Policy and Internet

Response Scales in Voting Advice Applications: Do Different Designs Produce Different Outcomes?

Martin Rosema and Tom Louwerse

Voting Advice Applications (VAAs) represent popular election campaign tools in many countries, enabling voters to discover which party or candidate provides the best match with their political preferences. This article examines the effects of design choices on these tools by focusing on the response scale that is used to measure the policy positions of parties and voters. We analyze the impact of scale length on the advice generated by these tools using user data from a VAA developed for the 2014 Dutch local elections. We transform the original 101-point scale into several alternative scale formats and determine if this leads to a different voting recommendation. We also examine the suitability of alternative scales for creating spatial models, which are often employed by VAAs. We show that the response scale has a potentially profound impact on the resulting advice (with voters receiving different VAA outcomes depending on the scale length), except for voters with an extremist response style. The findings have practical implications for the design of these tools: outcomes should be presented as a preference list, rather than focusing on the "best match."

KEY WORDS: Voting Advice Application, policy preferences, attitude measurement, electoral behavior, response scale

Introduction

The rise of the Internet and the increase in the number of undecided voters have boosted the development and usage of Voting Advice Applications (VAAs), especially in countries with a multiparty system. In countries like Switzerland, Germany, and the Netherlands millions of voters visit such websites before election day—and even on the day itself—to find out which party or candidate provides the best match with their policy preferences. The widespread use of VAAs and their potential impact on citizens' voting behavior give the developers a great responsibility when it comes to the design. When developing a VAA, they have to make numerous choices, such as selecting statements, choosing a method to determine party or candidate positions, and deciding about the format that is used to present the advice to users. Previous research has shown that such design choices have a strong impact on the advice that is provided. For instance, which statements are included will influence how often a particular political party is

Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

listed as the best match (Walgrave, Nuytemans, & Pepermans, 2009). Furthermore, the method that VAAs adopt to transform users' answers into a voting recommendation has a strong impact on the voting recommendation that it generates (Louwerse & Rosema, 2014). This article focuses on another element of the design of VAAs, which has received scarce attention in the academic literature, namely the answer scale that is used to measure users' policy preferences and to position political parties or candidates. More specifically, we explore if the choice for a particular scale length has an effect on the advice that users receive and to what extent alternative response scales are suitable for creating the spatial models that VAAs sometimes employ.

VAAs are well-known and widely used in many countries today (for reviews about their history and usage, see Cedroni & Garzia, 2010; Garzia & Marschall, 2012), and research about the design, purpose, and effects of VAAs has matured in recent years (see, e.g., Garzia & Marschall, 2014; Rosema, Anderson, & Walgrave, 2014). The topics addressed in research about VAAs are diverse and include, among others, the selection of statements (Lefevere & Walgrave, 2014; Walgrave et al., 2009), the coding of party positions (Gemenis, 2013; Gemenis & van Ham, 2014; Trechsel & Mair, 2011), and the spatial models used in VAAs (Germann, Mendez, Wheatley, & Serdült, 2015; Louwerse & Otjes, 2012; Otjes & Louwerse, 2014). These are all elements of the design of VAAs. This type of research has shown that design choices can have a great impact on the advice that the tools generate, and that the use of spatial models leads to high demands on the data used for constructing the models that underlie the advice. Other studies have focused on the consequences of VAA usage, such as effects on political knowledge (Schultze, 2012), political interest (Kruikemeier, van Noort, Vliegenthart, & de Vreese, 2014), electoral turnout (Dinas, Trechsel, & Vassil, 2014; Garzia, De Angelis, & Pianzola, 2014; Gemenis & Rosema, 2014), electoral volatility (Ladner, Fivaz, & Pianzola, 2012), and party choice (Walgrave, Van Aelst, & Nuytemans, 2008; Wall, Krouwel, & Vitiello, 2014). Although the results are mixed, on the basis of these studies we can safely state that VAAs have an impact on voters' behavior, both in terms of electoral turnout and in terms of party choice. Voters who are hesitating about their choice at the polls appear to be particularly affected (Ruusuvirta & Rosema, 2009; Wall et al., 2014). Still other topics addressed in this growing literature include profiles of VAA users (Hanel & Schultze, 2014; Marschall & Schultze, 2015; Wall, Sudulich, Costello, & Leon, 2009), normative foundations of VAAs in theories of democracy and citizenship (Anderson & Fossen, 2014; Fossen & Anderson, 2014), and methodological challenges of VAA research (Pianzola, 2014).

Previous studies have created much insight into the effects of design choices, but some issues have remained poorly understood. One of these is the relevance of the answer scales that VAAs adopt to measure policy preferences. Sometimes users have to choose between "in favor" and "against" or between "agree" and "disagree" in response to a statement. The advantage of such binary choices is that the meaning of the categories is clear and the choice is simple. Some other VAAs present a 5-point agree–disagree scale, which suggests that preferences are

measured with more precision. Still other VAAs use scales that distinguish between more positions, sometimes as many as a hundred. Especially when more than two categories are used, the question arises if users, parties or candidates, and developers assign the same meaning to all those categories. Making meaningful comparisons between party positions and user positions in order to create a voting recommendation is obviously crucial. Apart from the number of scale positions, there are also differences with respect to the labeling of the answer categories and how the tools deal with the possibilities of a neutral position and "don't know" or "no opinion" option. In this article we aim to assess whether it makes any difference what response scale format is used in these tools by focusing on a key element, namely the number of answer categories or scale length.

We approach this matter in two ways. The basis for our analyses are data from the Dutch VAA called *De Stem Van*, which asked voters and parties to indicate their positions with respect to a statement on a continuum where two verbal labels were used to assign meaning to both end points. No numerical labels were visible for the users, but the software coded the responses in terms of a 0–100 scale. First, we analyze if using a scale with fewer positions would have resulted in different advice. This is done by recoding user and party data via several linear transformations into scales with fewer positions and next comparing the voting advice that would then have been generated (cf., Louwerse & Rosema, 2014). This implies that in this study we only focus on the *mechanical effect* of scale length, that is, the consequences of having less precision available for the measurement. There is a body of research that shows that the number of positions a scale has, the availability of a mid-point, and the labels that are used also have *psychological effects*, such as the tendency to opt more or less often for an extreme response option or to choose the middle category (see, e.g., Tourangeau, Rips, & Rasinski, 2000). So the effects of scale length can be stronger than what we report in this study if one would also take such other elements into account. Second, we analyze the suitability of different response scales for constructing dimensions in a spatial model, because many VAAs present their advice in this way (Louwerse & Rosema, 2014). Before presenting these analyses, however, we first elaborate on the theoretical foundations of the use of different response scales in VAAs and discuss what is known more in general about the effects of response scale formats.

This article provides a first step in analyzing the effects of scale length. We acknowledge that our approach, recoding users' and parties' answers on a 101-scale to shorter scales and observing effects on the advice provided, cannot take into account all aspects relating to scale length. In particular, our recoding assumes *linearity*, that is, that positions on the long scale can be linearly transformed into positions on a shorter scale. This assumption might not hold if we were, in an experimental setting, to compare users' positions on a long and a short scale. In other words, presented with scales of different lengths, users and parties might change their preferred position in a nonlinear way. We highlight this limitation of the study early on, so as to be clear on what the current study

can and cannot do. Our findings, based on our recoding exercise, suggest small to moderate effects of scale length on VAA advice provided. In the conclusion, we provide suggestions for further research in this area, using experimental designs.

Answer Scales and Spatial Models in VAAs

There is clearly no agreement among the designers of VAAs about the type of answer scale that is best used to gauge the policy preferences of citizens and parties or candidates. We take the national elections in the Netherlands in 2012 as an example to illustrate the diversity. The oldest and most popular VAA, *StemWijzer*, asked users to indicate if they "agree" or "disagree" with each statement, while also providing a "neither" option. So this response scale corresponds with a 3-point scale. Users also had the option to indicate that they do not know their position on an item ("skip this question"). Furthermore, after they have answered all the statements users got the opportunity to indicate what statements they consider most important. The statements selected were then given a double weight, which means that one could argue that in practice for voters the scale operated in a similar way as a 5-point scale: Agreeing or disagreeing with a party on an important issue resulted in a score of plus or minus 2, while agreeing or disagreeing with a party on another issue resulted in a score of plus or minus 1; choosing the neutral category led to a score of 0. Another tool from the same organization, *Stemmentracker*, is an example of a test that adopted a binary scale: users could only choose between "in favor" and "against" (or skip the question). The mechanism of these calculations is comparable with the way scale positions are used in the directional model of issue voting (Rabinowitz & Macdonald, 1989).

Another popular VAA in the Netherlands, *Kieskompas*, adopted a 5-point scale to measure the policy preferences of its users and presents five standard answer options listed under each statement: "fully agree," "agree," "neutral," "disagree," and "fully disagree." Additionally, users had the option to select the answer category "no opinion." *Kieskompas* uses the logic of a spatial theory of issue voting (Enelow & Hinich, 1984), because these scales are used to calculate a position in a spatial model in which distances indicate dissimilarities. *Kieskompas* was certainly not the only VAA to use this scale format. There have been many other VAAs in use in national elections in the Netherlands. An inventory made at the website www.kieswijzerkiezer.nl listed no less than 34 different VAAs for the 2012 Dutch parliamentary elections. Among these, a 5-point scale with a neutral mid-point was the most popular response scale format. But other formats were also used. For instance, the tool *Kieswijzer 2012* presented three or four alternative policy options for each statement, thus making use of a categorical instead of an ordinal answer scale format. Still other VAAs used an ordinal response scale, but expanded the number of response options.

VAAs in other countries have often adopted comparable response scales. For example, the German *Wahl-O-Mat 2014* adopted a 3-point scale with the response options "agree," "neutral," and "disagree." The Swiss *Smartvote 2011* presented

its users with a 4-point scale ("ja," "eher ja," "eher nein," "nein") complemented with a "no answer" option. Furthermore, users could also indicate the importance of each statement on a 5-point scale. The scale made use of symbols: $++$, $+$, $=$, $-$, and $--$. In Flanders *De Stemtest 2014* made use of a binary scale ("agree" vs. "disagree" or "no opinion"). Another VAA in Flanders, called *De Stem van Vlaanderen*, made use of a graphical scale instead of a numbered scale and coded the position of the slider, which users could move along the continuum to indicate their position on each statement, in terms of a 101-point scale.

When the scores of users and parties are transformed into a voting recommendation, many VAAs follow the logic of so-called spatial models (Mendez, 2012; Wagner & Ruusuvirta, 2012). *Kieskompas* is a clear example, with its result screen that plots a horizontal and vertical axis and positions parties and user in the resulting two-dimensional space. *Smartvote* in Switzerland, with the so-called *spider diagram*, also transforms the answers into positions on a limited number (six or seven) of dimensions. Although this approach is appealing for political scientists in particular, because it links up with the widely applied spatial framework used to analyze party competition and voting behavior (Downs, 1957; Enelow & Hinich, 1984), the spatial approach as employed in VAAs has also received criticism (Germann et al., 2015; Louwerse & Otjes, 2012; Louwerse & Rosema, 2014; Otjes & Louwerse, 2014). Moreover, scholars who are familiar with the literature on issue voting are aware that spatial models using Euclidean distance, which VAAs closely link up to, are not the only way in which attitude scales that are used to assess policy preferences can be modeled regarding vote choice (Mendez, 2012).

An important argument against the use of spatial models based on Euclidean distance is given by those who posit that the response scales of the underlying items can only be meaningfully transformed into spatial dimensions if the positions of the scale represent a set of alternative policy options, an assumption that is usually not met in practice (Rabinowitz & Macdonald, 1989; Stokes, 1963). In other words, if a statement about building a new road in a particular town has, say, a response scale with seven positions along a continuum, the interpretation of the scale is that voters who position themselves at different values want something different. An alternative interpretation is that such a scale has only two alternative policy positions—for example, in favor or against building the road—while the actual position indicates the strength of the preference for either choice option (building vs. not building the road). Building on this type of argumentation, Rabinowitz and Macdonald (1989) formulated a directional theory of issue voting, in which a scalar product of party position and voter position is central instead of the distance between party and voter on the same scale. Using data from VAAs, Mendez (2012) compared how often these alternative algorithms led to an accurate prediction of citizens' real voting preferences and found that the directional model performed slightly better than the Euclidean distance model. However, one should be aware that the purpose of VAAs is not to predict existing vote preferences, but to provide a well-reasoned assessment of how well each party or candidate matches with citizens' political preferences. In that sense

the debate about which of these "logics" is more appropriate for VAAs is still open.

In this article we are interested in the effects of the response scale, in particular scale length, and focus on the more traditional spatial models using Euclidean distance, because these have been central in VAAs more often. Readers should be aware, though, that we thus limit our focus, and in addition to the effects that we observe there will also be effects stemming from the choice of the type of issue voting model that underlies the calculations. Furthermore, we are interested in what we referred to above as the *mechanical effect* of scale length. Self-evidently, a scale that has only two or three positions allows for less precision when expressing political preferences than, say, an 11-point or 100-point scale. In addition to those effects changes in scale length and labeling can have all sorts of other effects on the answers given by respondents, which we have referred to as *psychological effects*. The tendency to use the middle category or extreme categories are well-known examples of this (see, e.g., Moors, 2008; Tourangeau et al., 2000; Weijters, Cabooter, & Schillewart, 2010). Any estimate of the size of scale length effects on the advice that a VAA generates therefore has to be viewed as an estimate of the minimum of such effects.

Response Scales and Response Styles

In the literature about VAAs the response scales have not received much attention, but they have in other domains. When deciding about the response scale format for a closed question, which are understandably more common than open-ended questions (Schuman & Presser, 1981, chap. 1), several choices have to be made, such as how many response options to provide (scale length), how to label the response options (labeling), in which order to present the response options (ordering), and how to visually present the response options (visual presentation), assuming that the information is transferred visually, which corresponds with how VAAs operate in practice. All these elements, and their advantages and disadvantages, have been analyzed in past research in other contexts than VAA research.

In its simplest form a scale will have two positions, for example, giving the choice between "in favor" and "against" or between "agree" and "disagree." The advantages of such binary scales are that they are relatively easy to read and interpret by respondents, because the number of alternatives is limited and the meaning of the values is usually clear, and consequently the questions take less time than with alternative answer formats. These facts, combined with experimental findings that patterns observed with alternative longer scales tend to be rather similar, have led some scholars to recommend this scale format, in particular when respondents have to answer a long list of questions (Dolnicar, 2003).

An important characteristic of the 2-point scale is that it has no mid-point that represents a neutral position, so respondents are forced to choose between either sides. The matter of including a mid-point or not is relevant not only for

distinguishing between a 2-point scale and a 3-point scale, but also for longer scales. While analyzing this matter for questions about ideology in terms of a left–right continuum, Kroh (2007) concluded that the answers about left–right ideology become more valid when a mid-point is included, and hence an 11-point scale is preferable to a 10-point scale. This suggests that some respondents really consider themselves as neither left-wing nor right-wing, and when forced to choose either side their answers are less stable and do not reveal actual ideological positions.

However, the use of a mid-point also has disadvantages. This has to do with the fact that the substantive meaning of a mid-point is not always clear and the interpretation of its meaning differs among respondents; some interpret it as a neutral position, others as reflecting having no opinion. Another disadvantage is that when labels are used, how often the mid-point is chosen depends on the label used: it is more often chosen when the label is "neutral" than when the label is "no opinion" (Nadler, Weston, & Voyles, 2015). So even if a middle category has been shown to be meaningful for some items (e.g., left–right ideology), this is not necessarily true for all items. When designers of survey questionnaires make their choices about scale length and labeling, they also have to consider if a "no opinion" or "don't know" option is offered (Giljam & Granberg, 1993; Mondak & Davis, 2001). If it is included, it is important to present it visually as something separate from the regular answer scale, because otherwise respondents get confused about the mid-point of the scale: the visual mid-point is then interpreted as the conceptual mid-point (Tourangeau, Couper, & Conrad, 2004).

In surveys, policy preferences are often assessed using scales that have more positions than two or three, varying from four or five positions up to about a hundred (as, e.g., with so-called feeling thermometers, which have been widely used in national election studies to measure evaluations of parties and candidates). Several scholars have found that at the individual level, answers on binary scales are less reliable than answers on scales with more positions, such as 5- or 7-point scales. When the number of choice options exceeds 10, however, the reliability of the scale appears to decrease again, and hence 100-point scales appear to lead to more random error than 10-point scales (Preston & Colman, 2000). These findings suggest that scales that allow for much fine-tuning in practice lead to more noise instead of more precise measurement. This is confirmed in the study on the measurement of left–right ideology by Kroh (2007), who concluded that an 11-point scale gives more valid results than a 101-point scale.

Instead of listing a fixed number of positions, scales can also be presented as a line between two extremes, with respondents being allowed to choose any point on that line (often transformed into a 101-point scale). The widespread use of Internet-based surveys seems to have made this scale format more popular, as it can easily be offered on a computer screen with a slider. In practice these slider scales operate in very similar ways as ordinal scales, as shown in a study comparing these formats using print questionnaires (Dolnicar & Grün, 2007) and in another study that compared the use of a slider with radio buttons in an online

survey (Roster, Lucianetti, & Albaum, 2015). However, these so-called visual analogue scales do require more time than ordinal scales and more often lead to missing data, which led Couper, Tourangeau, Conrad, and Singer (2006) to recommend against their use, unless respondents can make fine-tuned distinctions on the items in question.

There has been much research about the effects of scale length, building on the idea that the number of answer categories may influence the answer that respondents provide (Gaskell, O'Muircheartaigh, & Wright, 1994). Other elements of the question design also matter. For example, instead of listing all answer categories at once, it is possible to make use of branched questions, for instance by first asking respondents if they agree or disagree with a statement and next asking about the intensity of their (dis)agreement. This has consequences for the time needed to answer questions (branched questions take longer to answer) as well as the answers that are provided (branched questions lead to more extreme answers) (Gilbert, 2015). The direction of a scale (the order in which answer categories are listed), can also have an impact on what answers are provided (Yan & Keusch, 2015), as can all kinds of elements of the layout (Couper, Tourangeau, Conrad, & Crawford, 2004; Tourangeau, Couper, & Conrad, 2007). However, the most important choice regarding the response scale appears to be the scale length and labeling.

This brief review of the academic literature on response scales underlines that answers provided by participants in surveys are not merely reflections of true opinions (cf. Zaller, 1992). The answers depend at least in part on the format of the response scale that is used. The answers that respondents provide to survey questions also depend on yet another factor. Individuals not only differ in terms of the opinions and preferences they have, but they also differ in the way they use answer scales in surveys. These patterns have become known as *response styles*. In their review, Van Vaerenbergh and Thomas (2013) distinguished eight different response styles, including two that are presumably most well-known: acquiescence response style and extreme response style. The former refers to the tendency to give positive answers and thus agree with statements, while the latter refers to the tendency to opt for answers at the end points of a scale. The sources of response patterns cannot only be found in features of the stimulus (Weijters et al., 2010), but also in characteristics of the respondents, such as demographic characteristics like age, education, and personality (Greenleaf, 1992; Leeper, 2014; Van Vaerenbergh & Thomas, 2013). Furthermore, the strength of the effect of response styles depends on a range of factors, including scale length (Kieruj & Moors, 2010), mode of data collection (Ye, Fulton, & Tourangeau, 2011), and the country in which data are collected (Van Herk, Poortinga, & Verhallen, 2004).

For research about VAAs, the question arises what the implications are of the effects that response scales have. In this case what is of paramount interest is not whether at the aggregate level the opinions that the electorate at large seems to have is accurately reflected by particular scales, or to what extent these answers are influenced by the scale length, but primarily if the response scale has an impact on the advice received by users. On the one hand, it is theoretically

possible that short scales lack precision and consequently lead to other recommendations than those one might consider appropriate. On the other hand, it is also possible that long scales include more "noise" and hence that the recommendations are based not just on true preferences, but on this noise. In both scenarios the scale length has an effect and hence the central question we address in this article is to what extent the scale length of the response scale in VAA questionnaires has an effect on the recommendation that users receive. For VAAs it is not that important if the preference with respect to a specific policy issue is adequately captured, but above all if the advice that is provided is adequate. We therefore want to analyze the effect of scale length on the outcome of these tools.

Data and Methods

To analyze the impact of the response scale on the advice that VAAs generate, we work with data from a VAA that was introduced in the Netherlands in 2014 and targeted at the local elections: *De Stem Van*. These three words, which can be translated into English as "the voice of" or "the vote of," were complemented with the name of the municipality in question, for example, *De Stem Van Enschede*. This VAA was an initiative of a policy research and consultancy company oriented at the public sector, *Necker van Naem*, in collaboration with the Dutch online media company *NU.nl* and the Belgian company *iVox*, which already had experience with VAAs in Belgium (*De Stem Van Vlaanderen* and *La Voix des Belges*).

To indicate their policy preferences, for each statement users had to move a slider along a horizontal continuum (see Figure 1). No numerical values were presented and labels were used only at both ends of the scale, while points and thin vertical lines were presented on the continuum. Users could not skip items or register a do not know answer. The starting position of the slider was exactly in the middle and it had to be moved before users were able to go to the next statement (they could move it to the left or right and then back to the starting position in the middle). The software coded the position of the slider in terms of a scale with values ranging from 0 to 100, and hence in practice this VAA made use of a 101-point scale. Political parties had been invited beforehand to indicate their positions in an identical way, so using the same screens and the same software. The match between user and party was calculated with a formula that took into account the width of the positions of parties and user along the continuum. The data emerging from this tool provide a good opportunity to observe the implications of changing the response scale, which can be done by merging sets of answer positions and then analyzing the consequences for the advice generated.

We kindly received information from *De Stem Van* about the party positions and the answers of 2,500 randomly selected users in three Dutch municipalities: Apeldoorn, Nijmegen, and Zwolle. In these three municipalities *De Stem Van* was the only VAA available, and consequently the usage figures were relatively high and comparable with those of *StemWijzer* in other municipalities.[1] Because VAAs are very popular tools in the Netherlands, we can safely conclude that the users

Figure 1. Screenshot From VAA *De Stem Van*, Including the Response Scale Used for Each Statement.

are not just a small segment of political sophisticates. In these municipalities the VAA comprised 25 or 26 statements.

We acknowledge that not all participants may have answered the statements of the VAA in a serious manner. Instead, people might "click through" to look at the "advice" page. The best practice would be to filter out nongenuine answers based on the time it took to complete the tool (Andreadis, 2014). Unfortunately, the VAA did not include a timer, so we cannot filter out responses on that basis. Users could not "click through" the VAA (by answering "don't know" or any other option consistently), as users had to move the slider that was used to register their answer. The quickest way to go through the VAA would therefore be to slide to either extreme for each question. Therefore, we filter out participants who gave a very extreme answer (higher than 90 or lower than 10) more than 90 percent of the time. This reduces our sample from 7,500 to 7,318 (97.6 percent of the original sample).[2]

In the below analysis, the user and party data from the 101-point scale are recoded into different scales of various lengths: 11, 10, 9, 8, 7, 6, 5, 4, 3, and 2 points. The Appendix shows exactly how the scales have been recoded. Note that for some scale lengths one category was larger than the other ones, for example, when recoding a 101-point scale into a 4-point scale we map 25 original scale points on each of the new scale points, but are then left with one remainder (because $101/4 = 25$ with a remainder of 1). In our analysis any remainder has been consistently assigned to the lowest scale point (or scale points if the remainder exceeded 1). We have, however, replicated our analyses by assigning remainders to the highest scale points, which yields very similar results for all of the reported analyses.

Our analysis consists of three steps. First, we provide descriptive results regarding the congruence of recommendations between pairs of scales. Second, we analyze the determinants of this degree of congruence using a multilevel regression model. Third, we look at the impact of differing scale lengths on the use of low-dimensional spatial representations, a technique that is used in a large minority of VAAs.

To calculate the degree of match between parties and users, and subsequently determine which party provides the best match, different metrics can be used. Because the results all point in the same direction, we will report the Euclidean distance only, but we also calculated figures for the City block distance and the Agreement (or Exact match) method (cf., Louwerse & Rosema, 2014, pp. 293–296). The Euclidean and City block distances are well-known metrics. Their application is straightforward as we do not have any weights or missing values to consider. The Agreement method, which is used by *StemWijzer*, basically awards 1 point for every statement on which party and user have exactly the same position. Arguably this metric only makes sense for scales with very few answer options, and indeed this metric shows much lower congruence scores than the Euclidean and City block distance.

We use two measures of congruence to assess the impact of the different response scales (cf., Louwerse & Rosema, 2014, pp. 296–297). First, we look at the percentage of users that received the same "best match" for two response scales. In the event of ties, we consider any overlap between the set of top matches for two response scales as a match. This gives an initial indication of how the most visible part of the result can change as a result of limiting the range of the response scale. The second measure of the similarity between two response scales is the correlation between the scores of all parties. The rationale is that the "best match" might easily change, because of the limited number of statements involved in a VAA and the ideological similarities between some parties. Party scores for two methods might be highly correlated, even if the "best match" changes. Each calculation method provides a (dis)similarity metric between the user and each party. We calculated the correlation between user–party dissimilarity scores for any two response scales. For example, we calculate the (dis) similarity of each user with party A on a 2-point scale and the (dis)similarity of each user with party A on an 11-point scale, then we calculate the correlation between both measures, and in a final step we determine the average correlation across all parties. As we are looking at the same statements, parties, users, and method for calculating the advice, we should expect that these correlations are high. If we find correlations below 0.8 that would demonstrate quite a substantial impact of the length of the response scales on the advice provided.

In the next step, we focus on the determinants of the congruence of recommendations. We ran a multilevel linear regression model seeking to explain the correlation of party scores for a single user in each dyad of scales, for example, the correlation of scores for user number 47 according to the 101- and 7-point scales. For each of the 7,318 users in total, we observe 55 dyads of response scales. Our analysis takes the dyadic nature of these data into account

by including random effects for the participant, the dyad, and the two response scales (Gilardi & Füglister, 2008). Our explanatory variables are a measure of the difference in scale length,[3] the extremism of the answers provided,[4] and whether both scales are both odd or both even (or not). We also include a dummy variable for the municipality.

The last part of our analysis concerns the use of a low-dimensional model to present VAA results. The idea is that VAA statements tap into underlying issue dimensions that structure the political competition between parties. The Electoral Compass family of VAAs, for example, generally uses two dimensions: left–right and progressive–conservative (Otjes & Louwerse, 2014). One argument in the debate on answer scale lengths in VAAs is indeed whether they are suitable for constructing a low-dimensional model. Most VAAs that use such a model do indeed use slightly longer answer scales (4 or 5 points), whereas some other VAAs that do not use these models use short scales (2 or 3 points).

We estimate a low-dimensional model based on users' positions on each of the statements. We should note that the items were not selected with any kind of spatial model in mind and local politics may display less of a structure than national politics. Initial analyses indeed showed low scalability of these models, with Mokken scaling analysis, the preferred model of scaling, resulting in dimensions consisting only of a few statements (cf., Germann & Mendez, 2016; Germann et al., 2015). As constructing many scales with only two or three items each would not make sense from the idea of providing a low-dimensional representation of the political space to users, we opted to create a two-dimensional model. We focus on the case of Nijmegen, which shows relatively high congruence scores in the descriptive analysis. We perform the scaling analysis based on the 11-point answer scales, using a factor analysis. We assign each statement to the scale which shows the highest factor loading, provided this is at least 0.4 (which implies a 0.4 correlation between the statement and the factor). In this way, we obtain two dimensions with three and four items, respectively. The Loevinger's H for these scales is (mostly) low (0.26 and 0.28 on the user data, 0.58 and 0.66 on the party data), which means that we should interpret these results with some caution. Next we use this selection of items to calculate party and issue positions on the scales, according to the 11, 5, and 3-point answer scales. Therefore, the differences that we find are due to differences in the number of answer categories, not differences in the composition of the scales.

Results

To get an idea of how parties and users used the 101-point scale of *De Stem Van*, we display all positions taken on issues in each VAA (see Figure 2). All scale positions are used to a certain extent by the users, which is important for our analysis. The extreme positions are relatively popular both with parties and voters. The same is true for the scores of 25 and 75, which is presumably due to the visual guides plotted at those positions that serve as a kind of anchor. The

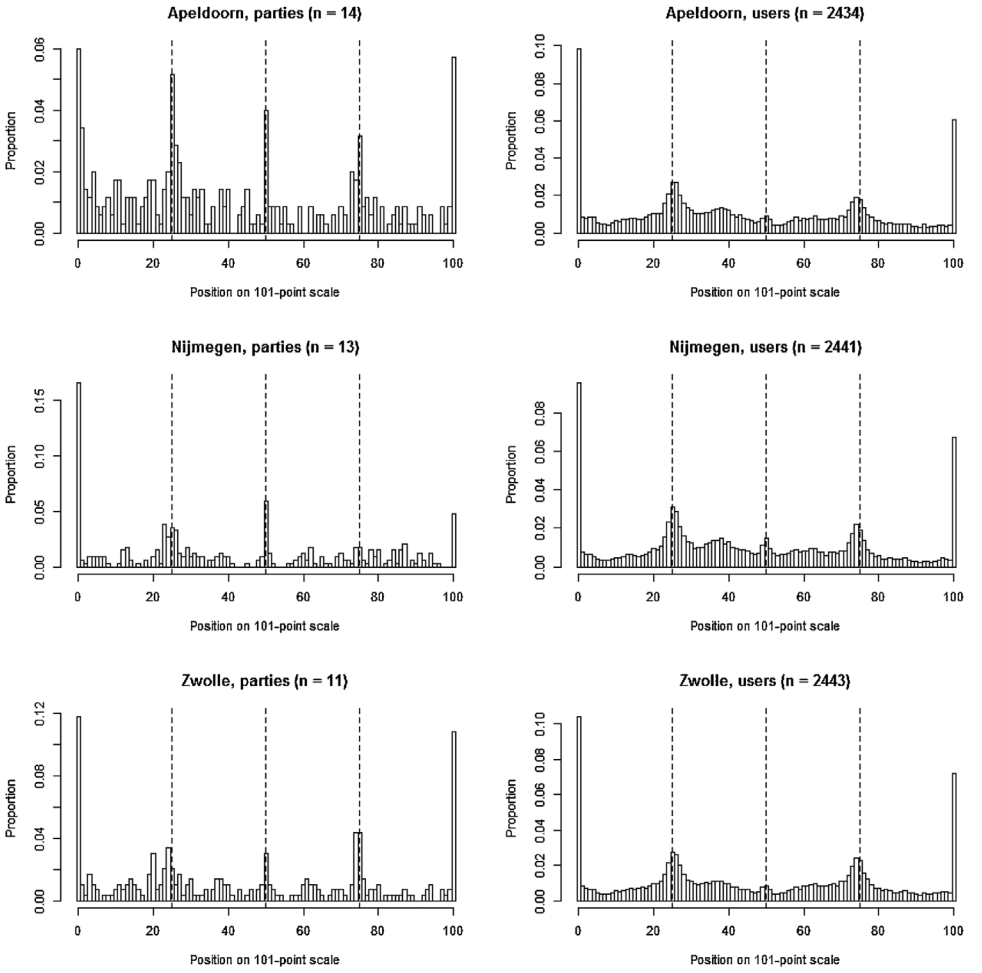


Figure 2. Distribution of Positions on 101-Point Scale for Parties and Users.
Notes: The n refers to the number of parties and users, respectively. Each party or user took a position on 25 or 26 statements, which are combined in these figures.

50 percent mark is relatively popular as well among parties, but not among voters. Perhaps parties that wanted to position themselves in the middle took more effort in precisely finding that position than voters did—recall that the slider had to be moved, so the 50 percent mark could only be reached by moving the slider away from the middle position and then back.

Congruence of Matches

If we reduce the scale length, how does this affect the advice given by the VAA? As outlined above, we recoded parties' and users' answers to an 11, 10, 9, 8, 7, 6, 5, 4, 3, and 2-point scale. Our first measure concerns how many users get the same best match with different scale lengths. Table 1 presents the mean scores

over the three municipalities of the percentage of users that do receive the same top match, using the Euclidean distance metric.

The main finding is that as scale length decreases, the percentage of users who receive the same advice as with the original 101-point scale reduces significantly. About 89 percent of users receive the same party as best match under the 101- and 11-point scales. This is a relatively good result, given that the top match might easily shift. The results for a 10, 9, 8, and 7-point scale are also quite similar to the 101- and 11-point scales; and the 5-point scale provides at least 73 percent of the users with the same top advice as under longer scales. If we go down any further, however, we see increasing differences between scales. A 2-point scale would have provided only about half of the users with the same best match. While this is a large difference, it is not entirely surprising, as moving from a 101- to a 2-point scale is substantial and a lot of information is (apparently) lost in the process.

The differences between the longer scales (101 to 7 points) are relatively modest compared to the differences between the shorter scales (4 to 2 points). The 3- and 2-point scales, for example, only provide the same top advice in 54 percent of the cases. This seems to suggest that the difference between shorter scales is larger than between longer scales. The differences are indeed particularly strong for the 2-point scale: at most 56 percent get the same result as under any of the other scale lengths. A 3-point scale shows higher congruence with all other scales.

We do not observe large differences between odd-numbered scales, which have an explicit middle category, and even-numbered scales, which do not. Sometimes, an odd-numbered scale shows higher congruence with other odd-numbered scales, for example, the 5-point scale matches slightly better with the 7-point scale than with the 6-point scale. But these effects are small and not consistent across all scales. All in all, we do not observe a strong "middle category effect."

Our second measure of the degree to which different scales result in different outcomes is the correlation between the recommendations. Each calculation method provides a (dis)similarity metric between the user and each party. The

Table 1. Proportion of Users Receiving the Same Best Match for Different Scales

	101	11	10	9	8	7	6	5	4	3	2
101	1.00	0.89	0.89	0.88	0.86	0.86	0.81	0.73	0.71	0.64	0.49
11	0.89	1.00	0.88	0.84	0.84	0.86	0.82	0.73	0.70	0.67	0.51
10	0.89	0.88	1.00	0.87	0.85	0.85	0.81	0.78	0.72	0.65	0.52
9	0.88	0.84	0.87	1.00	0.84	0.81	0.80	0.77	0.71	0.65	0.50
8	0.86	0.84	0.85	0.84	1.00	0.81	0.79	0.72	0.79	0.64	0.51
7	0.86	0.86	0.85	0.81	0.81	1.00	0.79	0.73	0.71	0.66	0.51
6	0.81	0.82	0.81	0.80	0.79	0.79	1.00	0.72	0.71	0.68	0.55
5	0.73	0.73	0.78	0.77	0.72	0.73	0.72	1.00	0.65	0.58	0.52
4	0.71	0.70	0.72	0.71	0.79	0.71	0.71	0.65	1.00	0.61	0.56
3	0.64	0.67	0.65	0.65	0.64	0.66	0.68	0.58	0.61	1.00	0.54
2	0.49	0.51	0.52	0.50	0.51	0.51	0.55	0.52	0.56	0.54	1.00

Notes: Euclidean distance metric. The columns and rows display the scale length (number of scale positions). The table entries display mean scores over three municipalities.

correlations of this matrix between different scales are displayed in Table 2. The findings largely confirm the patterns that we observed when looking at the best match. There are high correlations between the 101- and 5-point scales ($r \geq 0.95$). The correlations are even still quite high for the 4-point scales ($r \geq 0.91$). We observe lower correlations between the 2- or 3-point scales and longer scales ($0.61 \geq r \geq 0.89$). Again, we see no evidence of a "middle category effect."

We repeated the above analysis for the City block distance and the Agreement method (figures not reported). The City block distance yielded results very similar to the Euclidean distance, albeit generally slightly lower. The Agreement method showed large degrees of incongruence, which was to be expected as this method will only see a party and user as being in agreement when they take exactly the same position. That is much harder on a 101-point scale than on a 2-point scale, resulting in seemingly almost arbitrary advice produced with the 101-point scale and other very long scales.

Taken together, our analysis suggests that moving from a 101-point to a 5-point scale results in a relatively small change in advice. The recommendations given under those longer scales correlate strongly. Still, about 25 percent of the users would have received different "top advice" under a 5-point scale than under the 101-point scale. Incongruence increases when moving to a 4, 3, or 2-point scale. It is noteworthy that the correlation between the 4- and 3-point scales ($r = 0.82$) is *lower* than between the 101- and 4-point scales ($r = 0.92$). Overall, the shorter scales show relatively low degrees of congruence with *all* other scales, including other short scales.

Explaining Congruence Between Response Scales

The previous analyses suggest that there are considerable differences in the advice users get, depending on the length of the scale used. As we reduce scale length, the correlation with the original party–user matches declines. A multilevel linear regression model confirms these patterns (see Table 3).

Table 2. Pearson's Correlation Between User Dissimilarities With Parties for Different Scales

	101	11	10	9	8	7	6	5	4	3	2
101	1.00	0.99	0.99	0.99	0.98	0.98	0.97	0.95	0.91	0.83	0.61
11	0.99	1.00	0.99	0.99	0.98	0.98	0.98	0.96	0.92	0.86	0.64
10	0.99	0.99	1.00	0.99	0.98	0.98	0.98	0.97	0.93	0.85	0.65
9	0.99	0.99	0.99	1.00	0.98	0.97	0.98	0.97	0.93	0.86	0.64
8	0.98	0.98	0.98	0.98	1.00	0.97	0.97	0.96	0.95	0.85	0.65
7	0.98	0.98	0.98	0.97	0.97	1.00	0.97	0.95	0.92	0.87	0.66
6	0.97	0.98	0.98	0.98	0.97	0.97	1.00	0.96	0.93	0.89	0.68
5	0.95	0.96	0.97	0.97	0.96	0.95	0.96	1.00	0.92	0.83	0.65
4	0.91	0.92	0.93	0.93	0.95	0.92	0.93	0.92	1.00	0.84	0.70
3	0.83	0.86	0.85	0.86	0.85	0.87	0.89	0.83	0.84	1.00	0.68
2	0.61	0.64	0.65	0.64	0.65	0.66	0.68	0.65	0.70	0.68	1.00

Notes: The columns and rows display the scale length. The table entries display mean correlations.

Table 3. Multilevel Linear Regression Model of Correlation of Matches Between Scale Dyads

	Model 1
Intercept	0.88***
	(0.02)
Municipality	
Nijmegen	0.03***
	(0.00)
Zwolle	0.01***
	(0.00)
Difference in scale length	−0.82***
	(0.07)
Extremism of user answers	0.30***
	(0.00)
Both scale lengths odd/even	0.01
	(0.00)
AIC	−820,136.84
BIC	−820,016.88
Log likelihood	410,079.42
Num. obs.	402,490
Num. groups: participant_id	7,318
Num. groups: dyad	55
Num. groups: scale 2	10
Num. groups: scale 1	10
Var: participant_id (Intercept)	0.00
Var: dyad (Intercept)	0.00
Var: scale 2 (Intercept)	0.00
Var: scale 1 (Intercept)	0.00
Var: Residual	0.01

Note: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Our analysis shows that scale length and extremism are significantly related to the outcome variable, the correlation between two sets of party scores. We plot the expected values for each of the variables in Figure 3. We see some differences between municipalities, with Nijmegen generally showing higher correlations. The substantive effect size is, however, low, amounting to only a 0.03 increase in the correlation coefficient.

We observe larger differences for the difference in scale length. The larger the difference between the two scales, the higher the correlation of recommendations for two scales. The effect is about −0.40 between the largest difference (101- vs. 2-point scale) compared to scales of a very small difference. The level of extremism in a user's answers, which ranges from 0 to 0.5, also impacts upon the correlation between a dyad of recommendations. The most extreme answer pattern would include answers of 0 or 100 on the original scale only. The results clearly suggest that more extreme answer patterns result in a higher correlation of recommendations under different scales. This effect is quite substantial, with very moderate answers showing a correlation below 0.8, while the most extreme users in our sample would, on average, receive correlations of party scores of above 0.9.

We find no effect to be discerned of whether both scales have an odd or even number of categories. Recall that odd-numbered scales have a clear middle

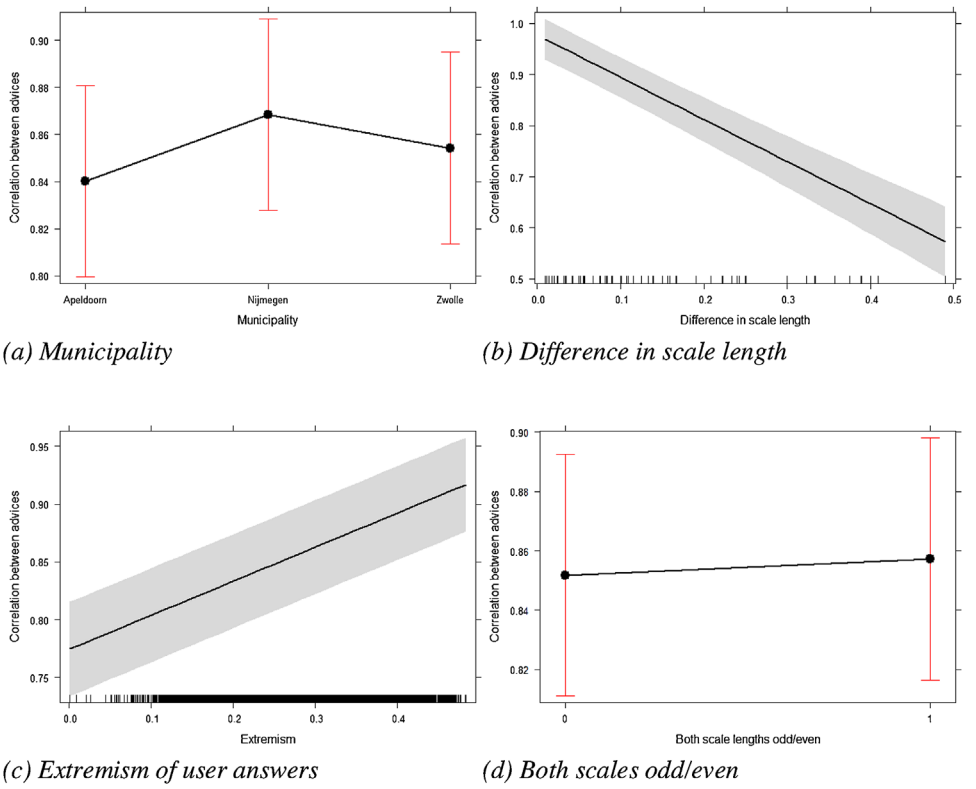


Figure 3. Expected Values of Correlation of Recommendations.
Notes: Effect plots of Model 1 in Table 3. All other variables are kept at their mean (for categorical variables: modal) value.

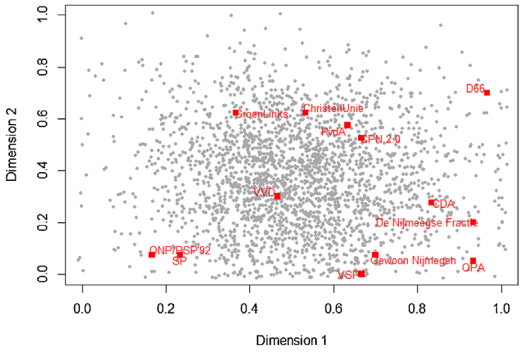
category, which is not present in even-numbered categories. Thus, if the presence of a "middle category" had an impact, we would find that the correlation of scores obtained under two dimensions that both do not have a middle category (or both have) would be higher compared to a situation where one dimension has a middle category (e.g., a 5-point scale) and the other has not (e.g., a 6-point scale). Our results show, however, no significant difference. The reason might be that the effect of the middle category is diminished by the requirement of *De Stem Van* for users to move the slider away from the exact mid-point of the scale. As the descriptive data illustrated, this resulted in only a small peak in the distribution of answers around this middle point. As the middle category seems hardly overused in our data, it is not too surprising that we find little difference between odd- and even-length scales.

Answer Scale Length and Spatial Models

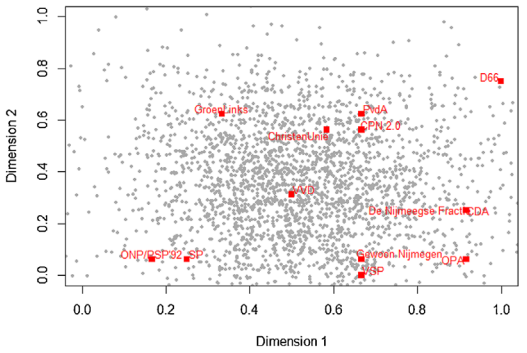
The previous analyses have used the Euclidean distance metric to calculate the VAA result. This basically treats each statement as a separate dimension in a high-dimensional space. Many VAAs, however, use a low-dimensional spatial

representation as (one of) the way(s) in which the advice is presented. How do differences in answer scale length affect the ability to construct such low-dimensional spaces?

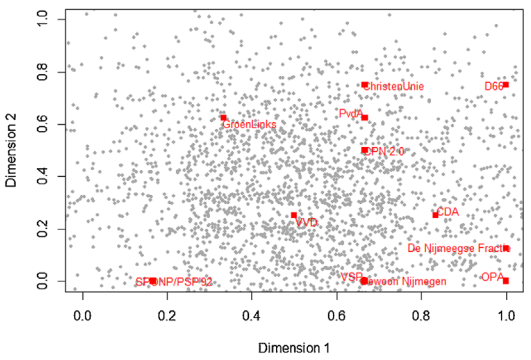
Figure 4 presents three spatial representations of user and party positions in one of our cases (Nijmegen). In fact, the party and user positions as well as the spatial model used are *the same* for all three cases. The only difference is the number of answer categories used, which varies from 11, to 5 and 3. The dimensions have been recoded to run from 0 to 1 to aid comparison. Dimension 1



(a) 11 point answer scale



(b) 5 point answer scale



(c) 3 point answer scale

Figure 4. Three Spatial Models of Party Competition in Nijmegen.

roughly corresponds with a right–left scale, while Dimension 2 seems to set apart more governmental and protest parties.

Although the configuration of parties is roughly the same in the three models, there are a number of differences. First, the number of unique positions that users can occupy is naturally smaller when using 3-point answer scales (in the figure a little "jitter" is added to prevent points being plotted exactly on top of each other). When using the 3-point scale, we thus seem to throw away some data. Second, contrary to expectation, we do not see a clustering of parties and users in the center of the political space on the shorter scales. If anything, user positions seem to be a little more spread out in the 3-point answer scales-model. Third, some parties seem to be at slightly different places in the three models, for example, the ChristenUnie which is more moderate in the 5-point model, while the VVD is somewhat more extreme in the 3-point model.

All in all, correlations between the matches that users receive with each party in these three models is high ($0.97 \geq r \geq 0.89$), although the percentage of users receiving the same top advice ranges between 66 percent and 76 percent. Bearing in mind that the only thing that differs between the three models is the length of the answer scales, this does show that answer scale length does have an effect, especially on the "top advice."

If we construct *separate spatial models* for each answer scale length, the correlation between matches declines ($0.83 \geq r \geq 0.70$) and only 26 percent to 39 percent of users would receive the same top match. This reflects that in separate models, the dimensions obtained might differ. The low level of congruence may furthermore be the result of low coefficients of scalability. It proved difficult to summarize users' responses in a low-dimensional model, even when limiting ourselves to the best-matching 7 out of a total of 25 statements being included in the dimensions. When scalability is poor, small changes in the underlying data may have large results in terms of the scaling solution obtained and therefore the advice provided to users. In line with previous work in this area, we would therefore recommend to make sure that the data are suitable for constructing a low-dimensional model (Louwerse & Rosema 2014; Otjes & Louwerse, 2014). If this is the case, it is unlikely that the number of answer categories will make a large difference. If, however, scalability is low, differences in answer category length will likely result in hugely different recommendations.

Conclusions

Previous research showed that the advice that is generated by a VAA not only depends on the answers that citizens provide in response to the statements incorporated, but also on the method that these tools use to transform the answers in a voting recommendation (Louwerse & Rosema, 2014). Indeed, when the calculation method of *Kieskompas* was applied with the data from *StemWijzer*, a majority of the voters received a different voting recommendation. This means that the design choices in VAAs have a substantial impact on the outcome. There are other studies that have shown that such design effects also occur as a result of

other elements of the design, such as statement selection (Lefevere & Walgrave, 2014; Walgrave et al., 2009). In this article we have expanded this strand of research by analyzing the relevance of the response scale format that VAAs employ. To what extent are the results of VAAs sensitive for the number of positions that the response scale includes?

We have analyzed this matter on the basis of data from a new VAA in the Netherlands, *De Stem Van*, which mapped the policy preferences of political parties and voters on a continuum where for each statement users had to select a position on the screen with a slider (visual analogue scale). These positions were coded by the VAA on a scale from 0 to 100, thus creating a 101-point scale. In our analyses we transformed the answers from voters as well as the positions of political parties on those same items into a different format, such as an 11-point scale, 5-point scale, or a simple dichotomy (2-point scale). Although the effects of limiting the number of scale positions from 101 to 11 did not have particularly strong consequences, further reducing the number of scale positions did result in substantive effects. For example, had *De Stem Van* coded the user responses in terms of "favor," "against," or "neither" (effectively creating a 3-point scale), a format that is comparable to the one used by *StemWijzer*, about one-third of the users would have received a voting recommendation for a different party. Self-evidently, this large number has to some extent to do with the relatively large number of parties that compete in Dutch elections. Clearly, in a two-party system the figure would be much smaller. But still, we believe that even in the Dutch context this is quite a sizable proportion of the electorate. Furthermore, we have shown that the advice is also strongly affected by scale length if a VAA makes use of a low-dimensional spatial model, unless the items correlate strongly and the scalability of the dimensions is good. For such models identifying items that correlate strongly enough to the underlying latent dimensions, however, appears to be a stronger challenge than identifying the optimal scale length.

Basically, these findings mean that we have found further evidence for the claim that the design of a VAA has a substantive impact on the advice that the tool generates when citizens express their policy preferences. We should also note, however, that the effects that we have observed in this research are considerably smaller than the effects regarding the calculation method and the spatial model adopted by a VAA (Louwerse & Rosema, 2014). We should also note that some characteristics of the VAA we used might impact on our results. Users had to provide a substantive answer on the 101-point scale and furthermore were not allowed to simply keep the slider in the middle position. This presents a somewhat different environment from other VAAs. Had, for example, an explicit no opinion button been available, users might have had missing answers on more items, potentially increasing the impact of scale length, as fewer items would have been included in the advice calculations.

One of the questions that arise is what the implications are for designers of VAAs. Should they perhaps avoid scales with few answer options, because these are unable to adequately capture the nuances of voters' opinions and consequently may result in "wrong" voting advice? Or should designers avoid response scales

with many answer options, because such answers will include "noise" and consequently the subsequent advice is to some extent misleading? After all, it is difficult to convincingly contend that citizens are able to meaningfully distinguish between a hundred different positions with respect to a particular policy topic. On the basis of our analysis we cannot answer such questions. There is no *a priori* reason why fewer or more answer categories would be better. To address this topic, we also need insight, for example, in matters such as the degree to which citizens can meaningfully distinguish between a number of intensities or positions when it comes to the type of statements included in a VAA and the reliability of such scales (would the same voters give the same answer when asked to give a response twice?). Furthermore, it depends also on the degree to which the number of response scale positions facilitates or inhibits an identical interpretation of the substantive meaning of each position among parties and voters. And, last, voters might be more likely to choose a no opinion/do not know option (when available), if the scale length offers too little or too much choice. These matters might be addressed in future research. It is of paramount importance to know if voters and parties (or those who code party positions), interpret scale values of VAA items in comparable ways. It has already been argued that voters may interpret the middle category in VAAs in different ways (Baka, Figgou, & Triga, 2012) and the same could well be true for the other scale positions. For these reasons scales with many positions should be treated with some skepticism.

What we can conclude at this point, is that the response scale to be adopted is one of the choices that designers of VAAs face and that the choice matters. Hence, further research about the properties of alternative response scales would be valuable. More specifically, experimental research in which comparable questions are complemented with different response scales is necessary to increase our insights in this topic. Such research is vital for understanding the "psychological effects" of scale length, such as the tendency to use or avoid the middle category and the tendency to choose extreme answer categories. One difficulty of such an approach would be, however, that it will not be as straightforward to compare the advice provided to an individual user under different lengths of scales. It would, after all, be problematic to offer voters scales of different lengths to map their position on the same issues. Therefore, such an analysis would almost necessarily shift the analysis from the individual level to the aggregate level: how often is each party recommended under different scale lengths?

One point of advice can already be made for the developers of VAAs. If the advice is indeed rather strongly dependent on design choices, such as statement selection, response scale format, and calculation method, it would be unwise to pretend that any VAA can really tell which party provides the best match. As we have argued before, listing a small set of parties instead of a single party as best choice option would do more justice to what we know about VAA design.

Martin Rosema, Ph.D., Assistant Professor of Political Science, Department of Public Administration, University of Twente, Enschede, the Netherlands [m.rosema@utwente.nl].

Tom Louwerse, Ph.D., Assistant Professor of Political Science, Institute of Political Science, Universiteit Leiden, Leiden, the Netherlands.

Notes

1. Source: Personal communication with Peter Joosten from *Necker Van Naem*, June 1, 2016.
2. The results are not greatly affected by the filter. If we include all observations, the reported values in Tables 1 and 2 change by less than 0.02 units. We note that users might have used other strategies to quickly "click through" the VAA, such as moving the slider only a bit on each question, but we cannot filter out these answer patterns without making strong assumptions about what constitutes a reasonable answer.
3. We measure the difference in scale length as follows:

$$scalediff_{x,y} = \left|\frac{1}{x} - \frac{1}{y}\right|$$

   where x and y stand for the two scale lengths. For example, the difference measure for the 101- and 9-point scales is $|\frac{1}{100} - \frac{1}{9}|$, for the 9- and 5-point scales (0.08) and for the 9- and 2-point scales (0.38). Note that this measure is much larger for the difference between a 3- and 2-point scales (0.17) than the difference for the 10- and 11-point scales (0.009). This is exactly what we want to capture as the reduction in precision of the scales is much larger in the former case.
4. Extremism is measured as the average absolute distance of each of the user's answers to the scale midpoint (50). We divide this measure by 100, so it runs from 0 to 0.5, which is better comparable to the range on which the dependent variable is measured (0–1).

References

Anderson J., and T. Fossen. 2014. "Voting Advice Applications and Political Theory: Citizenship, Participation, and Representation." In *Matching Voters With Parties and Candidates: Voting Advice Applications in a Comparative Perspective*, eds. D. Garzia and S. Marschall. Colchester: ECPR Press, 217–26.

Andreadis, I. 2014. "Data Quality and Data Cleaning." In *Matching Voters With Parties and Candidates: Voting Advice Applications in Comparative Perspective*, eds. D. Garzia and S. Marschall. Colchester: ECPR Press, 79–92.

Baka, A., L. Figgou, and V. Triga. 2012. "'Neither Agree, Nor Disagree': A Critical Analysis of the Middle Answer Category in Voting Advice Applications." *International Journal of Electronic Governance* 5 (3/4): 244–63.

Cedroni, L., and D. Garzia, eds. 2010. *Voting Advice Applications in Europe: The State of the Art*. Napoli, Italy: Scriptaweb.

Couper, M.P., R. Tourangeau, F.G. Conrad, and S.D. Crawford. 2004. "What They See Is What We Get: Response Options for Web Surveys." *Social Science Computer Review* 22 (1): 111–27.

Couper, M.P., R. Tourangeau, F.G. Conrad, and E. Singer. 2006. "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment." *Social Science Computer Review* 24 (2): 227–45.

Dinas, E., A.H. Trechsel, and K. Vassil. 2014. "A Look Into the Mirror: Preferences, Representation and Electoral Participation." *Electoral Studies* 36: 290–97.

Dolnicar, S. 2003. *Simplifying Three-Way Questionnaires—Do the Advantages of Binary Answer Categories Compensate for the Loss of Information?* Australia and New Zealand Marketing Academy (ANZMAC) CD Proceedings.

Dolnicar, S., and B. Grün. 2007. "How Constrained a Response: A Comparison of Binary, Ordinal and Metric Answer Formats." *Journal of Retailing and Consumer Services* 14: 108–22.

Downs, A. 1957. *An Economic Theory of Democracy*. New York: Harper & Row.

Enelow, J.M., and M.J. Hinich. 1984. *The Spatial Theory of Voting: An Introduction*. Cambridge: Cambridge University Press.

Fossen, T., and J. Anderson. 2014. "What's the Point of Voting Advice Applications? Competing Perspectives on Democracy and Citizenship." *Electoral Studies* 36: 244–51.

Garzia, D., A. De Angelis, and J. Pianzola. 2014. "The Impact of Voting Advice Applications on Electoral Participation." In *Matching Voters With Parties and Candidates: Voting Advice Applications in Comparative Perspective*, eds. D. Garzia and S. Marschall. Colchester: ECPR Press, 105–14.

Garzia, D., and S. Marschall. 2012. "Voting Advice Applications Under Review: The State of Research." *International Journal of Electronic Governance* 5 (3/4): 203–22.

Garzia, D., and S. Marschall, eds. 2014. *Matching Voters With Parties and Candidates: Voting Advice Applications in Comparative Perspective*. Colchester: ECPR Press.

Gaskell, G.D., C.A. O'Muircheartaigh, and D.B. Wright. 1994. "Survey Questions About the Frequency of Vaguely Defined Events: The Effects of Response Alternative." *Public Opinion Quarterly* 58 (2): 241–54.

Gemenis, K. 2013. "Estimating Parties' Policy Positions Through Voting Advice Applications: Some Methodological Considerations." *Acta Politica* 48 (3): 268–95.

Gemenis, K., and C. van Ham. 2014. "Comparing Methods for Estimating Parties' Positions in Voting Advice Applications." In *Matching Voters With Parties and Candidates: Voting Advice Applications in Comparative Perspective*, eds. D. Garzia and S. Marschall. Colchester: ECPR Press, 33–48.

Gemenis, K., and M. Rosema. 2014. "Voting Advice Applications and Electoral Turnout." *Electoral Studies* 36: 281–89.

Germann, M., and F. Mendez. 2016. "Dynamic Scale Validation Reloaded: Assessing the Psychometric Properties of Latent Measures of Ideology in VAA Spatial Maps." *Quality & Quantity* 50 (3): 981–1007.

Germann, M., F. Mendez, J. Wheatley, and U. Serdült. 2015. "Spatial Maps in Voting Advice Applications: The Case for Dynamic Scale Validation." *Acta Politica* 50: 214–38.

Gilardi, F., and K. Füglister. 2008. "Empirical Modeling of Policy Diffusion in Federal States: The Dyadic Approach." *Swiss Political Science Review* 14 (3): 413–50.

Gilbert, E.E. 2015. "A Comparison of Branched Versus Unbranched Rating Scales for the Measurement of Attitudes in Surveys." *Public Opinion Quarterly* 79 (2): 443–70.

Giljam, M., and D. Granberg. 1993. "Should We Take Don't Know for an Answer?" *Public Opinion Quarterly* 57 (3): 348–57.

Greenleaf, E.A. 1992. "Measuring Extreme Response Style." *Public Opinion Quarterly* 56 (3): 328–51.

Hanel, K., and M. Schultze. 2014. "Analyzing the Political Communication Patterns of Voting Advice Application Users." *International Journal of Internet Science* 9 (1): 31–51.

Kieruj, N.D., and G. Moors. 2010. "Variations in Response Style Behavior by Response Scale Format in Attitude Research." *International Journal of Public Opinion Research* 22 (3): 320–42.

Kroh, M. 2007. "Measuring Left-Right Political Orientation: The Choice of Response Format." *Public Opinion Quarterly* 71 (2): 204–20.

Kruikemeier, S., G. van Noort, R. Vliegenthart, and C.H. de Vreese. 2014. "Unraveling the Effects of Active and Passive Forms of Political Internet Use: Does It Affect Citizens' Political Involvement?" *New Media & Society* 16 (6): 902–20.

Ladner, A., J. Fivaz, and J. Pianzola. 2012. "Voting Advice Applications and Party Choice: Evidence From Smartvote Users in Switzerland." *International Journal of Electronic Governance* 5 (3/4): 367–87.

Leeper, T.J. 2014. "Cognitive Style and the Survey Response." *Public Opinion Quarterly* 78 (4): 974–83.

Lefevere, J., and S. Walgrave. 2014. "A Perfect Match? The Impact of Statement Selection on Voting Advice Applications' Ability to Match Voters and Parties." *Electoral Studies* 36: 252–62.

Louwerse, T., and S. Otjes. 2012. "Design Challenges in Cross-National VAAs: The Case of the EU Profiler." *International Journal of Electronic Governance* 5 (3/4): 279–97.

Louwerse, T., and M. Rosema. 2014. "The Design Effects of Voting Advice Applications: Comparing Methods of Calculating Matches." *Acta Politica* 49 (3): 286–312.

Marschall, S., and M. Schultze. 2015. "German E-Campaigning and the Emergence of a 'Digital Voter'? An Analysis of the Users of the Wahl-O-Mat." *German Politics* 24 (4): 525–41.

Mendez, F. 2012. "Matching Voters With Political Parties and Candidates: An Empirical Test of Four Algorithms." *International Journal of Electronic Governance* 5 (3/4): 264–78.

Mondak, J.J., and B.C. Davis. 2001. "Asked and Answered: Knowledge Levels When We Will Not Take 'Don't Know' for an Answer." *Political Behavior* 23 (3): 199–224.

Moors, G. 2008. "Exploring the Effect of a Middle Response Category on Response Style in Attitude Measurement." *Quality & Quantity* 42 (6): 779–94.

Nadler, J.T., R. Weston, and E.C. Voyles. 2015. "Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires." *Journal of General Psychology* 142 (2): 71–89.

Otjes, S., and T. Louwerse. 2014. "Spatial Models in Voting Advice Applications." *Electoral Studies* 36: 263–71.

Pianzola, J. 2014. "Selection Biases in Voting Advice Application Research." *Electoral Studies* 36: 272–80.

Preston, C.C., and A.M. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104: 1–15.

Rabinowitz, G., and S.E. Macdonald. 1989. "A Directional Theory of Issue Voting." *American Political Science Review* 83 (1): 93–121.

Rosema, M., J. Anderson, and S. Walgrave, eds. 2014. "Voting Advice Applications." Special Symposium in *Electoral Studies* 36: 240–297.

Roster, C.A., L. Lucianetti, and G. Albaum. 2015. "Exploring Slider vs. Categorical Response Formats in Web-Based Surveys." *Journal of Research Practice* 11 (1): 1–15. http://jrp.icaap.org/index.php/jrp/article/view/509/413.

Ruusuvirta, O., and M. Rosema. 2009. "Do Online Vote Selectors Influence Electoral Participation and the Direction of the Vote?" Paper presented at the ECPR Conference, September 10–12, Potsdam, Germany.

Schultze, M. 2012. "Effects of Voting Advice Applications (VAAs) on Political Knowledge About Party Positions." *Policy and Internet* 6 (1): 46–68.

Schuman, H., and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Stokes, D.E. 1963. "Spatial Models of Party Competition." *American Political Science Review* 57 (2): 368–77. Reprinted in 1966 in: A. Campbell, P.E. Converse, W.E. Miller, and D.E. Stokes. *Elections and the Political Order*, New York: John Wiley & Sons, 159–79.

Tourangeau, R., M.P. Couper, and F. Conrad. 2004. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68 (3): 368–93.

Tourangeau, R., M.P. Couper, and F. Conrad. 2007. "Color, Labels, and Interpretive Heuristics for Response Scales." *Public Opinion Quarterly* 71 (1): 91–112.

Tourangeau, R., L.J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Trechsel, A.H., and P. Mair. 2011. "When Parties (Also) Position Themselves: An Introduction to the EU Profiler." *Journal of Information Technology & Politics* 8 (1): 1–20.

Van Herk, H., Y.H. Poortinga, and T.M.M. Verhallen. 2004. "Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries." *Journal of Cross-Cultural Psychology* 35 (3): 346–60.

Van Vaerenbergh, Y., and T.D. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25 (2): 195–217.

Wagner, M., and O. Ruusuvirta. 2012. "Matching Voters to Parties: Voting Advice Applications and Models of Party Choice." *Acta Politica* 47 (4): 400–22.

Walgrave, S., M. Nuytemans, and K. Pepermans. 2009. "Voting Aid Applications and the Effects of Statement Selection." *West European Politics* 32 (6): 1161–80.

Walgrave, S., P. Van Aelst, and M. Nuytemans. 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., A. Krouwel, and T. Vitiello. 2014. "Do Voters Follow the Recommendations of Voter Advice Application Websites? A Study of the Effects of Kieskompas.nl on Its Users' Vote Choices in the 2010 Dutch Legislative Elections." *Party Politics* 20 (3): 416–28.

Wall, M., M.L. Sudulich, R. Costello, and E. Leon. 2009. "Picking Your Party Online—An Investigation of Ireland's First Online Voting Advice Application." *Information Polity* 14 (3): 203–18.

Weijters, B., E. Cabooter, and N. Schillewart. 2010. "The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Rresponse Category Labels." *International Journal of Research in Marketing* 27: 236–47.

Yan, T., and F. Keusch. 2015. "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey." *Public Opinion Quarterly* 79 (1): 145–65.

Ye, C., J. Fulton, and R. Tourangeau. 2011. "Research Synthesis: More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice." *Public Opinion Quarterly* 75 (2): 349–65.

Zaller, J.R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Appendix

Recoding of 101 Scale Into Shorter Response Scales

101	11	10	9	8	7	6	5	4	3	2
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
11	2	1	1	1	1	1	1	1	1	1
12	2	2	1	1	1	1	1	1	1	1
13	2	2	2	1	1	1	1	1	1	1
14	2	2	2	2	1	1	1	1	1	1
15	2	2	2	2	1	1	1	1	1	1
16	2	2	2	2	2	1	1	1	1	1
17	2	2	2	2	2	1	1	1	1	1
18	2	2	2	2	2	2	1	1	1	1
19	2	2	2	2	2	2	1	1	1	1
20	3	2	2	2	2	2	1	1	1	1
21	3	2	2	2	2	2	1	1	1	1
22	3	3	2	2	2	2	2	1	1	1
23	3	3	2	2	2	2	2	1	1	1
24	3	3	3	2	2	2	2	1	1	1
25	3	3	3	2	2	2	2	1	1	1
26	3	3	3	2	2	2	2	1	1	1
27	3	3	3	3	2	2	2	2	1	1
28	3	3	3	3	2	2	2	2	1	1
29	4	3	3	3	2	2	2	2	1	1
30	4	3	3	3	3	2	2	2	1	1
31	4	3	3	3	3	2	2	2	1	1
32	4	4	3	3	3	2	2	2	1	1
33	4	4	3	3	3	2	2	2	1	1
34	4	4	3	3	3	2	2	2	1	1
35	4	4	4	3	3	3	2	2	2	1
36	4	4	4	3	3	3	2	2	2	1
37	4	4	4	3	3	3	2	2	2	1
38	5	4	4	3	3	3	2	2	2	1
39	5	4	4	4	3	3	2	2	2	1
40	5	4	4	4	3	3	2	2	2	1
41	5	4	4	4	3	3	2	2	2	1
42	5	5	4	4	3	3	3	2	2	1
43	5	5	4	4	3	3	3	2	2	1
44	5	5	4	4	4	3	3	2	2	1
45	5	5	4	4	4	3	3	2	2	1
46	5	5	5	4	4	3	3	2	2	1
47	6	5	5	4	4	3	3	2	2	1
48	6	5	5	4	4	3	3	2	2	1
49	6	5	5	4	4	3	3	2	2	1
50	6	5	5	4	4	3	3	2	2	1
51	6	5	5	4	4	3	3	2	2	1
52	6	6	5	5	4	4	3	3	2	2
53	6	6	5	5	4	4	3	3	2	2
54	6	6	5	5	4	4	3	3	2	2
55	6	6	5	5	4	4	3	3	2	2
56	7	6	5	5	4	4	3	3	2	2
57	7	6	6	5	4	4	3	3	2	2
58	7	6	6	5	4	4	3	3	2	2
59	7	6	6	5	5	4	3	3	2	2
60	7	6	6	5	5	4	3	3	2	2
61	7	6	6	5	5	4	3	3	2	2
62	7	7	6	5	5	4	4	3	2	2
63	7	7	6	5	5	4	4	3	2	2
64	7	7	6	6	5	4	4	3	2	2
65	8	7	6	6	5	4	4	3	2	2
66	8	7	6	6	5	4	4	3	2	2
67	8	7	6	6	5	4	4	3	2	2
68	8	7	7	6	5	5	4	3	3	2
69	8	7	7	6	5	5	4	3	3	2
70	8	7	7	6	5	5	4	3	3	2
71	8	7	7	6	5	5	4	3	3	2
72	8	8	7	6	5	5	4	3	3	2
73	8	8	7	6	6	5	4	3	3	2
74	9	8	7	6	6	5	4	3	3	2
75	9	8	7	6	6	5	4	3	3	2
76	9	8	7	6	6	5	4	3	3	2
77	9	8	7	7	6	5	4	4	3	2
78	9	8	7	7	6	5	4	4	3	2
79	9	8	8	7	6	5	4	4	3	2
80	9	8	8	7	6	5	4	4	3	2
81	9	8	8	7	6	5	4	4	3	2
82	9	9	8	7	6	5	5	4	3	2
83	10	9	8	7	6	5	5	4	3	2
84	10	9	8	7	6	5	5	4	3	2
85	10	9	8	7	6	6	5	4	3	2
86	10	9	8	7	6	6	5	4	3	2
87	10	9	8	7	7	6	5	4	3	2
88	10	9	8	7	7	6	5	4	3	2
89	10	9	8	8	7	6	5	4	3	2
90	10	9	9	8	7	6	5	4	3	2
91	10	9	9	8	7	6	5	4	3	2
92	11	10	9	8	7	6	5	4	3	2
93	11	10	9	8	7	6	5	4	3	2
94	11	10	9	8	7	6	5	4	3	2
95	11	10	9	8	7	6	5	4	3	2
96	11	10	9	8	7	6	5	4	3	2
97	11	10	9	8	7	6	5	4	3	2
98	11	10	9	8	7	6	5	4	3	2
99	11	10	9	8	7	6	5	4	3	2
100	11	10	9	8	7	6	5	4	3	2
101	11	10	9	8	7	6	5	4	3	2

Note: Scales have been recoded using the cut function in R.