	—	—	—	0.49	—	—	—	0.36
26	0.53	—	—	—	—	0.50	—	—	—	—	0.48	—	—
27	0.68	—	—	—	—	0.62	—	—	—	—	0.58	—	—
28*	—	—	—	—	—	0.40	—	—	—	—	0.39	—	—
Scale *h*	0.64	0.58	0.59	0.57	0.55	0.59	0.53	0.51	0.51	0.51	0.54	0.43	0.43

Note: Items with * are reversed.

another' (van Schuur, 2003, p. 149). As a rule of thumb, the scales with $h > 0.3$ are considered acceptable, although strong scales should attain $h > 0.5$. As ordering in terms of difficulty implies that items are in the same direction, I recoded all items so that positive values would always reflect a right-wing or pro-EU position. I then performed a Mokken scale analysis by setting the threshold for item selection at 0.3, 0.4 and 0.5, which correspond to weak, moderate and strong scales, respectively.

The results in Table 4 largely confirm the results of the factor analysis. When the threshold for selecting pairs of items was set at $h > 0.5$, the Mokken scaling procedure resulted in five strong scales. The first scale includes all the items of the EU Profiler EU scale, the second includes three items on the 'taxes versus spending' dimension, the third includes all four items of the 'society, religion and culture' EU Profiler spider graph, the fourth includes three items about law and order and immigration and the fifth comprises two out of the three environmental policy items. By relaxing the threshold to $h > 0.4$, an additional item is included in each of the first two scales, without changing their interpretation. Moreover, two law and order items are added on the third scale effectively making it a 'authoritarian versus libertarian' scale, whereasile the fourth scale becomes more of a mixed bag. Finally, relaxing the threshold to $h > 0.3$ results in three scales: a slightly expanded EU scale with the addition of the rather irrelevant item on bank bailouts (which nevertheless has the lowest h coefficient in the scale) and two scales corresponding to the 'authoritarian versus libertarian' and 'taxes versus spending' sub-dimensions of the L–R. Finally, in none of the scaling procedures did items 3, 4 and 15 attain the minimum acceptable value of 0.3 on the h coefficient, which attests to their limited scalability, largely confirming the results obtained by factor analysis and the inspection of item-rest score scatter plots.

The final step in examining the EU Profiler scales involves validation against alternative measures. To do so, I compare the EU Profiler placements on the EU and L–R scales with the placements given by the 2006 Chapel Hill expert survey (Hooghe *et al*, 2010). As the Pearson product–moment correlation coefficient reflects dispersion rather than agreement between two measures, I use the concordance correlation coefficient ρ_c (Lin, 1989). The advantage of using ρ_c in assessing agreement between two methods of estimating parties' policy positions (Gemenis, 2012, p. 601). In essence, the concordance correlation coefficient is the product of two components: Pearson's r, which measures accuracy, and the bias correction factor C_b, which measures precision.

Table 5 presents the results of these comparisons. The results point out that the EU Profiler scales have a fairly good concordance with the respective expert survey scales when we consider that we should not expect a near perfect concordance between the two measures, given that the 2006 Chapel Hill expert survey was conducted nearly 3 years before the 2009 EP election, and that its EU scale measures party leader positions instead of party positions. A look at the components of the concordance correlation coefficient reveals that the

Table 5: Comparing the EU Profiler and expert survey party placements

Scale	*N*	ρ_c	*95% CI*	*r*	C_b
EU Profiler pro/anti-EU	143	0.826	[0.767, 0.872]	0.833	0.992
EU Profiler L–R	143	0.785	[0.720, 0.837]	0.842	0.933
Modified L–R	142	0.835	[0.780, 0.878]	0.859	0.972

Note: ρ_c: concordance correlation coefficient; 95% CI: z-transform 95% confidence interval; r: Pearson product–moment correlation coefficient; C_b: bias correction factor.

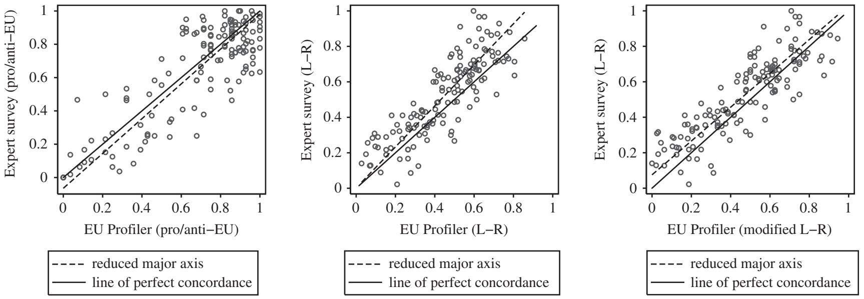

Figure 3: Comparing EU Profiler and expert survey party placements.

disagreement between EU Profiler and expert survey placements are mostly because of random measurement error. This is mostly true for the EU scale rather than the L–R, which has a lower C_b indicating some bias. Moreover, Figure 3 shows that this bias is not distributed evenly across the L–R scale. The slope of the reduced major axis on the middle scatter plot in Figure 3 indicates that extreme right parties (as placed by the experts) are more likely to be placed towards the centre by the EU Profiler compared with extreme left parties.

Could the EU Profiler L–R scale be improved in light of the analyses presented in this section? The last line in Table 5 and the right scatter plot in Figure 3 evaluate an alternative L–R scale on the basis of the EU Profiler data. Notwithstanding the evidence in favour of a multidimensional interpretation of the L–R scale given by the factor analysis and Mokken scale analysis, the modified L–R simply drops the items that were shown to be problematic in Figure 2 (items 3, 4, 8, 15 and 17). The results in Table 5 and Figure 3 point out that the modified scale reduces the centrist bias on the extreme right of the scale and results in greater concordance with expert surveys, although the overlapping confidence intervals indicate that this increase is not large enough to be statistically significant. However, the argument put forward by the analyses in this section is that the placement of political parties by VAAs can be improved by quasi-inductive techniques, reliability analysis and rigorous validation of the used scaling methods.

Conclusions

This article used examples primarily from EU Profiler to highlight some methodological issues faced by VAAs in estimating the positions of political parties. I argue that third-party users of VAA data should acknowledge these issues, although this article does not imply that VAAs are inherently problematic. To the contrary, VAAs have several attractive features and exhibit a lot of potential for creating data that could enable us to answer many important empirical questions in comparative politics. For instance, this article showed that despite all the aforementioned methodological concerns, when compared with the 2006 Chapel Hill expert survey, the EU Profiler party placements appear to be more valid in comparison with the placements derived from the Euromanifestos Project (cf. Veen, 2011, p. 280). Moreover, most of the issues identified in this article can be dealt with in a fairly straightforward manner. For this reason, the article concludes with a set of recommendations for VAA designers.

Regarding the phrasing of stem statements, VAA designers could avoid using double-barrelled and quantitative statements, as well as statements that are too general. As shown in the article, sometimes double-barrelled statements arise from the (legitimate) need to present fiscal attitudes as a trade-off between conflicting choices, but in other instances they should be avoided. In addition, giving responses on Likert item response keys on quantitative statements requires increased cognitive efforts, which may lead to satisficing (Krosnick, 1991), whereas vague statements may lead to coding unreliability. In terms of designing response scales, VAA designers could consider avoiding offering a midpoint for controversial questions, but consider offering a reference point that indicates the *status quo*. More generally, the findings from the survey design literature can be used to achieve best practice in the area of statement phrasing and response scale formulation.

Turning to the issue of coding statements into response scales, VAA designers can be more explicit in reporting coding reliability. This would require switching from the current practice of coding by ‘consensus’ to a coding procedure involving two rounds. In the first round, multiple coders would work on the same parties and statements independent of one another, and in the second round any inconsistencies among the coders could be solved by consensus. This two-round coding process would not only enable the calculation of meaningful intercoder reliability statistics, but also it would make ‘reasonable to assume that post-coding reconciliation improves the reliability of the data beyond the reliability of data generated by any one individual observer’ (Krippendorff, 2004, p. 219). In addition, VAAs, which make use of parties’ self-placements, could be more explicit about reporting the concordance between party self-placement and placement by the VAA research team. This way, we can have a measure of the degree to which parties may be trying to manipulate their placement in VAAs in order to achieve more voting recommendations. Coding reliability also relates to transparency and the need for research to be replicable. VAAs could make worthy investments by producing detailed coding handbooks, which can be made available publicly in their websites. In this respect, EU Profiler is a best practice example as it allows users to see the exact statements (and their source) that have been used to code each party position.

Finally, with regard to the assumptions about the dimensionality of political space and scale validity and reliability, VAA designers can combine *a priori* deductive reasoning with quasi-inductive statistical techniques as outlined in this article. More specifically, VAA designers can implement the informal (but informative) graphical approach outlined in the previous section in order to check how well individual items fit into the scales utilised by VAAs. Moreover, quasi-inductive techniques such as factor analysis and Mokken scale analysis can be used to improve the validity and reliability of the resulting scales (see, also Groot, 2003, pp. 26–28). Since different scaling methods may lead to different voting recommendations (Louwerse and Rosema, 2011), it is advisable to follow some of best practices outlined in the dimensionality and scaling literature.

Although VAAs are used by an increasing number of countries and in an increasing number of electoral contexts, the methodological discussions about regarding their use has been rather limited. I hope that this article will open a wider debate on how VAAs can estimate the positions of political actors in a valid and reliable way, and that VAA designers might consider some of the recommendations proposed here. Following methodological practices such as the ones outlined in this article in phrasing stem statements, constructing response keys, presenting coding decisions, reporting intercoder reliability, dealing with missing values and assessing scale validity and reliability will increase confidence in the use of VAAs, both from the perspective of the public and the perspective of third-party users of VAA data.

Acknowledgements

Earlier versions of this article were presented at the Annual Meeting of the Dutch and Flemish Political Science Associations (Amsterdam, 9–10 June 2011), and at the workshop on Political Parties, Issue Competition and Public Policy at SciencesPo (Paris, 30 March 2011). I would like to thank the participants of these meetings as well as Zoe Lefkofridi, Martin Rosema, Markus Wagner and the anonymous reviewers of the journal for their useful comments. André Krouwel kindly provided answers to my questions regarding the scales used by the EU Profiler. Any errors remain my responsibility. Supplementary material and replication files are available online through my departmental page.

About the Author

Kostas Gemenis is an Assistant Professor in the Department of Public Administration at the University of Twente. His work has been published in journals such as *Electoral Studies*, *Comparative European Politics*, *Party Politics* and the *European Political Science Review*.

References

Benoit, K. and Laver, M. (2012) The dimensionality of political space: Epistemological and methodological considerations. *European Union Politics* 13(2): 194–218.

Bradburn, N.M., Sudman, S. and Wansink, B. (2004) *Asking Questions*, Revised edn. San Francisco, CA: Jossey-Bass.

Budge, I. (2001) Theory and measurement of party policy positions. In: I. Budge, H-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds.) *Mapping Policy Preferences*. Oxford, UK: Oxford University Press, pp. 75–91.

Cronbach, L. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3): 297–334.

DeVellis, R.F. (2011) *Scale Development: Theory and Applications*, 3rd edn. Thousand Oaks, CA: Sage.

Dinas, E. and Gemenis, K. (2010) Measuring parties' ideological positions with manifesto data: A critical evaluation of the competing methods. *Party Politics* 16(4): 427–450.

Duncan, O.D. and Stenbeck, M. (1987) Are Likert scales unidimensional? *Social Science Research* 16(3): 245–259.

Fivaz, J. and Nadig, G. (2010) Impact of Voting Advice Applications (VAAs) on voter turnout and their potential use for civic education. *Policy and Internet* 2(4): 167–200.

Garzia, D. (2010) The effects of VAAs on users' voting behaviour: An overview. In: L. Cadroni and D. Garzia (eds.) *Voting Advice Applications in Europe*. Naples, Italy: ScriptaWeb, pp. 13–47.

Gemenis, K. (2012) Proxy documents as a source of measurement error in the Comparative Manifestos Project. *Electoral Studies* 31(3): 594–604.

Gemenis, K. and Dinas, E. (2010) Confrontation still? Examining parties' policy positions in Greece. *Comparative European Politics* 8(2): 179–201.

Groot, L. (2003) Een kritische evaluatie van de StemWijzer 2002. *Tijdschrift voor Beleid, Politiek en Maatschappij* 30(1): 20–30.

Gschwend, T. and Proksch, S.-O. (2010) Improving the measurement of policy preferences in surveys: Bringing the status-quo back in. Paper presented at the REPCONG-Workshop Policy Representation Europe; 28–29 May, Mannheim, Germany.

Hansen, J.M. (1998) Individuals, institutions, and public preferences over public finance. *American Political Science Review* 92(3): 513–531.

Hartwig, P. (2010) The internet, political participation and election turnout: A case study of Germany's, www.abgeordnetenwatch.de. *German Politics and Society* 28(3): 156–175.

Hattie, J. (1985) Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9(2): 139–164.

Hix, S., Noury, A. and Gérard, R. (2006) Dimensions of politics in the European Parliament. *American Journal of Political Science* 50(2): 494–511.

Hodge, D.R. and Gillespie, D.F. (2004) Phrase completion scales. In: K. Kempf-Leonard (ed.) *Encyclopedia of Social Measurement*, Vol. 3, Amsterdam, the Netherlands: Elsevier, pp. 53–62.

Hooghe, L. *et al* (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, L., Marks, G. and Wilson, C. (2002) Does left/right structure party positions on European integration? *Comparative Political Studies* 35(8): 973–976.

Jacoby, W. (1991) *Data Theory and Dimensional Analysis*. Thousand Oaks, CA: Sage.

Johns, R. (2005) One size doesn't fit all: Selecting response scales for attitude items. *Journal of Elections, Public Opinion and Parties* 15(2): 237–264.

Johns, R. (2010) Likert items and scales. Colchester: UK Data Archive. Survey Question Bank Methods Fact Sheet 1.

Katsanidou, A. and Lefkofridi, Z. (2012) Citizen representation in the European Union: Policy congruence in the 2009 EP election. In: S. Banducci *et al* (eds.) *An Audit of Democracy in the EU*. Florence, Italy: European University Institute, pp. 153–174.

Kleinnijenhuis, J. and Krouwel, A. (2008) Simulation of decision rules for party advice websites. Amsterdam: Free University Amsterdam (mimeo).

Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S. and Frey , T. (2006) Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research* 45(6): 921–956.

Krippendorff, K. (2004) *Content Analysis: An Introduction to Its Methodology*, 2nd edn. Thousand Oaks, CA: Sage.

Krosnick, J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5(3): 213–236.

Krosnick, J.A. and Presser, S. (2010) Question and questionnaire design. In: P.V. Marsden and J.D. Wright (eds.) *Handbook of Survey Research*, 2nd edn. Bingley, UK: Emerald, pp. 263–314.

Ladner, A. and Pianzola, J. (2010) Do voting advice applications have an effect electoral participation and voter turnout? Evidence from the 2007 Swiss federal elections. In: E. Tambouris, A. Macintosh and O. Glassey (eds.) *Electronic Participation*. Berlin, Germany: Springer, pp. 211–224.

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

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

Likert, R. (1932) A technique for the measurement of attitudes. *Archives of Psychology* 22: 1–55.

Lin, L.I. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45(1): 255–268.

Louwerse, T. and Rosema, M. (2011) The design effects of voting advice applications: Comparing methods of calculating results. Paper presented at the ECPR General Conference; 25–27 August, Reykjavik, Iceland.

Lowe, W. (2008) Understanding Wordscores. *Political Analysis* 16(4): 356–371.

Marschall, S. and Schmidt, C. (2008) Preaching to the converted or making a difference? Mobilizing effects of an internet application at the German general election 2005. In: D. Farrell and R. Schmitt-Beck (eds.) *Non-party Actors in Electoral Politics*. Baden-Baden, Germany: Nomos, pp. 259–278.

McDonald, R.A., Thurston, P.W. and Nelson, M.R. (2000) A Monte Carlo study of missing items methods. *Organizational Research Methods* 3(1): 71–92.

Nezi, R., Sotiropoulos, D.A. and Toka, P. (2010) Attitudes of Greek parliamentarians towards European and national identity, representation and scope of governance. *South European Society and Politics* 15(1): 79–96.

Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*, New edn. London: Pinter.

Pellikaan, H., van der Meer, T. and de Lange, S. (2003) The road from a depoliticized to a centrifugal democracy. *Acta Politica* 38(1): 23–49.

Ramonaitė, A. (2010) Voting advice applications in Lithuania: Promoting programmatic competition or breeding populism? *Policy and Internet* 2(1): 117–147.

Schwarz, D., Schädel, L. and Ladner, A. (2011) Pre-election positions and voting behaviour in parliament: Consistency among Swiss MPs. *Swiss Political Science Review* 16(3): 533–564.

Sijtsma, K. (2009) On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74(1): 107–120.

Škop, M. (2010) Are the Voting Advice Applications (VAAs) telling the truth? Measuring VAAs' quality. In: L. Cedroni and D. Garzia (eds.) *Voting Advice Applications in Europe*. Naples, Italy: ScriptaWeb, pp. 199–230.

Spector, P.E. (1992) *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage.

Spirling, A. and McLean, I. (2007) UK OC OK? Interpreting optimal classification scores for the U.K. House of Commons. *Political Analysis* 15(1): 85–96.

Talonen, J. and Sulkava, M. (2011) Analyzing parliamentary elections based on voting advice application data. In: J. Gamma, E. Bradley and J. Hollmen (eds.) *Advances in Intelligent Data Analysis X*. Berlin, Germany: Springer, pp. 340–351.

Trechsel, A.H. (2010) EU-Profiler: Positioning of the parties in the European Elections [computer file]. European University Institute [distributor] V5.

Trechsel, A.H. and Mair, M. (2011) When parties (also) position themselves: An introduction to the EU Profiler. *Journal of Information Technology and Politics* 8(1): 1–20.

van der Eijk, C. (2001) Measuring agreement in ordered rating scales. *Quality & Quantity* 35(3): 325–341.

van Schuur, W.H. (2003) Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis* 11(2): 139–163.

Veen, T. (2011) Positions and salience in European Union politics: Estimation and validation of a new dataset. *European Union Politics* 12(2): 267–288.

Volkens, A. (2007) Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies* 26(1): 108–120.

Wagner, M. and Ruusuvirta, O. (2012) Matching voters to parties: Voting advice applications and models of party choice. *Acta Politica*, 47(4): 400–422.

Walgrave, S., Nuytemans, M. and Pepermans, K. (2009) Voting advice applications and the effect of statement selection. *West European Politics* 32(6): 1161–1180.

Walgrave, S., van Aelst, P. and Nuytemans, M. (2008) 'Do the vote test': The electoral effects of a popular vote advice application at the 2004 Belgian elections. *Acta Politica* 43(1): 50–70.

Wall, M., Sudulich, M.L., Costello, R. and Leon, E. (2009) Picking your party online: An investigation of Ireland's first online voting advice application. *Information Polity* 14(3): 203–218.

Wheatley, J., Carman, C., Mendez, F. and Mitchell, J. (2012) The dimensionality of the Scottish political space: Results from an experiment on the 2011 Holyrood elections. *Party Politics*, in press.

Appendix A

Table A1: Examples of double-barrelled or quantitative statements in recent European VAAs

VAA	*Country*	*Year*	*Statement*
Wahlkabine	Austria	2008	Should the names of sex offenders in Austria be recorded on a special register and made public?
Wahlkabine	Austria	2008	Should the video surveillance of public spaces (eg., spaces, public buildings and transport, and so on) be *further increased* in Austria?
Help Me Vote	Greece	2010	The electoral system must be changed and single-member constituencies should be established in order to strengthen the relationship between citizens and MPs and eliminate the need for campaign spending.
Help Me Vote	Greece	2010	The police should use *more strict* repressive measures to protect the property of citizens.
Vote Match	UK	2010	The Government should be compelled by law to halve the budget deficit within 4 years.
Vote Match	UK	2010	All Government departments should face some of the spending cuts necessary to reduce the budget deficit.
StemWijzer	The Netherlands	2010	People who become unemployed receive unemployment benefits. The amount they receive in the first few months needs to be *increased*, but the total period of the benefits has to be *much shorter*.
StemWijzer	The Netherlands	2010	Car ownership has to become *cheaper* and the use of the car more *expensive*. Therefore there needs to be a kilometre levy.
StemWijzer	The Netherlands	2010	There should be *stricter* punishment for people who use violence.

Table A1 *continued*

VAA	*Country*	*Year*	*Statement*
Smartvote	Switzerland	2011	Should Switzerland legalize the consumption of hard and soft drugs as well as the possession of such drugs for personal consumption?
Smartvote	Switzerland	2011	Should access to 'facilitated naturalization' via the Federation be made *more difficult*?
Smartvote	Switzerland	2011	Would you essentially welcome *greater restriction* on the tax competition that applies between the cantons?

Note: The elements of double-barrelled statements are underlined; quantitative elements are marked in italics.

Appendix B

Items in the EU Profiler

1. Social programmes should be maintained even at the cost of higher taxes.
2. Greater efforts should be made to privatise health-care services in the [country].
3. State subsidies for créches and child care should be increased substantially.
4. Immigration policies oriented towards skilled workers should be encouraged as a means of fostering economic growth.
5. Immigration into the [country] should be made more restrictive.
6. Immigrants from outside Europe should be required to accept our culture and values.
7. The legalisation of same sex marriages is a good thing.
8. Religious values and principles should be shown greater respect in politics.
9. The decriminalisation of the personal use of soft drugs is to be welcomed
10. Euthanasia should be legalised.
11. Government spending should be reduced in order to lower taxes.
12. The EU should acquire its own tax-raising powers.
13. Governments should bail out failing banks with public money.
14. Governments should reduce workers' protection regulations in order to fight unemployment.
15. The EU should drastically reduce its subsidies to Europe's farmers.

16. Renewable sources of energy (for example, solar or wind energy) should be supported even if this means higher energy costs.
17. The promotion of public transport should be fostered through green taxes (for example, road taxing).
18. Policies to fight global warming should be encouraged even if it hampers economic growth or employment.
19. Restrictions of civil liberties should be accepted in the fight against terrorism.
20. Criminals should be punished more severely.
21. On foreign policy issues, such as the relationship with Russia, the EU should speak with one voice.
22. The European Union should strengthen its security and defence policy.
23. European integration is a good thing.
24. [Country] is much better off in the EU than outside it.
25. The European Union should be enlarged to include Turkey.
26. The European Parliament should be given more powers.
27. Individual member states of the EU should have less veto power.
28. Any new European Treaty should be subject to approval in a referendum in the [country].