SCUOLA NORMALE SUPERIORE

Scuola Normale Superiore
Department of Political and Social Sciences

PhD in Political Science and Sociology
Department of Political and Social Sciences

Empowering Citizens in the Digital Age

–

A Systematic Evaluation of Voting Advice Applications and Best Practices for their Design

Thesis of

Sebastianus Cornelis Jacobus Bruinsma

Supervisors:

Prof. Filippo Tronconi .....

Candidate:

Sebastianus Bruinsma .....

Abstract

Voting Advice Applications (VAAs) are online tools that match voters and parties based on common positions on a series of issues. Starting as tools for political education, VAAs have nowadays become relevant political actors. Besides providing information, they also raise electoral turnout, improve political knowledge and influence party choice. The degree in which they do so depends on their design. In this thesis, I view the design of VAAs as a process and focus on two steps: the questionnaire and the visualizations. For the questionnaire I focus on two aspects: the scales used to position users on the political map and the formulation of the questions. I assess the scales using various data-reduction techniques and rank them on unidimensionality, quality and reliability. I find that most scales score insufficient, though the score depends on the construction method used for the scale. The cause of these problematic scales is that VAA users often apply simplification methods. This means they do not always understand the question or use the response categories as intended. This results in problematic scales, resulting in a political map that is difficult to interpret. Also, I find that altering the questions in the main questionnaire to have either a positive or negative formulation, not only influences the responses of the users but the match between the user and the party as well. For the visualization, I run an online experiment in which I ask users to answer questions related to various kinds of visualization. I find that not only do they have difficulty to finish some basic tasks, they also have diverging interpretations of popular VAA visualizations. My main conclusion is therefore that the design of VAAs cannot be neutral. Also, the underlying information used to calculate the match and visualize the political map are often troublesome. Yet, I also show designers can use simple methods to improve their VAAs. This is important as VAAs are likely to become even more popular than they are already at this moment.

Foreword

The first time I became aware of Voting Advice Applications was during the elections of 2002 in the Netherlands. Despite being too young to vote, I found it a nice pastime to click my way through them. In those days, I based my appreciation of the result on whether I liked the party's logo - things have only changed little since then. And so have Voting Advice Applications. Despite new visualizations and behind-the-scenes improvement, they remain questionnaires with visuals. Yet, a sizeable amount of people take it seriously - sometimes even too seriously - which makes me very happy and rather concerned at the same time. This thesis is (hopefully) confirmation of that trust for the sceptics and maybe a slight wake-up call for the believers.

At the SNS, I would like to thank Filippo Tronconi, Lorenzo Mosca, and Marco Deseriis for being supervisors with an enviable amount of patience and useful comments. A similar word of thanks goes to Fernando Mendez and Vasilis Manavopoulos for helping to develop, design and run Stem-Consult and make the EUVox data available. Without them, no data, and without data, no thesis. Also, Kees Aarts, helped me to write and structure my research proposal which became the foundation of this thesis. To my family, I wish to say that their explicit and inexplicit support at all hours helped, though I often wonder if they knew what I was up to during the last years. Yet, as it allowed many visits to Florence, I think they did appreciate it.

Then, I would like to thank Kostas Gemenis. To list all the things I should thank him for takes too much time, but I can say that without his persistence, helpful remarks, and guiding, this thesis would not and could not have been written - I would probably not have started the whole project in the first place.

Finally, I would like to end with a brief quote, that I think nicely sums up the spirit of the thesis:

> "When you are studying any matter or considering any philosophy, ask yourself only what are the facts and what is the truth that the facts bear out. Never let yourself be diverted either by what you wish to believe, or by what you think would have beneficent social effects if it were believed. But look only, and solely, at what are the facts."
>
> — Bertrand Russell

Table of Contents

List of Figures

List of Tables

1 | Introduction

Citizens form the foundation of democracies. In the ideal situation, they take part in the democracy and learn what they need to run it with success. Such a democracy is both desirable in a moral and practical sense. As all citizens take part, its policies are close to the requirements of the citizens and the citizens can carry them out in a manner that suits them. Moreover, as all citizens are aware of what is, and what is not, possible, they are more content with the policies they carry out. This leads to a satisfied society, which is less difficult to govern than a disgruntled one (Irvin and Stansbury 2004).

Still, this ideal democracy seems to be far away from everyday reality. In most western democracies, the level of political interest is in decline (Zmerli and Meer 2017). Also, electoral turnout — for most the only way in which they take part in political life — falls year by year (Blais 2000). Besides, citizens are often irrational (Kahneman 2011), misled with ease (Caplan 2007) and lack the required political knowledge to make political decisions (Delli Carpini and Keeter 1997). In the words of Brennan (2011): "(...) [citizens] are incompetent, ignorant, irrational, and morally unreasonable about politics (...) despite that, they hold political power" (p.700).

This lack of political interest and political knowledge has led to widespread debates among scholars (Dalton 1988, 2014; Skocpol and Fiorina 1999; Norris 1999). Most scholars link it to the gradual erosion of the traditional social structures that started in the 1970*s*. This erosion severed the link between the voting behaviour of a citizen and their religion, class and cultural background (Franklin, Mackie, and Valen 1992). In this new situation, citizens became *issue voters*. This meant that party choice now depended on how close they were to a party (Downs 1957; Rabinowitz and MacDonald 1989). In other words, citizens now chose parties not because of who they are, but because of what issues the party supports.

The key concept of issue voting is *information*. If citizens do not know what the positions of the various parties are, they are unable to make a meaningful vote (Bartels 1986). Besides, issue voting requires citizens to have and develop their own opinions on these issues (Walgrave, Aelst, and Nuytemans 2008a). Yet, many citizens lack either time, resources or the knowledge to do so. Instead, they use electoral "shortcuts". Instead of reading electoral manifestos they use candidate characteristics, endorsements or polls (Popkin 1991; Lupia 1994). The reason for this is the high cost they have to pay for information. One of the reasons this cost is so high is because of the large number of political parties and candidates that can take part in a party system. For example, during the 2017 elections in the Netherlands, voters had to choose between 1,116 candidates spread over 28 parties. And, during the 2016 general election in Ireland 551 candidates participated. Obtaining information on all these parties

and candidates is difficult. This is especially so when they are new or not well-known. Also, the widening "gap" between citizens and politicians makes them even less inclined to get the knowledge they need (Thomassen 2015). The result of this is a steady dwindling of the level of political knowledge (Dalton 2000).

This decrease in political knowledge is not only a problem for the individual citizen, but also for society as a whole. As democracy depends on citizens voting in line with their opinions, an insufficient level of political knowledge threatens it. Thus, scholars and politicians have searched for new ways to increase the level of political knowledge of citizens. One way has taken them online. This because of the positive relationship between the use of the internet and political knowledge and participation (Dalrymple and Scheufele 2007; Boulianne 2009; Brennan 2011; Dimitrova et al. 2014). Here, tools like podcasts, social networks, and weblogs allow citizens new ways to find relevant political information (Kaid and Holtz-Bacha 2008). While most of these techniques are online forms of traditional political communication, others are unique to the internet. Unlike the others, they exploit its possibilities to communicate political information in a way that traditional media cannot do.

1.1 Voting Advice Applications

An example of these techniques are the online tools known as Voting Advice Applications (VAAs). These are websites launched during elections where voters can compare their views on with those of the political parties. A VAA does this by requiring both the user and the party to respond to a common questionnaire. This questionnaire contains questions relevant to the election. The topics it covers try to distinguish between the various parties included in the VAA. The number of questions, as well as the range of topics, differs between each VAA. In the end, the VAA combines the responses of the user with those of the political parties and calculates a match between them. This match is also known as the *advice* of the VAA and is where the VAA derives its name from. The VAA can present this advice in different ways. Examples are bar graphs, two-dimensional political maps or spider graphs. Which one it uses depends on the goal of the designer. The difference between the information a VAA provides and the information that voters receive from other forms of political communication is that the latter is general, while the information the VAA provides is specific. Information from a party manifesto will be the same for everyone, while the information given by a VAA will be unique for each user. This is because the VAA bases its advice on the users' responses to the questions in the VAA. As the VAA tailors its advice to the user, using VAAs lowers the cost of information. Partly because of this, the usage of VAAs has taken a flight over the previous decade. In countries like Germany, the Netherlands and Switzerland, up to 40% of voters uses a VAA during elections (Marschall 2014).

VAAs can benefit society in two ways. First, they can increase the level of political interest and political knowledge of the voters. Second, they can increase the level of electoral participation and electoral turnout. About the first, Kamoen, Holleman, Krouwel, et al. (2015) and Schultze (2014) show that VAAs improve the awareness voters have of which issues are relevant during the elections, while Marschall (2005) and Marschall and Schmidt (2010) find that they can inspire voters to look for further information. About the second, the current consensus is VAAs increase the chance that their users go out to vote (Dinas,

Trechsel, and Vassil 2014; Gemenis and Rosema 2014; Garzia, De Angelis, and Pianzola 2014; Garzia, Trechsel, and De Angelis 2017; Germann and Gemenis 2018). Another effect VAAs have is they are able to influence the party choice of its users. This effect ranges from 2 to 3 percent in Belgium (Nuytemans, Walgrave, and Deschouwer 2010; Walgrave, Aelst, and Nuytemans 2008b) to about 6 percent in Germany (Marschall 2005) to over 10 percent in Switzerland (Ladner, Felder, and Fivaz 2010; Ladner, Fivaz, and Pianzola 2012). This effect is particularly strong for voters who doubt between certain parties or who have not decided upon a single party yet (Kleinnijenhuis et al. 2017).

1.2 Design Difficulties

Voting Advice Applications are thus suitable to help counter the downward trend of electoral turnout and political knowledge of voters. But, especially because VAAs do influence turnout, knowledge, and party choice, it becomes relevant whether the advice they provide is "correct". In other words, do VAAs match the users to those parties closest to them? It is here the design aspect becomes relevant. For though the basic structure of the VAA is simple, there are different ways in which the VAA can arrive at a match. Designers can alternate the way in which they position the parties, calculate the match between user and party, or visualize the advice. Moreover, they can decide whether to include or exclude certain parties and which issues to include in the questionnaire of the VAA. Each of these decisions influences the design of the VAA and the advice it provides.

For the statements, from early on it was clear that the advice the VAA gives depends on the type of statements included in the VAA (Lefevere and Walgrave 2014; Walgrave, Nuytemans, and Pepermans 2009). Other factors are the wording of the questions (Holleman, Kamoen, Krouwel, et al. 2016), and the type of available response options (Baka, Figgou, and Triga 2012; Rosema and Louwerse 2016). For example, Lefevere and Walgrave (2014) show that including more left-right statements advantages parties with more extreme positions on this dimension, while Holleman, Kamoen, Krouwel, et al. (2016) find users respond different to questions worded positive or negative. Moreover, the way designers position the parties on the statements varies as well. To begin with, the VAA designers have to decide which parties to include. This decision alone might lead VAA users to ignore certain parties as the VAA does allow the user to compare themselves with them. This threatens the supposed "neutrality" of VAAs (Fossen, Anderson, and Tiemeijer 2012). Then, to position the parties, the designers have several options. The most common option is to have the parties position themselves. Yet, while simple, this might cause parties to lie about their true position to gain a more favourable position in the VAA (Ramonaitė 2010). For example, a party that holds an unfavourable position on taxes might choose to be neutral or not respond to a question on taxes. This can increase the number of matches it receives. Other methods, like a combination between party self-placement and expert positioning (Krouwel, Vitiello, and Wall 2012; Krouwel and Elfrinkhof 2014) or an iterative approach in which coders decide on positions over a number of rounds (Gemenis 2015), try to address these problems. Yet, unclear decision-making and high costs plague them. Finally, designers can use different algorithms to calculate the match between the user and the party (Mendez 2012, 2014b). Finally, in the case where the VAA present the match in the form of a political map, it matters which

spatial model the designers choose to construct it (Louwerse and Rosema 2014).

VAAs are thus by no means *neutral*: the decisions taken by the designers shape the way the VAA functions and the advice it produces. While most VAA designers accept this, few of them seem to feel the need to explain to the user why they made certain design choices or explain what their effect is. Moreover, the perceived neutrality is also challenged when one considers VAAs rely on a single model of democracy: that of social choice (Anderson and Fossen 2014). Other forms of democracy, like deliberative or contestatory democracy, are not included in the picture VAAs give of the electoral context. The picture of the political landscape VAAs promise to deliver is thus not complete. In the worst case, this could lead to the VAA undermining voter competence instead of strengthening it by presenting a simplistic idea of what the electoral context looks like (Fossen and Brink 2015). Moreover, the increasing number of VAAs designed by (commercial) media organizations raises the question if they design the VAA for the benefit of the voter, or to provide the organization with free data (Laaksonen, Nelimarkka, and Haapoja 2016, pp.11-13).

1.3 A Proposal

So, does this mean we should give up on VAAs? While VAAs do have problems, their unique approach to provide voters with custom information makes them suited to serve as tools for voter empowerment in the digital age. The question is thus if can we help VAAs to overcome some problems that plague their design and suggest solutions on how to address them? In other words: *is it possible to improve VAAs so that they can fulfil their promises?* While addressing all potential problems is impossible, here I will attempt to make a start by looking at the most prominent design parts of the VAA. To do so, I will consider the design of a VAA not as a series of separate steps but as a design process in which each of the steps influences the other. How this process looks like and what its benefits are, is the subject of the next chapter. Here I also go deeper into the history of VAAs and their problems and will provide an overview of the various types of VAAs that are available.

2 | Theory & Framework

In this chapter, I will set up the thesis and provide all the material needed for the other chapters. I will start by telling the story of Voting Advice Applications, where they came from, how they have developed, and where they are going. I will also touch upon several points that have been at the centre of the debate that has surrounded VAAs since their start. Then, I will turn to the various types of VAAs that exist. I will discuss the many design choices the designers have to make and the many options they have. Using cluster analysis, I will then arrive at 6 different types of VAAs. These types reveal two important design choices: the questionnaire and the visualization. It is at this point that I will introduce the main research questions that will guide this thesis. Afterwards, I will discuss the framework that structures the thesis. I will conclude with a brief overview of what the next chapters will bring.

2.1 The Story of Voting Advice Applications

The story of VAAs starts in 1989 with the launch of *Stemwijzer* as a civic education tool for high school students in the Netherlands. Consisting of a book and a floppy-disk, Stemwijzer aimed to increase the political knowledge of the students. It also tried to shift their attention from the personality of the candidates to the contents of their manifestos (Fossen, Anderson, and Tiemeijer 2012). After the successful online implementation of a similar tool in Finland in 1996, Stemwijzer received its first internet-based version in 1998, drawing 6500 visitors (Graaf 2010, p.36). Yet, it was the rise of Pim Fortuyn and his LPF during the elections of 2002 that brought about the definitive breakthrough. Curious how Fortuyn's new party related to the traditional parties, *Stemwijzer* voters visited more than 2 million times during the 11 weeks leading up to the election. During snap elections a few months later, Stemwijzer reached a similar number of users in only 7 weeks (Graaf 2010, p.42). Based on this success, the German *Wahl-O-Mat* copied the *Stemwijzer* concept during the 2002 elections in Germany. Reaching 3.6 million users, designers in Bulgaria, Flanders and Switzerland soon followed. At the same time, the first alternatives appeared. In Switzerland, *smartvote* challenged the *Stemwijzer* based *Politarena* , while in the Netherlands *Stemwijzer* faced competition from *Kieskompas*. The latter attempted to address some criticisms aimed at VAAs that had started to appear since 2002 (Groot 2003c,b). Most important, it introduced a system of checks on the answers that parties gave to the questions. These answers were now checked by the designers. This meant parties could no longer make their positions appear different than from what they were. *Kieskompas* also introduced a two-dimensional political map, instead of a simple listing of

percentages as *Stemwijzer* had done. Immediately after its launch it was as successful as *Stemwijzer* and remains its main competitor. Not long after the launch of *Kieskompas* the first VAA in the European context appeared in 2009. The *EU Profiler* offered versions for 30 countries and attracted 2.5 million users. In 2014 *euandi* and *EUVox* followed in up and were also successful (Garzia, De Angelis, and Pianzola 2014). At the same time, VAAs began appearing for national referendums, student union councils, and elections for the chamber of commerce. Besides, they spread outside of Europe and made the transition to the East-Asian context with versions appearing in Taiwan, Japan and South Korea (Liao and Chen 2016). Currently, the appearance of *Stemwijzer* indicates the start of the electoral campaign in the Netherlands and leaders of the political parties attend its launch. In Flanders, VAAs have at times had their own television show, while in Germany, more than 13 million users used the *Wahl-O-Mat* during the elections in 2013.

The success of VAAs has also opened them for criticism. Reacting to the finding that VAAs can influence vote choice (Wall et al. 2009; Mahéo 2016), political parties started to question the choices of the designers of the VAAs. They were especially critical when they feared the VAA might be disadvantageous to them. This was the case in the Netherlands during the elections of 2017. Then, the party "Geen Peil" objected to the way the designers of *Stemwijzer* positioned their party. After a much-publicised debate, the party requested that *Stemwijzer* drop them from the VAA. This happened soon thereafter. I myself received a complaint from a party that was not included in the *Stem-Consult* VAA launched during the same elections. The party argued that because it was not included in the VAA, this affected its chances during the elections. And this was at odds with the stated intents of the VAA to be neutral and independent. Accompanying these issues was an increased media focus on the security of the VAA websites themselves. After revelations over vulnerabilities by a collective of ethical hackers, *Stemwijzer* decided to re-evaluate its security and *Kieskompas*, launched a few weeks later, placed emphasis on its increased security.

Another source of criticism came from the community of VAA designers themselves. One strand of criticism was conceptual and focused the simplistic view of politics most VAAs adopt. Of the three separate linkages between parties and voters — programmatic, clientelistic and charismatic — VAAs only focus on the programmatic content. Thus, they only capture one side of the political space. Also, VAAs even simplify this programmatic link due to the limited number of questions and the limited response format. Another critique is that users can combine responses in which a future government spends more while taking less. Impossible to achieve in reality, this can make voters disillusioned by reality. Not limited in their responses, they have unrealistic assumptions about what parties can do (Korthals and Levels 2016). Likewise, the issues appearing in the VAAs are often ambiguously formulated, not discriminating enough to allow for a meaningful distinction between the various parties (Fossen, Anderson, and Tiemeijer 2012), and not representative of the current issues at play (Groot 2003b). Besides, the issues included were often the more traditional issues. Issues on raising taxes are almost always present, while less conventional issues — like the country withdrawing from NATO — are rarely included. New parties focusing on such issues are thus unable to distinguish themselves in VAAs. The most serious conceptual critique is the presumed neutrality of VAAs. Not only did parties try to influence their position in the VAA — regardless of whether this would fit with their actual ideas (Ramonaitė 2010) —

but VAAs also presented users with a biased picture of the political space (Fossen and Brink 2015). The basis of this picture is the underlying idea the designers have about the political space — ideas of which they may or may not be aware of. One of the results of this is that VAAs only focus on a social choice picture of democracy. Other options, like deliberative and contestatory models, are thus left aside (Fossen and Anderson 2014).

Another line of criticism is methodological. This is the critique stemming from debates about the many design choices designers of VAAs have to make. For example, as response options, designers can offer anything from a *agree-disagree* to a five-point Likert scale to a ten-point thermometer. They can position parties by asking the parties themselves, read the party manifestos or use the positions of experts. They can use different algorithms to calculate the match between the user and the party and can show these in a wide range of visualizations. None of these choices follows from the other. This would not be a problem, be it not that the choices themselves are not neutral. In other words, the design choices of the VAA not only influence its look and feel but also its results. The type of questions and response scales (Lefevere and Walgrave 2014), the methods of party positions and the calculation of the matches (Gemenis 2015; Mendez 2012; Louwerse and Rosema 2014) all influence the advice the user receives. Early on in VAA research, Walgrave, Nuytemans, and Pepermans (2009) showed that different selections of questions led to different matches for the user. More important, these selections benefited or disadvantaged certain parties. Besides, the formulation of the questions and their complexity can cause different responses from the user (Gemenis 2013b; Holleman, Kamoen, Krouwel, et al. 2016; Camp, Lefevere, and Walgrave 2014). Also, Rosema and Louwerse (2016) show that the number of response categories also influences the response of the user. The algorithm that matches the user and the party is not neutral either. Not only are there different ways of how one can see the match, but each of these ways can also lead to a different result (Mendez 2017). Besides, visualizing the result can be problematic as well. This is especially the case for the VAAs using a political map. They often assume this map is similar for different countries - which is most likely not the case (Otjes and Louwerse 2014). Even within a single country, the original assumptions of the political space might not be compatible with how it actually looks like (Germann, Mendez, et al. 2015; Germann and Mendez 2016). Yet, even if designers are able to navigate through all these problems, the usefulness of the VAA depends on the user. For example, Alvarez, Levin, Mair, et al. (2014) find that sex, education, interest in new technologies, and political knowledge influence how users perceive the VAA. Users who embrace new technologies, users who believe that politics is complex, and users with a high political interest found the VAA more useful than other groups. Also, users at the extreme ends of the left-right scale, as well as those positioned in the centre, viewed the VAA as less useful. This is either because of their strong convictions or due to their weaker party attachments.

2.2 The Relevance of Design

The message is thus that the design of the VAA matters. Also, the design of the VAA is not a neutral affair. Through their choices, VAA designers can advantage of disadvantage certain parties. To better understand how they do so, I will now discuss the four main design choices designers of VAAs have to make. These are the choices related to: a) the questionnaire, b)

the party positioning, c) the matching, and d) the visualization. I will discuss each of these choices and discuss the options that VAA designers have.

2.2.1 Questionnaire

VAAs use questionnaires to measure the political position of the user. How the designer organizes their questionnaire is the first choice. Different sets of questions (Camp, Lefevere, and Walgrave 2014; Lefevere and Walgrave 2014; Walgrave, Nuytemans, and Pepermans 2009), different response scales (Rosema and Louwerse 2016) and the availability of a neutral position (Baka, Figgou, and Triga 2012) all influence way in which users respond to the questionnaire. There are four main aspects of a questionnaire: the response format, the number of response positions, the number of items, and if there is a "do not know"-option.

Number of Items. Most VAA questionnaires contain between 25 and 30 items. Some VAAs like the Swiss *Smartvote* offer the user a choice between a short VAA of 32 questions and a longer VAA of 59 questions. The number of questions in the questionnaire decides on how many topics the VAA can match them to a party. While longer questionnaires allow for a higher degree of accuracy, they also cause user fatigue. This can lead to users starting to satisfice to finish the VAA (Galesic and Bosnjak 2009; Blasius and Thiessen 2012). The result is less accurate responses leading to less accurate advice.

Vaalikone 2011	EUProfiler 2009
Should Finland seek membership in NATO?	*Finland should apply for NATO membership*
○ Yes, during this term	○ Completely agree
○ Yes, but not during this term	○ Tend to Agree
○ Yes, if Sweden is applying	○ Neutral
○ No, not during this term	○ Tend to disagree
○ Never	○ Completely disagree

Table 2.1: Differences between options and scale responses

Response format. Designers can use different response formats to measure the responses of the users. VAAs offer either options or scale responses to their items. Options mean that the user can choose from various answers. Scale responses mean that the user can say if (and sometimes to which degree) they agree or disagree with the questions. Table 2.1 shows an example for two Finnish VAAs that aim to measure the position of the user to the Finnish membership of NATO. *Vaalikone* offers the user five different options. Three of these options are affirmative (the user would like Finland to join NATO). Yet, each of these options has a different condition. Two options are negative, though one option delays the decision, while the other is definite. *EUProfiler* offers a five-point scale in which users can give up to five degrees of agreement. The main advantage of options is that these are the actual positions of the parties. Designers of VAAs can thus invite parties for their responses to a question and use these answers as the options. This allows for an exact match between what the user thinks are the best answers to the questions and an actual position of a party. Besides, using options instead of scales makes clear to the user that answers to political questions are not

clear-cut. Often, a user may wish to agree with a certain statement, but only on certain conditions, which the options can offer, while a scale cannot. Thus, options allow for more complex questions than the statements used for most scales.

Do Not Know. Several VAAs include a "do not know" option. In most, the result of choosing this would be like not responding. It is thus not a substantive answer as it does not provide any information about the position of the user. Note that this is different from the "neutral" option that many VAAs offer, which indicates the users can see both sides of the argument. The reason for designers of VAAs to include a "do not know" option is to prevent users from giving a "pseudo opinion". This happens when users feel pressured to provide an answer to a question they do not actually have an opinion on (Corbetta 2003, p. 135). Yet, with controversial or difficult questions, users might use the "do not know" option as a safe haven. As VAAs need a certain number of substantive responses, they often prohibit the user from choosing this option too often.

Number of Response Positions. Designers of VAAs use different numbers of response positions for users to choose from. For example, the Dutch *Stemwijzer* provides users with three options: "agree", "disagree", or "neither", resulting in a 3-point scale. Another VAA, the Austrian *Wahlkabine* offers only two options: "yes" or "no", resulting in a binary scale. VAAs based on the *Kieskompas* format offer five options in Likert-style format, ranging from "completely agree" to "completely disagree", with a central "neutral" option. While these three options are the most frequent, others are possible. The Irish *Pick Your Party* from 2007, for instance, allowed users to choose from a 21-point scale, with 10.5 as a middle option. The *De Stem van Vlaanderen*, a Flemish VAA from 2014, allowed for 101 response options by means of a slider. The number of response options is thus the result of a trade-off between simplicity and validity and reliability of the answers. Given too few response options, users might not have the opportunity to choose the option that represents their opinion. Given too many response options, they might fail to distinguish between them and choose arbitrarily (Nadler, Weston, and Voyles 2015; Krosnick and Presser 2010). Also, the designers have to decide whether to include a neutral option. Inclusion of such an option is a subject of discussion (Corbetta 2003, p. 168). While the option might allow users who can actually see both sides of the argument to give their opinion, it also allows those who do not want to state their actual opinion a place to hide. Yet, Baka, Figgou, and Triga (2012) found that when provided both a neutral option and "do not know", users chose the former as they felt it was more socially desirable. In that case, the neutral option does not record the user's genuine opinion. Also, Nadler, Weston, and Voyles (2015) notes that the true meaning of the neutral response category is often far from clear, with interpretations ranging between "no opinion", "don't care", "unsure", "neutral", "equal/both", and "neither".

2.2.2 Party Positions

Table 2.2 shows the different ways in which VAA designers can position parties. The first three methods depend on the manifestos political parties publish in the run-up to the elections. The first two of them rely on "salience theory", which holds that the more emphasis a party places on the issue, the more important it is for them. The *Manifesto Project* uses human coders and an elaborate codebook to hand-code these issues, while methods of *automated*

Type	Based on...
Comparative Manifesto Project	Salience of issues in manifestos (Budge 2001)
Automated Content Analysis	Counting of words in textual data (Grimmer and Stewart 2013)
Judgemental Coding	Party manifestos or proxy documents (Janda 1980)
Roll Call Method	Voting record of the party or candidate during the previous period (Ansolabehere, Snyder, and Stewart 2001)
Expert survey	Interviews or surveys of party experts (Mair 2001)
Candidate survey	Interviews or correspondence with individual candidates (Burden 2004; Francia and Herrnson 2007)
Elite survey	Interviews or correspondence with party elites (Power and Zucco Jr. 2009)
Voter survey	Surveys among voters (Eisinga and Franses 1996; Evans, Heath, and Lalljee 1996)
Kieskompas-method	Combination of elite survey and judgemental coding (Krouwel and Pol 2014)
Iterative Expert Survey	Iterative rounds of expert surveys (Gemenis 2015)

Based on Krouwel and Pol (2014), Gemenis (2015), and Benoit and Laver (2006).

Table 2.2: Overview of different of methods for obtaining party positions

content analysis use dictionaries (Laver and Garry 2000) or scaling methods (Laver, Benoit, and Garry 2003; Slapin and Proksch 2008). Still, not only is it questionable whether the frequency of mentioning an issue tells much about the position of that party (Laver 2001), both the Manifesto Project (Dinas and Gemenis 2010; Gemenis 2012; Hansen 2008) and various methods of automated content analysis have serious methodological issues (Budge and Pennings 2007; Grimmer and Stewart 2013; Bruinsma and Gemenis 2019). The third method, *judgemental coding*, instead requires coders to position the manifestos based on the whole text. This allows the coder to read between the lines and does not limit them to a rigid coding scheme. Yet, this type of coding is still dependent on a single coder. This is a problem as interpreting what a certain paragraph means is a very subjective exercise (Krippendorff 2004; Riffe, Lacy, and Fico 2005), and disagreement between coders can be persistent and hard to solve (Gemenis 2013b). Another method based on text is to establish positions based on *roll call votes*. This means the position of the party depends on their recorded votes in parliament (Clinton, Jackman, and Rivers 2004). Based on the assumption that history will repeat itself, the advantage of roll-call data is that there is often a large sample of data available on many topics. Even so, not all votes are on roll-call, and designers cannot include new issues or parties as there is no data on them. Also, parties can use the roll call strategically, which makes them less reliable (Hug 2010).

Instead of texts, we can also position parties data, parties based on interviews. These can

be interviews with *experts*, *candidates*, *elites*, or *voters*. *Expert* surveys have the advantage that they are cost-effective, easy to set-up and can cover a wide range of issues. This makes expert surveys the "gold standard" of party positioning (Marks et al. 2007). Still, experts might be selective in what sources they used to base their positions on (Krouwel and Pol 2014) and might have problems positioning new parties they know little about. *Elite* or *candidate* surveys might address the latter problem, but assumes that parties are unitary actors (elite) or know their own positions (candidate) (Krouwel and Pol 2014). Moreover, parties might be unclear about their opinion on purpose to get a more favourable position on controversial issues (Rovny 2012). While *Voter* surveys might clarify such blurring, they can suffer from a low-informed public (Steenbergen and Marks 2007; Evans, Heath, and Lalljee 1996).

The *Kieskompas-method* aims to address the problems with judgemental coding and elite surveys by combining them (Krouwel and Pol 2014). Thus, both party elites and the VAA designers position the party on the issues. If there is any disagreement, the designers point this out to the party elite and allow for a revision. This process reiterates until there is an agreement between them. While this method clarifies certain positions of the party and prevents them from being dishonest, it does not address the problem that party and designers can still have a different interpretation of the issues. Also, if parties and scholars fail to agree, the scholars have the final word. These often unstructured discussions are very dependent on the roles and positions of the various group members. Status, power, and prestige might be more important than actual knowledge, threatening the quality of the final decision (Gemenis 2015). To address this problem, Gemenis (2015) introduced the *iterative expert survey*. This technique, based on the Delphi method often used for policy forecasting (Dalkey 1969; Linstone and Turoff 1975), has multiple rounds in which experts position parties on a series of issues. After each round, information about the previous round, like the positions of the other experts and their justifications, are anonymously fed back to each of the experts. This process repeats until there is a desired level of agreement. While the advantages of this method are the discussion between experts is more structured and transparent, the disadvantage is that the studies are often time intensive and take certain knowledge to set up and moderate.

2.2.3 Matching

Matching is what differentiates VAAs from regular questionnaires. During the matching phase, the VAA compares the responses of the user to those of the parties included in the VAA. As with the positioning of the users, the matching can has an effect on the final advice of the VAA (Mendez and Wheatley 2014). Apart from the matching algorithm, it is relevant whether users and parties can apply *weighting*. These weights show if the topic carries significant importance for them, often doubling the impact of the response on the final match. As such, respondents who feel strong on environmental issues and attach more weight to them are more often matched to Green parties than would otherwise be the case. The choice of matching method is dependent on the way the designer thinks about the political space. Here, there are two different matching methods: *low-dimensional* and *high-dimensional*. In the latter, each of the questions or items in the VAA forms its own individual dimension, while the former clusters many items on overarching dimensions such as left-right or conservative-progressive. This leads to four options for designers. First, there is the proportion method,

which counts the proportion of statements where user and party agree. This method can use with three different metrics: the *agreement method*, *city-block distance* and *Euclidean distance* (Louwerse and Rosema 2014). The agreement method is the simplest:

$$A_{uv} = \sum_{i=1}^{n} w_{ui}\ a_{uvi} \tag{2.1}$$

here, a_{uvi} equals 1 when user u and party v provide a similar answer to statement i and equals 0 when the answers are dissimilar, n is the number of items in the VAA questionnaire, while w_{ui} equals the weight the user can assign to an issue. Another way of calculating the distances is by considering the positions parties have on an issue and then calculate the *city-block distance*. Here, the distance between two points is the sum of the absolute differences of their positions in space:

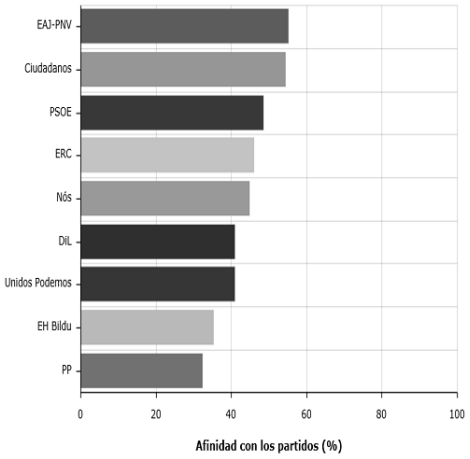
$$CB_{uv} = \sum_{i=1}^{n}\ w_{ui}|p_{ui} - p_{vi}| \tag{2.2}$$

where p_{ui} is the voters' position on issue i and p_{vi} the party's position, n is the number of questions in the VAA questionnaire, and w_{ri} is the weight the user assigns to the question. Finally, we can calculate the *Euclidean* distance between user and party. This distance is the straight line between the two points occupied by the user and the party on the two-dimensional political map:

$$E_{uv} = \sqrt{\sum_{i=1}^{n} w_{ui}(p_{ui} - p_{vi})^2} \tag{2.3}$$

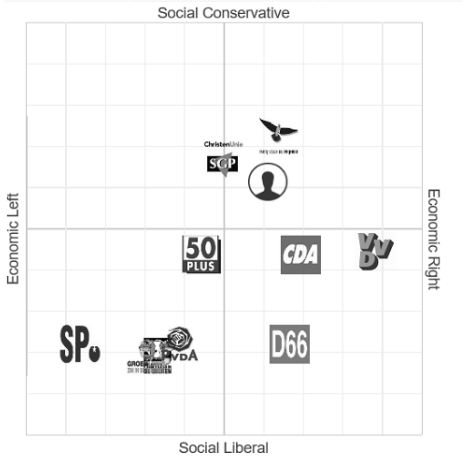
where p_{ui} is the voters' position on issue i and p_{vi} the party's position, n is the number of questions in the VAA questionnaire, and w_{ri} is the weight the user assigns to the question. Which metric designers use depends again on their assumptions on what the political space looks like. Benoit and Laver (2006) show that using *city-block* logic, one assumes the individual dimensions are independent of another. A position on one dimension says nothing about a position on the other. Euclidean logic meanwhile assumes the dimensions are dependent on another. Choosing a different metric can lead to different matches and thus to different results (Louwerse and Rosema 2014). The other three matching methods based on the *low-dimensional* model differ by the number of dimensions they use. The simplest of these is the one-dimensional space, in which the designer places all questions on a single dimension. Another method uses two dimensions to create a plot on which it positions both parties and users. Yet another method does something similar but uses many dimensions (for example 8 in the Swiss *Smartvote*) to position users and parties. In all three cases, designers have to combine separate issues into one or more dimensions. They can do this either *ex ante*, by assigning each issue to a certain dimension, or *ex post* by the use of dimension reduction methods. In both cases, the VAA calculates the final positions of both parties and users by summing the scores for the questions on each dimension and plotting them in the political map Mendez and Wheatley (2014).

2.2.4 Visualisation

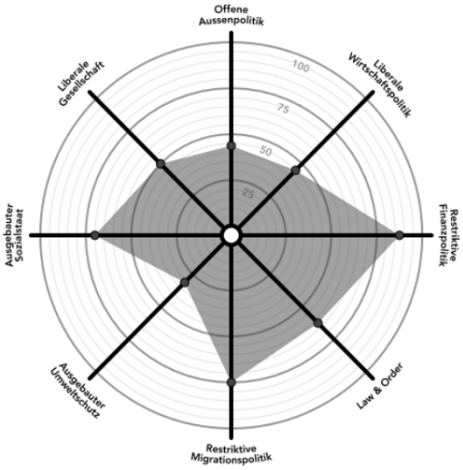


(a) Rank Order (Quantified)

(b) Alternative Rank Order (Quantified)



(c) 2-Dimensional Map



(d) 7-Dimensional Spidergraph

Figure 2.1: Four different types of visualisations employed by VAAs from the Netherlands, France, Spain and EUVox.

The visualization of the VAA influences the usefulness of the VAA to the user and the effect the VAA has on them. The choice of visualization depends on the matching method the VAA employs. For VAAs the most often used visualizations are: *single party*, *rank order (quantified)*, *rank order (not quantified)*, *2D-graph*, and *spider graph* (see Figure 2.1). The first is a mention of a single party or candidate that has the highest match with the user. The name of the party or candidate appears without the mention of others. The second and third options include all parties or candidates ranked in decreasing (or increasing) order. This can be either with a matching percentage or without. The fourth option shows a two-dimensional graph with both parties and user positioned on them. This does not give a direct match, but rather has the user find themselves in the political landscape. It thus leaves it up to them to decide which party is closest. The spider graph gives a web-like perspective of the user's position, overlaid with those of the parties. The more the areas of a particular party and those of the user overlap, the higher the degree of correspondence between the two.

2.3 Building an Inventory

While the previous discussion shows the possibilities the designers have, it does not answer the question of which possibilities they use. To know this, we would have to look at all VAAs launched and note down which choices they made. To do so, I will make an inventory of all VAAs I can find. This itself is problematic as VAAs are temporal events taken offline once the elections are over. This makes them hard to track down. Also, there is no principal authority or organization that coordinates VAAs. Thus, we do not even know what the total population is. To get a purposeful sample, we can then turn to several options. For example, we could search the internet for any mention of "VAA" or "Voting Advice Application", writing down the details and follow up leads if they appear. This would be analogous to *snowball sampling*. While this would give us a list of VAAs, we can not be sure whether the list is either exhaustive or comprehensive. Another approach would be to contact the designers of several well-known VAAs and ask them for a list of VAAs known to them. Yet, this would depend on the cooperation of the designers which could be difficult for older VAAs. So, I will use a third approach and use the already collected knowledge on the field of VAAs by the field itself. Thus, I turn to two overviews generated by two important players in the VAA field - the German *Bundeszentrale fur politische Bildung*[1] and the ECPR Research Network *Voting Advice Applications*[2]. The first is the governmental organization responsible for the German *Wahl-O-Mat*, while the second is the research network of VAA scholars from the European Consortium for Political Research (ECPR). Both have initiated overviews of VAAs on their respective websites constructed by enquiring their colleagues. Scholars can also contact both organizations to have new VAAs added to the list, update older information or correct mistakes. As a result, while neither can be exhaustive, they do represent the most complete list of VAAs available. Besides these lists, I will also include into the VAAs generated by the three largest consortia of VAAs: *PreferenceMatcher*[3], *Votematch*[4], and *Kieskompas*[5]. For further information, I do any of three things: a) visit the original website of the VAA, b. visit an archived version of the website[6], or c) take information on the VAA from relevant articles discussing the VAA. I count a VAA as unique as soon as there is a difference on at least one of the variables. This means I count VAAs designed for municipalities or regions a unique even though they are often launched at the same time and share a similar structure and design. Table 2.3 shows an overview of the codebook while Appendix A shows the codebook in full.

The final data-set contains 1099 different VAAs from 48 different countries. Figure 2.2 shows their distribution in a map. Here, we see that the majority of the VAAs originates from Europe. Other western democracies, like Australia, the United States, and Canada have VAAs as well, as do Japan and Taiwan in Asia. Also, several VAAs came from North African countries like Egypt, Sudan and Tunisia, as well as Brazil, Venezuela, Peru and

[1] `http://www.phil-fak.uni-duesseldorf.de/wahl-o-mat/en/links/`
[2] `http://vaa-research.net/`
[3] `http://www.preferencematcher.org/?page_id=18`
[4] `http://www.votematch.eu/`
[5] `https://home.kieskompas.nl/nl/wat-we-doen/`
[6] This is possible thanks to the *Wayback Machine*, maintained by the *Internet Archive* NGO, available at `https://web.archive.org/` and indexes and stores websites.

Section	Variable
Questionnaire	The **Response Format** of the VAA questionnaire
	The **Number of Response Positions**
	The **Number of Items** in the questionnaires
	Includes **"Do Not Know"**
Party Positioning	The **Source** of the party/candidate positions
Matching	Can **Users Weigh** their items?
	Measure for user-party **Distance**
	Number of Dimensions after aggregation
Visualisation	**Visualisation** method
Organisation	The **Designer** of the VAA
General Characteristics	**Year** of original appearance
	Name of the VAA
	Country for which the VAA was launched
	Platform on which the VAA was launched
	The total number of **Users** of the VAA
Electoral Characteristics	**Number of Parties or Candidates** in the VAA
	Number of Parties or Candidateson the ballot
	Number of Parties or Candidates in parliament
	Number of Electors
	Type of election for which the VAA was made

Table 2.3: Guiding questions for the coding of the different variables.

Election	N
European Parliament	153
Lower House	208
Municipality	464
Other	96
President	32
Region	110
Upper House	36

Table 2.4: Number of elections in the inventory

Mexico. Finland (402) contributes most VAAs, followed by the Netherlands (146), the United Kingdom (103) and the Czech Republic (53). Together, these four countries almost contribute half of the number of VAAs. This is because of the large number of VAAs designed for municipality elections (see Table 2.4). For example, 364 of the VAAs in Finland are for municipal elections of which 333 for the one in 2012. The same occurs for the Netherlands, were 66 of the VAAs were those designed for the municipal elections in 2010. This over-

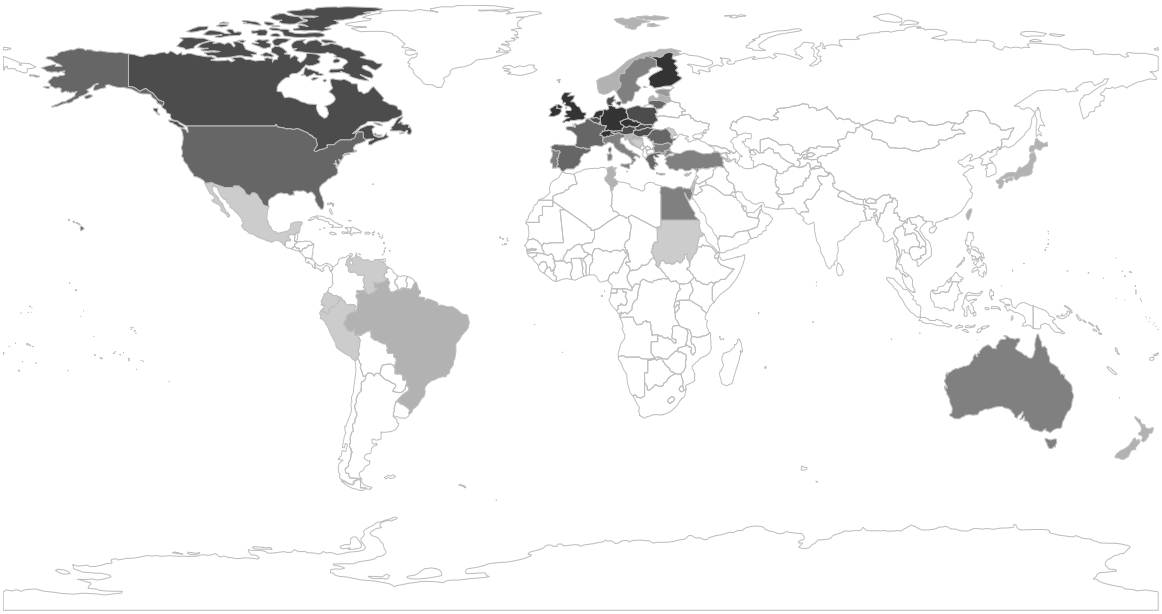


Figure 2.2: Distribution of the VAAs contained in the data-set. Darker colours indicate a higher number of VAAs which originated from that country.

representation is a result of the decision to count each VAA as separate as soon as they had a different number of parties or items. To address this problem, I assign weights to all entries that were versions of the same VAA. I did this by looking at the Type of Election, Country of Election (First Level) and Name of the VAA. I then identify variations by the Country of Election (Second Level). In the case of regions, the VAAs were only weighted if the launch occurred at the same time. I then calculate the weights based on the total number of VAAs in the group. Thus, when there are thirty VAAs in the group, each VAA gets a weight of $\frac{1}{30}$. All other VAAs get a weight of 1.

	University	Company	Media	NGO
University	63	–	–	–
Company	212	32	–	–
Media	17	21	412	–
NGO	188	87	2	260

Table 2.5: Collaboration between designers. Numbers on the diagonal are the number of times a VAA had only a single designer.

Table 2.5 shows the designers of the VAAs and their collaborations. Most collaborations took place between universities and companies and between universities and NGOs. Most media companies worked on their own when developing a VAA. In total, in 767 cases a single developer was responsible for a VAA.

2.3.1 A First Look: Applying MCA

For a better understanding of the data-set, I will run a Multiple Correspondence Analysis (MCA). MCA is a method to visualize tables based on the idea that a simple visualization

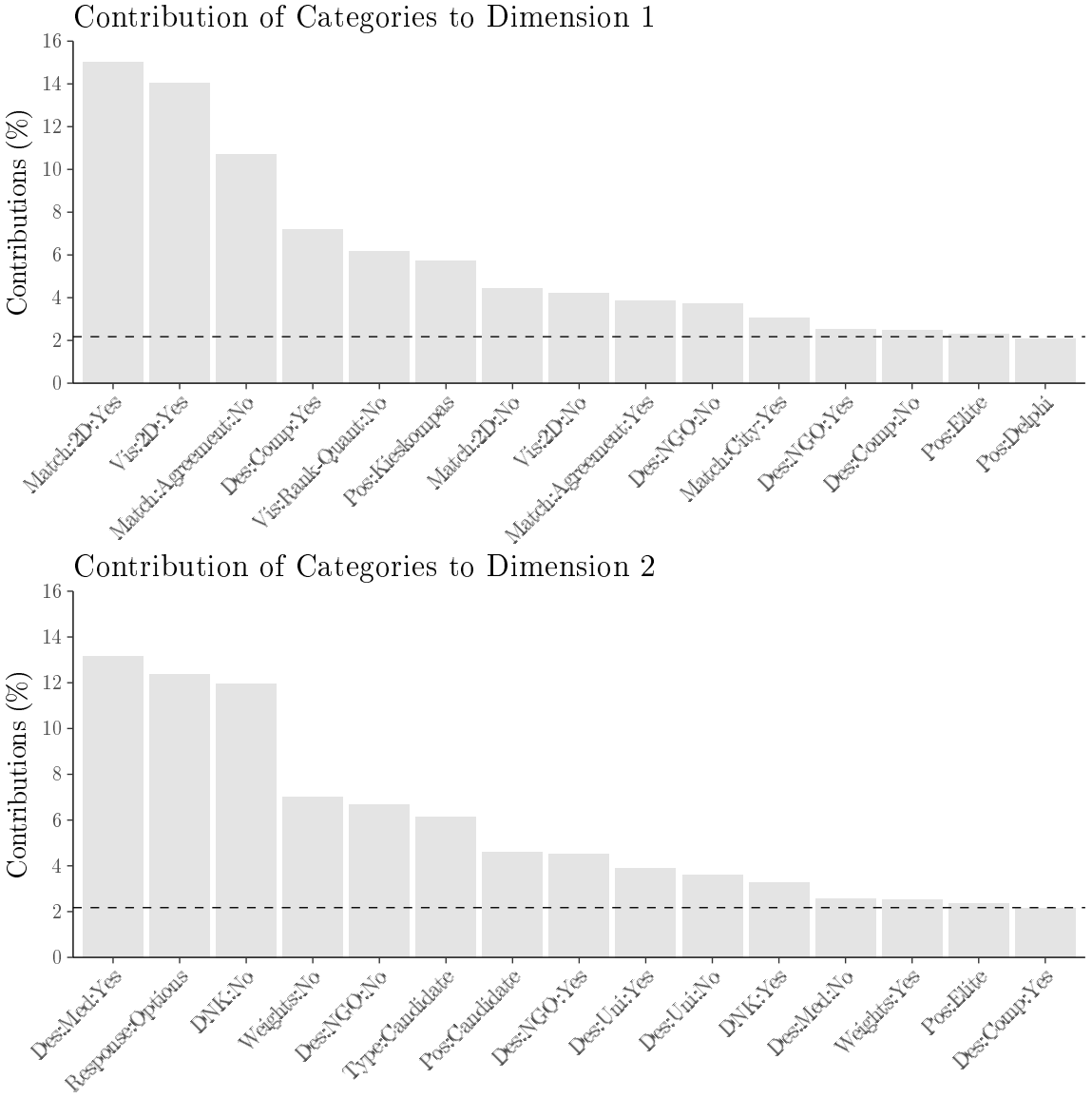


Figure 2.3: Contribution of the categories to the first two dimensions. The figure shows only the first 15 categories for each of the dimensions. The dashed reference line refers to the expected value if the contributions were uniform.

reveals more about the data-set than would at first meet the eye. As I will use MCA for later chapters as well, I will discuss the method more in-depth in the next chapter. For now, we only need to know that MCA aims to visualize the variance in the data-set. It does so by looking for those dimensions that explain as a large a part of this variance as possible. MCA then uses these dimensions to draw a two-dimensional map. To carry out MCA, I use the `FactoMineR` package (Lê, Josse, and Husson 2008) in R. This package allows me to identify quantitative and qualitative variables as well as row weights. Running the MCA results in a first dimension containing 38.63% and a second dimension containing 15.81% of the variance. To see what these dimensions mean, Figure 2.3 shows the contribution of each of the categories. The higher the contribution of a category, the higher its association with it. For the first dimension, two-dimensional matching and visualization showed the

highest contribution. For the second dimension, the highest contributions came from Media companies, response options and not offering a "Did Not Know" response. From this, we can interpret the first dimension as distinguishing between different types of matching and visualisation, while the second dimension distinguishes between different options related to how the questionnaire looks like and the type of designers that are behind the VAA.

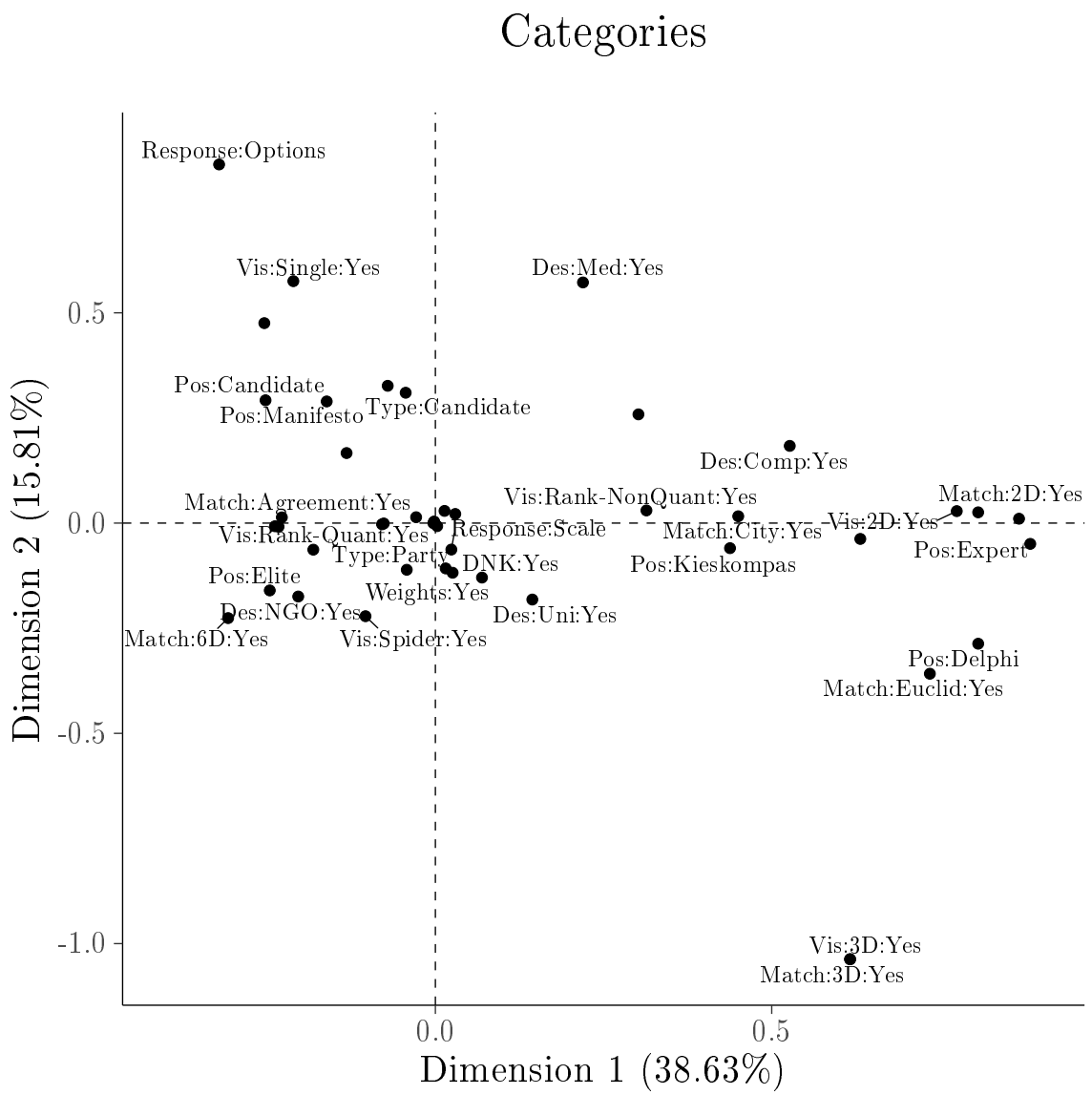


Figure 2.4: Plot for the categories on the 1st and 2nd dimensions. Note that labels are only shown for those categories where a feature is present for a VAA.

To further investigate this, Figure 2.4 plots the categories. For the first dimension, we can distinguish two broad clusters. The first starts on the left and contains response options and six-dimensional visualizations. Going to the centre, it also includes agreement-matching and elite, manifesto and candidate positioning. The other cluster starts to the right of the centre and extends to the right end of the dimension. This cluster contains Delphi, expert and Kieskompas positioning as well as two-dimensional matching and visualisation. For the second dimension, we find the use of response at the top and three-dimensional

matching and visualisation at the bottom. Also, a single visualisation and design by a media company belong to this dimension. Yet, labelling these dimensions is difficult. For the first dimension, we could say it distinguishes between Kieskompas-inspired VAAs featuring two-dimensional visualisations and party positioning methods and VAAs that use response options, candidate positioning and six-dimensional matching. The latter are characteristics that fit with VAAs developed in Switzerland and Finland. These VAA most often focus on individual candidates. The second dimension distinguishes between response options and three-dimensional visualisations. Both are rare in VAAs and are uncharacteristic of them most of the time. Also, the upper part of the map contains straightforward methods of visualisation and positioning, while the lower part contains more complex methods. As such, we could see this dimension as distinguishing between complex and non-complex types of VAAs.

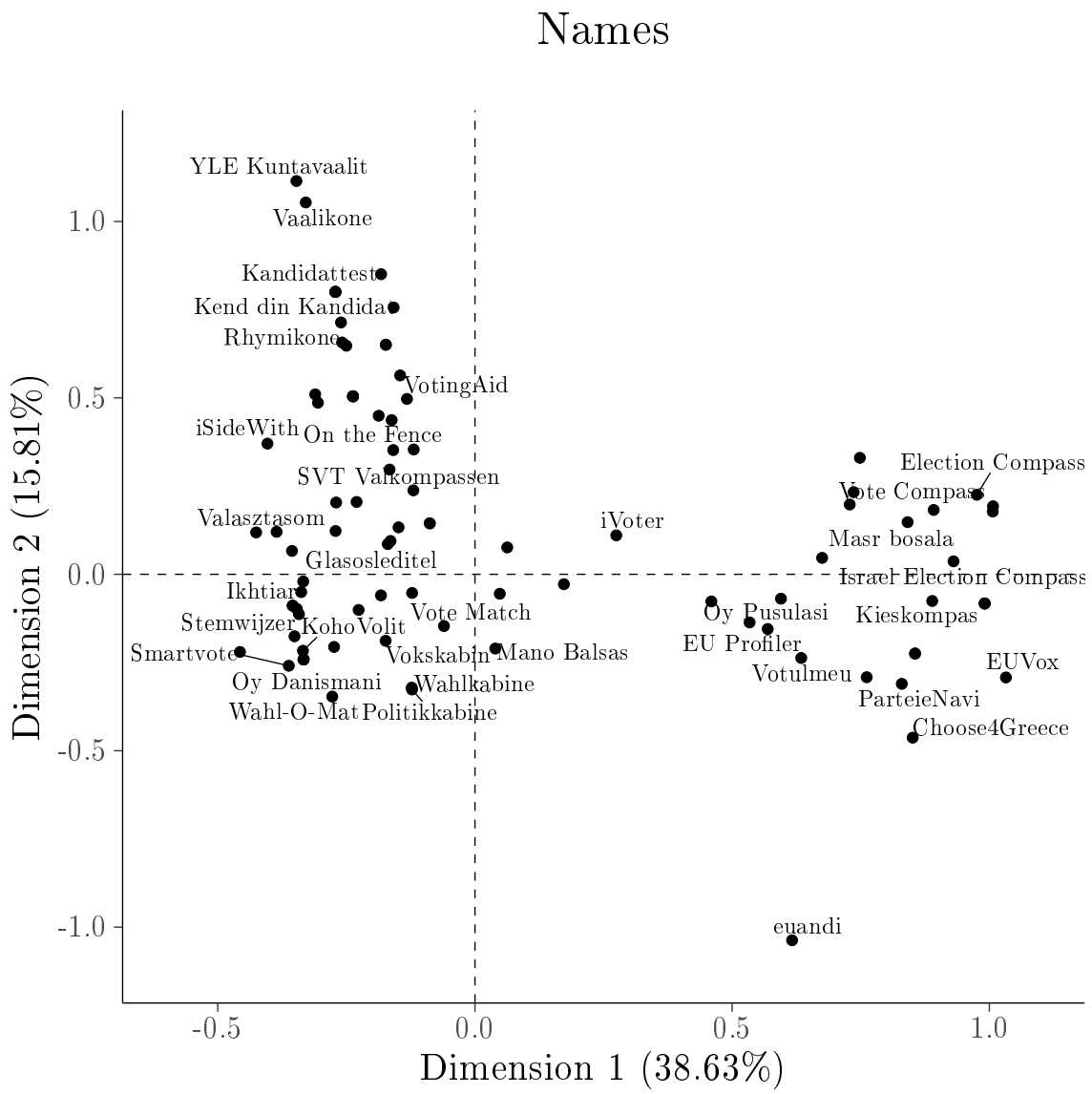


Figure 2.5: Plot for the individual VAAs on the 1st and 2nd dimensions. Note that labels are only shown for some of the VAAs to increase legibility.

Figure 2.5 supports this idea. This figure shows that the first dimension distinguishes VAAs based on the *Kieskompas* template from the others. The second dimension distinguishes the "YLE Kuntavaalit" and "Vaalikone" on the on hand from the "euandi" VAA on the other. The first two focused on candidates for the municipal elections in Finland and offered response options instead of response scales. The "euandi" focused on the European Parliament elections of 2014 and offered a standard Likert-response scale with a three-dimensional visualisation. As it was the only VAA to ever include such a visualisation, the position of the VAA is far from the others. Looking further we can distinguish on the upper side of the axis a group located in the quadrant formed by the null-line and an imaginary line drawn at -0.5. These VAAs are all related to the *Stemwijzer* family. Their main characteristic is elite positioning and simple rank-style visualisations. We can identify another group on the other side, which contains VAAs focussing on options and the individual candidates. I will label these the "Finnish" type of VAA.

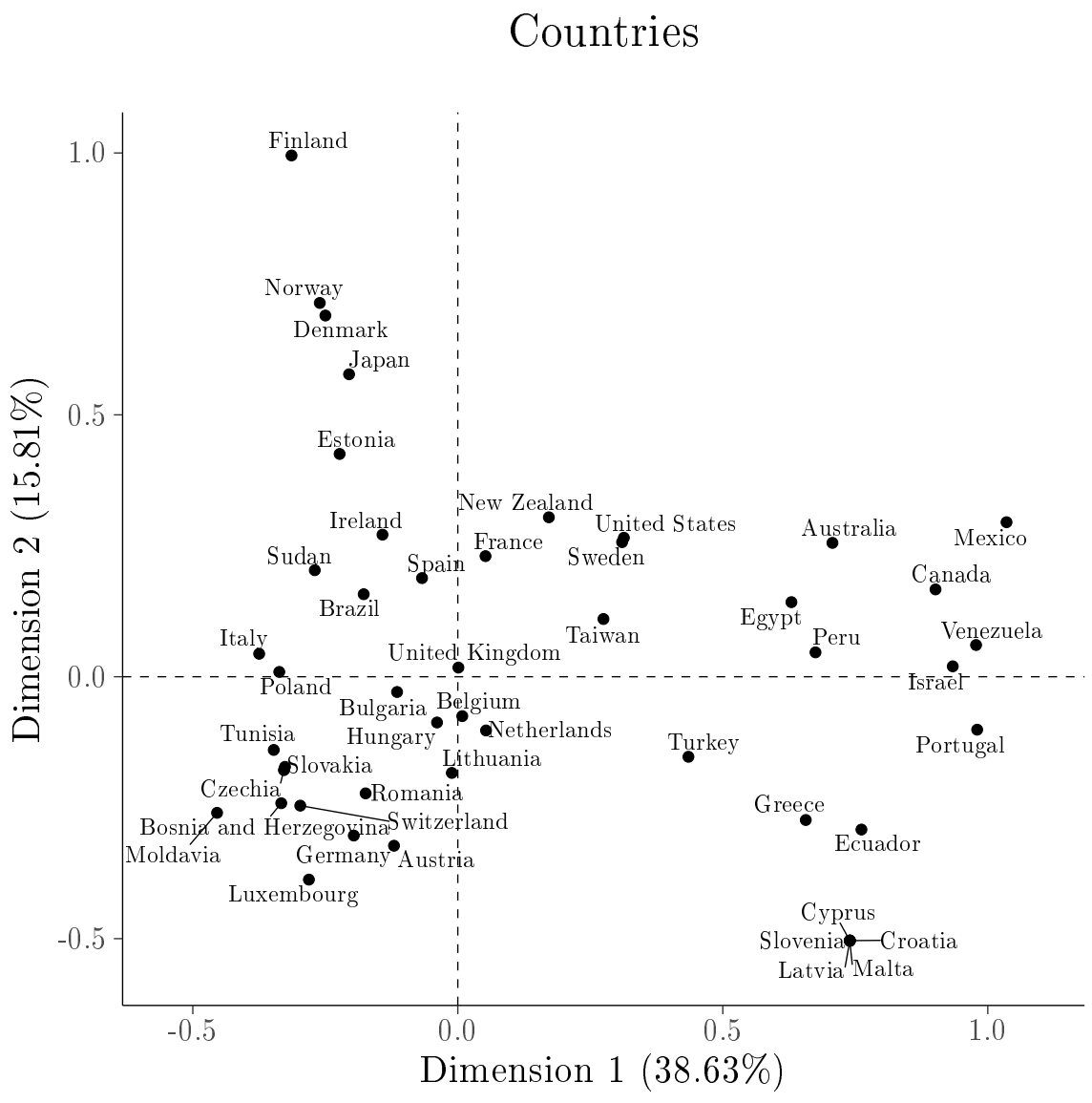


Figure 2.6: Plot for the countries on the 1st and 2nd dimensions.

Another interesting way to look at the data is to see from which countries the VAAs originated, which is what Figure 2.6 shows. Here, we see that the "Finnish" type of VAA comes from Finland, Norway, Denmark, Estonia and Japan. This suggests some geographical clustering (with the exception from Japan). For the other countries, a clear logic is lacking. Yet, it is interesting to see the Netherlands, were two of the groups of VAAs originate has a location close to the centre. This means VAAs from this country look much like the average VAA.

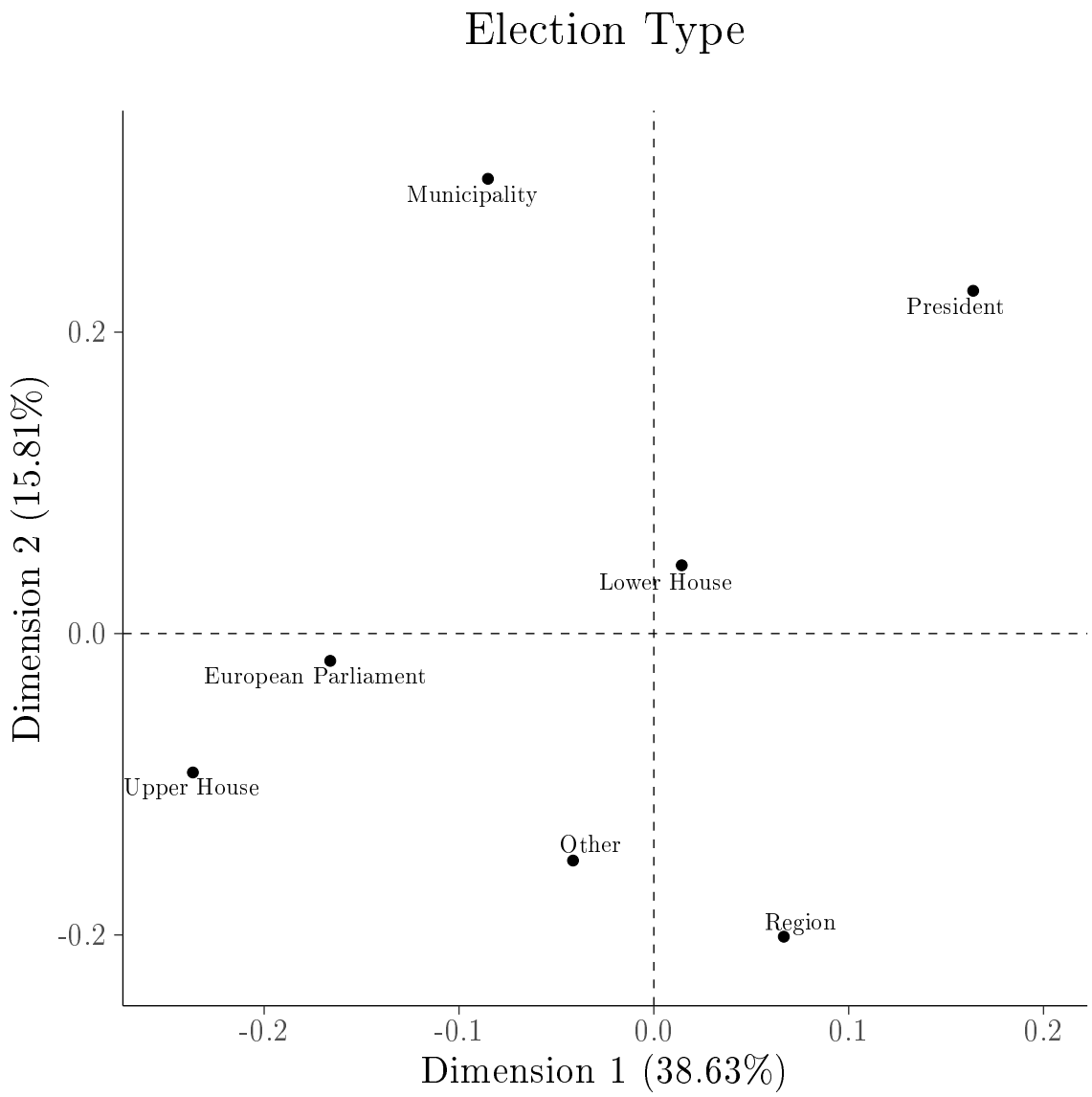


Figure 2.7: Plot for the elections on the 1st and 2nd dimensions.

Looking at the type of elections provides even more information. As Figure 2.7 shows, the "Finnish" type of VAA is most often made for municipalities, while the "Stemwijzer" type VAAs occurs more often for European Parliament and Upper House elections. This is because municipal and presidential elections focus more on the individual candidates than the other elections. This also explains why the "Finnish" type of VAA relates to a focus on Candidates, as Figure 2.5 shows.

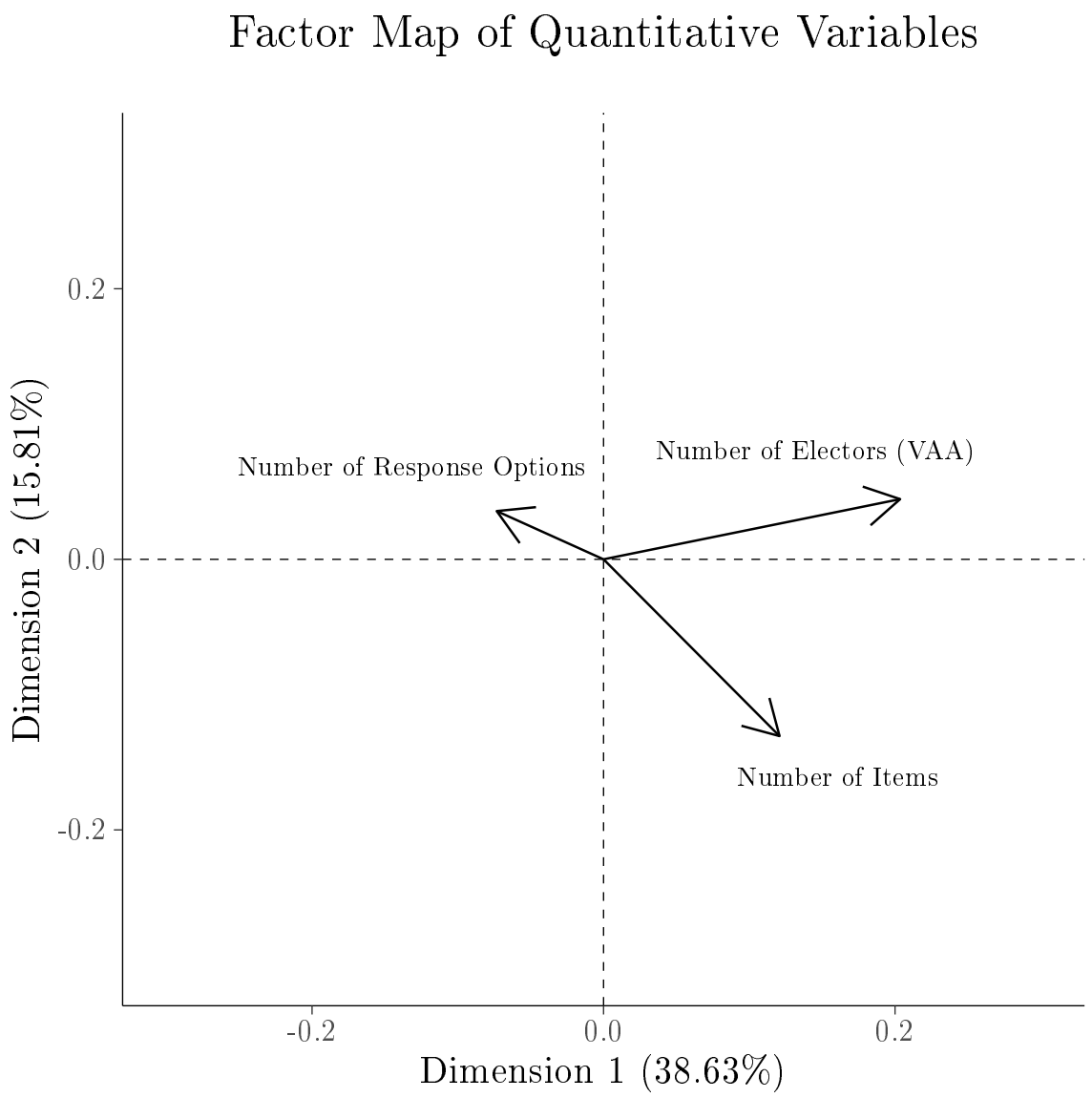


Figure 2.8: Association of the Quantitative variables with the first two dimensions.

Finally, we look at the behaviour of the quantitative variables (shown in Figure 2.8). Here, we see the number of response options is positively related to the second dimension and the negative part of the first dimension. This is most likely because the Finnish type of VAAs often has more than the standard 5 response options than the Kieskompas type of VAAs. The number of electors in the VAA is also related to the positive side of the second dimension and also to the positive side of the first dimension. This would mean that Kieskompas type of VAAs more often has a higher number of electors included. The number of items points towards those VAAs that were of a more complex nature based on Kieskompas logic.

So far, we have been able to identify three "types" of VAA by using MCA. The first type we labelled the "Finnish" type of VAA. This is a VAA for the municipal elections, designed by a media company that focuses on candidates. The second type of VAA bases itself on the "Stemwijzer" concept. NGOs develop these VAAs, which feature elite party positioning and simple quantitative rank visualisations. The third type are those VAAs based

on "Kieskompas". They feature two-dimensional maps and use either expert positioning or the Kieskompas-method.

2.3.2 A Second Look: Applying Cluster Analysis

To further pin down the various types of VAA, I conduct a type of cluster analysis. To do so, I use the principal components generated by the MCA (Husson, Josse, and Pagès 2010). This circumvents the problem of running a cluster analysis with categorical data. I run the cluster analysis with the `HCPC` (Hierarchical Clustering on Principle Components) function from in the `FactoMineR` package (Lê, Josse, and Husson 2008). HCPC differs from k-nearest neighbourhood clustering in that we do not need to specify the number of expected clusters in advance. This makes the method more suitable for exploratory research. Instead, the method uses Ward's criterion to find this number[7]. HCPC reaches Ward's criterion by minimising $\Delta_{ward}(A, B)$ in the equation:

$$\Delta_{ward}(A, B) = \frac{I_A I_B}{I_A + I_B} d^2(\mu_A, \mu_B) \tag{2.4}$$

with d representing the *Euclidean* distance, μ_A and μ_B the barycentres of cluster A and B, and I_A and I_B the cardinalities (number of elements) of each of the clusters. This method works in an ascending, bottom-up fashion. That is, it starts by defining each point as a separate cluster and works its way upward until it reached the cut-off point. We can either define this cut-off point ourselves or leave the choice to the algorithm. In that case, the cut-off is where the relative loss of inertia (the degree of variance in the data-set) is the highest. The algorithm calculates this as $\frac{i(n+1)}{i(n)}$, where i represents the inertia, and n the number of clusters. Here I use this second method. Also, I do not set any minimum or maximum to the number of clusters.

Running the HCPC algorithm results in 6 clusters, visualized in Figure 2.9. Looking at the distances between the individual VAAs and the centres of the clusters, I find the following VAAs to be most representative of them: Vaalikone (in red), Smartvote (in yellow), Stemwijzer (in green), Kieskompas (in Blue), EUVox (in purple), and euandi (in blue). There are a few points of interest here. First, three of the clusters line up with the clusters we identified earlier. Vaalikone is representative of the Finnish VAAs, Kieskompas of the Kieskompas VAAs and Stemwijzer of the Stemwijzer VAAs. Second, we find the algorithm splits up two of the clusters. It split the Stemwijzer group into a *Smartvote* and *Stemwijzer* cluster, and the Kieskompas group into a *Kieskompas* and *EUVox* cluster. Besides, these new clusters overlap for a considerable degree. This indicates a high amount of similarity between them. Third, the *euandi* VAA again forms its own cluster. Also, extra simulations show that only when we set the algorithm to find 3 clusters or less does this cluster. In that case, the algorithm adds it to the Kieskompas/EUVox cluster.

Figure 2.10 shows this in more detail. Starting from the left (which is opposite to what the algorithm does), we find that the first split is between the Vaalikone/Smartvote/Stemwijzer cluster and the EUVox/Kieskompas/euandi cluster. After this split, these clusters are stable for a long time until the euandi cluster splits off from the EUVox/Kieskompas cluster. Later,

[7] Other methods, like using the smallest, largest or average distance are also available.

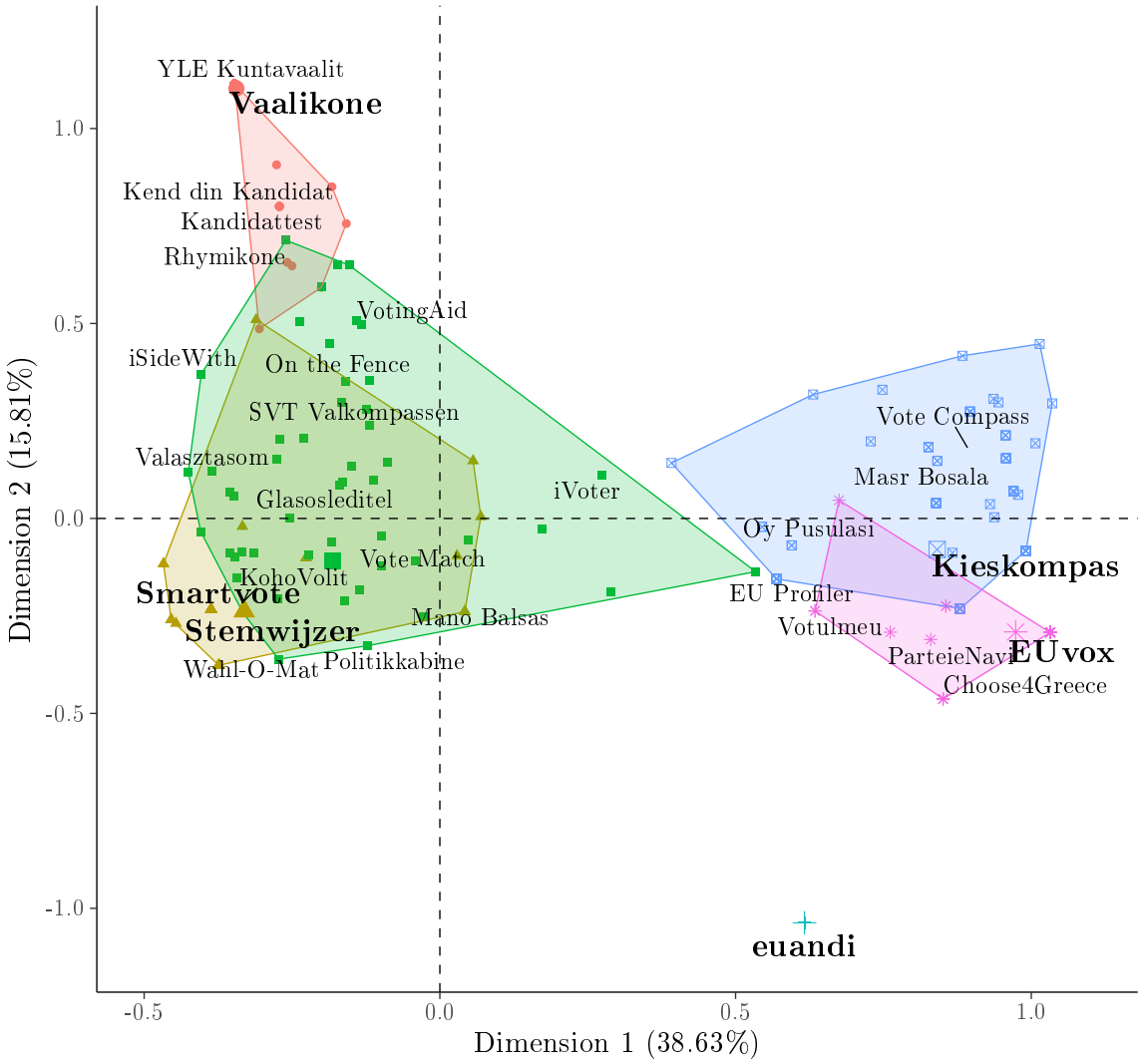


Figure 2.9: Clusters based on Hierarchical Clustering on Principal Components. Labels in bold are those cases that are most characteristic of their respective clusters. This means they are the closest case to the cluster of the respective centre.

the Vaalikone cluster splits off from the Smartvote/Stemwijzer cluster. Then, the Smartvote and Stemwijzer clusters split, and finally the EUVox and Kieskompas clusters separate. Going further, we see separations in the Stemwijzer and Kieskompas clusters. As these are below the cut-off point, they are not kept as separate clusters. Also, as the plot shows, further divisions are far from the ones that came before.

As before, to get a better idea of what defines these new clusters I take a look at which categories define them. Table 2.6 shows in the first column the percentage of individuals with a certain category in that cluster, the second the percentage of individuals in a cluster with a certain category, and the third the overall percentage of individuals belonging to that category. As such, 99% of the VAAs that launched in Finland are in cluster 1, 98% of the VAAs in cluster 1 are from Finland, and 37% of the VAAs in the data-set are from

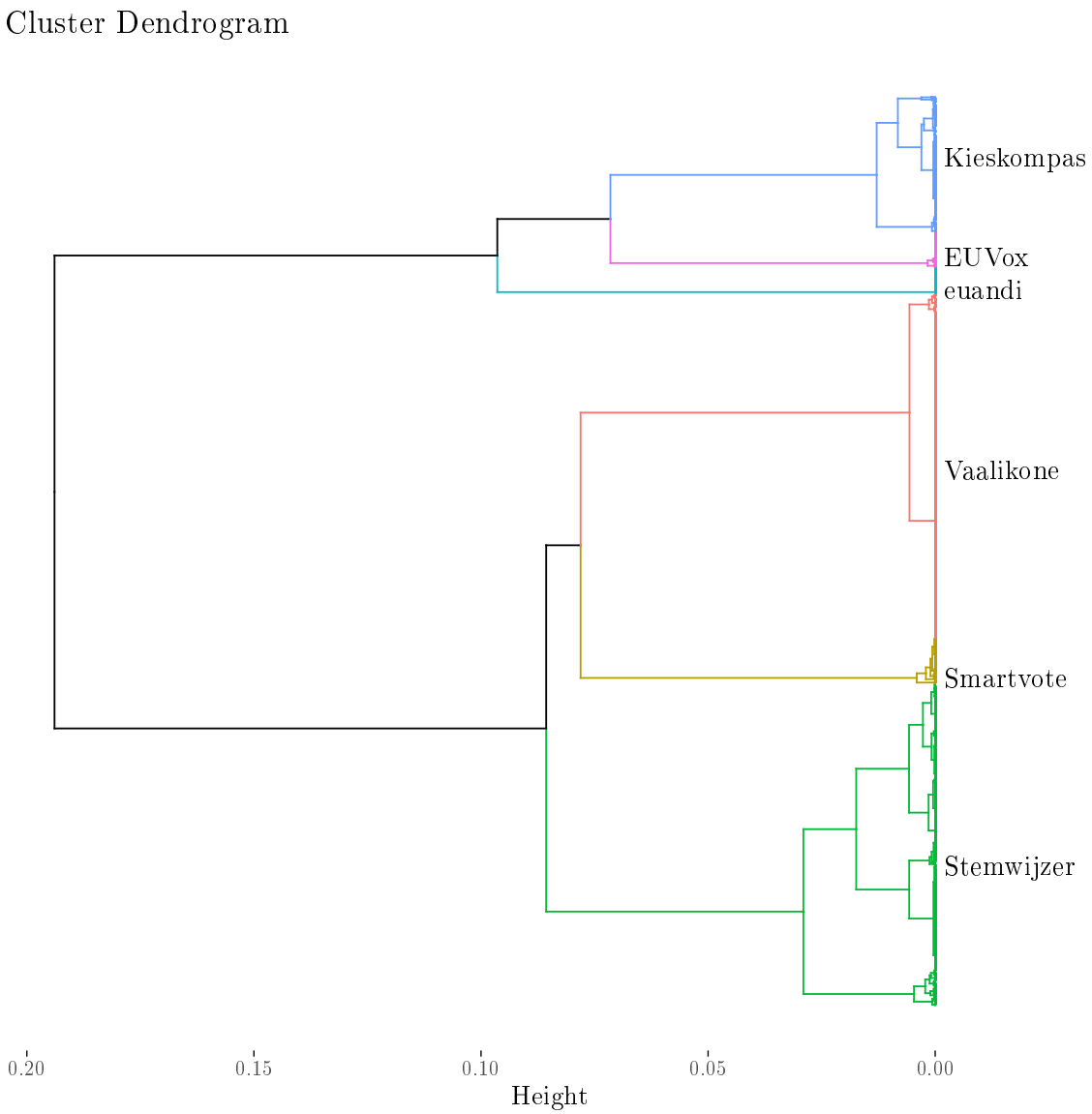


Figure 2.10: Dendrogram based on Hierarchical Clustering on Principal Components. The Height refers to the Euclidean distance between the centres of the clusters.

Finland. From this, we can conclude Finland is representative of this cluster. In the same vein, response options, design by a media company, not having a "Do Not Know" category and using candidate positioning are also representative of this cluster. Spider graphs, six-dimensional matching and design by an NGO define the second cluster. The VAA characteristic for this cluster is the Smartvote VAA from Switzerland. Elite positioning, scale responses and design by an NGO define the third cluster. Most representative here is the Vote Match VAA which is the international version of the Stemwijzer. The fourth cluster only contains the *euandi* VAA. This VAA defines itself with a three-dimensional visualization and Euclidean matching. A two-dimensional visualization and matching, using the Kieskompas method and design by a company define the fifth cluster. Not surprising, the Kieskompas VAA is the most representative of this cluster. The sixth cluster distinguishes itself with Delphi positioning and two-dimensional and Euclidean matching. The EUVox VAA is most representative of

Cluster	Category	Category/ Cluster	Cluster/ Category	Total
1	Finland	99	98	37
	Response Format – Options	98	98	37
	Designer - Media	93	100	40
	No DNK	87	99	42
	Positioning - Candidate	82	100	31
2	Matching - 6D	100	96	5
	Visual - Spider Graph	59	100	9
	Switzerland	98	74	4
	Smartvote	98	68	4
	Designer – NGO	11	88	41
3	Positioning - Elite	97	71	24
	Designer – NGO	71	88	41
	Designer – Media (No)	54	98	60
	Response Format – Scale	52	98	63
	Vote Match	100	35	11
4	Visual - 3D	100	100	3
	Matching - 3D	100	100	3
	euandi	100	100	3
	Matching - Euclidean	43	100	7
	Year - 2014	31	100	9
5	Matching - 2D	80	100	22
	Visual - 2D	80	100	22
	Designer – Company	75	98	23
	Positioning - Kieskompas	72	98	24
	Visual - Quant Rank (No)	99	67	12
6	Positioning - Delphi	100	98	4
	EUVox	100	77	3
	Matching - Euclidean	56	98	7
	Year - 2014	34	81	9
	Matching - 2D	18	100	22

Table 2.6: Description of the cluster by the variable categories. Category/Cluster refers to the percentage of individuals with a certain category in that cluster. Class/Cluster refers to the percentage of individuals in the cluster with a certain category. The total refers to the percentage of individuals with the category in the full data-set. The table shows only the first five categories for each cluster.

this cluster.

Table 2.7 shows an overview of these clusters based on the sections in Table 2.3. Each of these clusters represents a different type of VAA and different design choices that the designers made. This is important not only for designers but also for users, as these choices influence the information they take away from a VAA. For example, Stemwijzer type VAAs often use elite positioning which makes them dependent on the cooperation of the political

Cluster	Focus	Response	Position	Match	Visual	Design
Vaalikone	Cand.	Options	Candidate	Agree	Rank	Media
Smartvote	Cand.	5−point	Candidate	6D	Spider	NGO
Stemwijzer	Party	$2-3$ point	Elite	City	Rank	NGO
euandi	Party	5−point	Kieskompas	3D	3D	Uni
Kieskompas	Party	5−point	Kieskompas	2D	2D	Comp
EUVox	Party	5−point	Delphi	2D	2D	Uni

Table 2.7: Description of the 6 clusters with their main characteristics.

parties. Smartvote and Vaalikone VAAs give users an in-depth look in the stances of the individual candidates. Yet, they rarely offer a political map for users to position themselves on. Kieskompas and EUVox VAAs do show such political maps but differ in the way they position the parties. The euandi type VAA offers a three-dimensional matching and visualization. While this allows for an overview of the European political landscape, the result might be difficult to interpret.

Of all VAAs only a few offer an explicit description of these decisions on their website, let along a rationale for them. Why this happens so rarely is up for speculation. Possible explanations might be that the designer does not consider it necessary or useful. They might consider that users are not interested in the descriptions and will not read them. Or, the designer might not want to reveal the design choices to prevent copying. This can be the case when companies or media outlets design their VAA. Another explanation might be that the designer is unsure about their design choices or wants to hide uncertainties. Even so, it would seem reasonable for VAAs to include at least a minimal overview of its design choices. This would not only be helpful for the user, but for the designer as well.

2.4 Points of Research

If this brief analysis shows one thing it is that the vast variety of VAA design choices often comes down to three questions. First, how do VAAs position the users? Second, how do VAAs match the users with the parties or candidates? Third, how do VAAs visualize this match? As we can see from Tables 2.6 and 2.7, the methods of party positioning and the type of designer seem less relevant. If we aim to assess the quality of VAAs, it is thus on these three issues that we should focus. Starting with how to position the users, we find four of the six types of VAAs we identified using a five-point Likert scale in their questionnaire. More relevant is that the way in which the VAA positions the users is by the means of *scales*. Instead of calculating the total for all questions, these VAAs cluster together some of the questions into separate dimensions. They then use these dimensions to position the users on the political map or spider graph. As the political parties are also positioned on these scales, a lot depends on how the designers of the VAA construct these scales. So, it is necessary we are able to access the *quality* of the scales. Without high-quality scales, the positioning of both the user and the party is in jeopardy. The resulting visualisation will then provide little information and might even increase the confusion of the voter. Besides, knowing the quality of the scales is one thing, but it is more helpful if we also understand why the scales

have the quality they have. This way, we can identify where the problems with the scale lie and address these when possible. This leads to my first set of research questions:

RQ 1a What is the quality of the scales Voting Advice Applications use?
RQ 1b How can we explain the quality of these scales?

Both questions centre around the way in which the designers use the questions of the questionnaire to construct the scales. But this assumes that the questionnaire itself is unproblematic. Thus, any problem caused by the scales is only dependent on the construction of the scales. Even so, as Holleman, Kamoen, Krouwel, et al. (2016) point out, there are reasons to believe that the questionnaire itself can be problematic as well. Most relevant here is which questions the questionnaires have and which wording they use. Both the selection of questions (Walgrave, Nuytemans, and Pepermans 2009; Lefevere and Walgrave 2014) and the wording of the questions can influence the response of the user (Holleman, Kamoen, Krouwel, et al. 2016). Yet, if this influences the match between the user and the party is unknown. This leads to my second research question:

RQ 2 How does the wording of the questions influence the match between the user and the party?

After the questions and the scales, I turn to the visualizations. Until now, little the VAA community has largely ignored this aspect of VAA design. Yet, if users would interpret the visualisation in an alternative way then designers imagine they would, the point of constructing scales and questionnaires of quality loses some of its relevance. A valid and reliable VAA that users interpret in a different way than they should, is still problematic. So, my third research question is:

RQ 3 How does the visualization influence the way in which the user perceives their match with the parties?

These three questions assist me to assess the quality of the positioning of the users and the way in which they visualize this match. This allows me to say something about the quality of four of these types of VAAs. More important, it allows me to construct a toolset with which designers can test their VAAs.

2.5 Theoretical Framework

To provide a useful answer to these questions, it is beneficial I use a single framework in which I assess them. This will allow me to focus on specific aspects of the questions and allows me to compare the results between them with ease. The first problem here is that Voting Advice Applications did not form a research field of their own until the end of the 2000s. Until then the debates, like those between Groot (2003c, 2004, 2003b,a) and Schuszler, Graaf, and Lucardie (2003a,b) about the quality of the 2002 edition of the Dutch Stemwijzer, took place on the national level and had not yet found their way to scientific journals. A few years later, the first collective works began to appear (Cedroni and Garzia 2010; Garzia and Marschall 2014). At the same time, scholars began discovering the wealth of data generated by VAAs. Thus,

they used VAA data to study public opinion, cleavage theory, or new political dimensions (e.g. Mendez, Gemenis, and Djouvas 2014; Wheatley 2016; Garry, Matthews, and Wheatley 2017). Yet, most prominent was the branch of literature that focused on the methodological foundations of VAAs. These works questioned the questionnaires (Walgrave, Nuytemans, and Pepermans 2009; Camp, Lefevere, and Walgrave 2014), response options (Rosema and Louwerse 2016), party positioning methods (Gemenis 2013b) and matching algorithms (Louwerse and Rosema 2014; Mendez and Wheatley 2014; Mendez 2012). Still, while informative, most of these works and evaluations have been incidental. A common framework for the evaluation of VAAs is thus lacking. This makes it difficult for VAA designers to understand the influences of their design choices. Moreover, in the end, there is no agreed-upon way to test a VAA and assess whether it fulfilled its intentions. To overcome this problem, it would be beneficial if there was a way to assess the quality of VAAs. Such an assessment would call for a framework that designers could use to test both the individual steps of the VAA and the VAA as a whole. While there have been earlier attempts, these were either limited to only a limited number of aspects of the VAA (Škop 2010) or used broad criteria (Garzia, Trechsel, Vassil, et al. 2014).

Here, I will suggest seeing the design of VAAs as a process that consists of a series of steps. This focus on the process allows me to test not only the individual steps but also the process as a whole. Moreover, it allows me to apply certain standards to the process with which I can measure its quality. Figure 2.11 shows this process.

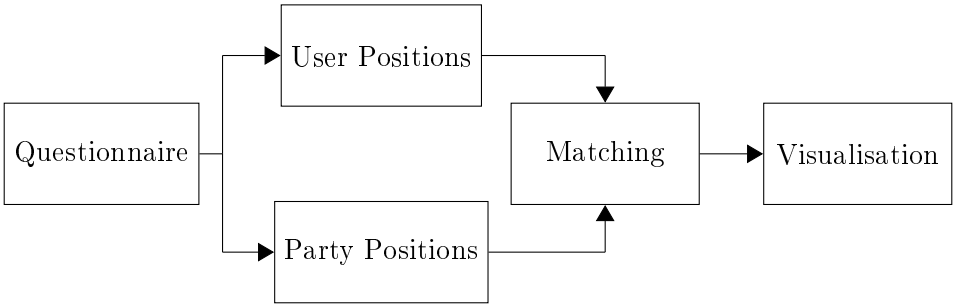


Figure 2.11: Inside the model section of the VAA design process

The first step in this process is the questionnaire. This contains the choice of topics, the wording of the questions, and the choice of response categories. For this step, Groot (2003c,b,a, 2004) was the first to point out problems with the selection of issues. Focusing on Stemwijzer, he identified issues that did not discriminate between different parties, important issues that were unincluded, and issue areas that were over-represented at the expense of others. In a reply, Schuszler, Graaf, and Lucardie (2003a,b), who were responsible for the design of Stemwijzer, pointed out that the cause of this was the self-imposed limit of 30 statements. They argued there always will be topics that are not included in the VAA. Besides, Stemwijzer consults both experts and parties to construct the questionnaire. According to the authors, this is enough to ensure a proper selection of issues. Yet, for a similar VAA in Belgium, Walgrave, Nuytemans, and Pepermans (2009) showed that the actual selection of issues matters. Not only does it influence the choices the user has to make, but it also influences the match they receive. Drawing on a data-set of 50 statements, Walgrave, Nuytemans, and Pepermans (2009) show that different configurations of the 36 statements that make up

their VAA have an impact on the match the user receives. Also, Ramonaitė (2010) points out that if a majority of the voters agree upon a certain issue while there is only a single party in favour, any VAA algorithm will favour that party. Later, Lefevere and Walgrave (2014) showed that including more left-right statements makes more left-wing voters receive advice for left-wing parties (and vice versa for right-wing voters). Also, doing this would favour parties with extreme positions on this dimension. Currently, most VAA builders select their issues by narrowing down a longer list of issues. This can happen either amongst themselves or in cooperation with the parties. The list of issues themselves comes from either by the designers, experts, or text analysis techniques like LDA (cf. Krouwel, Vitiello, and Wall 2012).

The VAA gets the user positions from the responses to the users to the questionnaire. How the VAA does this influences the responses (Krosnick and Presser 2010). Starting with the formulation of the questions, both Camp, Lefevere, and Walgrave (2014) and Gemenis (2013b) note a high amount of double-barrelledness and quantification. This not only complicates the statements but also makes it harder for users to understand what is being asked of them. Also, as a significant number of VAAs use Likert scales for their response options, negatively framed questions might lead users to provide incorrect answers. This is because they have to *disagree* with the statement if they hold a positive attitude towards the object of the statement. Indeed, Holleman, Kamoen, Krouwel, et al. (2016) find that framing the question in a positive or negative way influences the way users respond to the question. Besides, the layout of the questionnaire itself seems to matter as well. Holleman, Kamoen, and Vreese (2013) find that providing different headers above the statements (indicating to which area a certain statement belongs) influences users' responses while Rosema and Louwerse (2016) find that the number of response options offered to the users influences the final match the users receive. Moreover, Baka, Figgou, and Triga (2012) point out that even the response options themselves can be problematic. In a small-scale study, they show that users chose the middle — neutral — response not only when they had a true neutral opinion, but also when they lacked the knowledge to answer the question or disagreed with the assumptions of the statement itself. Moreover, as some users had similar reasons to choose a "I do not know" response, the true opinion of the user might be difficult to establish.

Like the positions of the users, VAAs establish the positions of the political parties using the same questionnaire. As we saw earlier, VAAs can use four ways to do so. The first is party self-placement, where parties position themselves on the statements. The second is expert positioning, where experts position the parties. The third is the *Kieskompas* method, which combines party self-placement and placement by experts. Fourth is the Delphi technique, where experts and non-experts iterate through several rounds they reach a certain level of agreement. Each of these methods has their own benefits and problems (Gemenis and Ham 2014) and leads to different positions. Which method the designer uses thus depends on which method they believe works best for the VAA.

Once the VAA knows the positions of the users and the parties it calculates the "match" between both. Mendez, Gemenis, and Djouvas (2014) shows that this is not a neutral matter. Choosing whether to match users using a city-block metric, Euclidean distance, directional or "hybrid" model alters the match the user receives. Most important is that there seems to be no single best way of matching that VAA designers have agreed upon. Also, Louwerse and Rosema (2014) point out that it differs whether VAA designers conceive of the match

as taking place in a high-dimensional or low-dimensional space. Around 90% of the users in their study would receive a different match if the VAA would use another matching algorithm. VAAs use high dimensional spaces if they show the user a list of agreement for each of the parties, they use low dimensional spaces if they reduce the total number of questions to a lower number of underlying dimensions and visualize them on a political map. Often this reduction of questions takes place before the launch of the VAA. Yet, often the ideas of the designers of what the underlying dimensions are, are not shared by the respondents. Thus, Germann, Mendez, et al. (2015) and Germann and Mendez (2016) suggest performing a Mokken Scaling Analysis to validate and adjust the scales using early user data. As such, the final version of the scales can have different questions than the original ones. Another approach, developed by Linden and Dufresne (2017) keeps all the questions included in the original scales but uses the early user data to provide weights to each question based on their contribution to a given dimension. Also, they propose two measures to estimate the validity of VAA scales. The first is the candidate-to-partisan proximity. This is the total distance between candidates and those users who indicated they would vote for the candidate. The second — candidate-to-candidate proximity — looks at the distances between the different candidates on the underlying dimension (Linden and Dufresne 2017, pp.12-14).

Partly depending on which matching method the VAA uses, the VAA visualizes the results through percentage lists, two-dimensional maps or spider graphs. While there are many different visualizations available, VAA designers have spent little time researching how the visualisation influences the user. Indeed, the assumption seems to be that the visualisation itself is "neutral". Yet, given that political spaces (used in the two-dimensional maps) rarely mean the same to different users (Benoit and Laver 2006), and the number of parties included in the visualisation likely influences how users interpret the positions of the parties (Dimara, Bezerianos, and Dragicevic 2017), this assumption seems questionable at the least. As such, visualisation is an (as yet) unexplored area of VAA design.

2.6 Set-Up of the Thesis

Corresponding with the framework set-up above, the three questions will cover each of the steps in the VAA design process. Deviating from the process, I will first discuss the user positions, then the questionnaire and finally the visualisations.

Chapter 3 will describe how I will measure the "quality" of the questionnaire, scales and visualisations. I will do so using the concepts of validity and reliability. These concepts are not only well-known in most social sciences, but are also flexible and I can measure them using a wide range of techniques. To measure the validity of the scales, I will draw on the advice of Blasius and Thiessen (2012) and use methods of visualization and data reduction. I will use Principal Component Analysis (PCA) and its categorical counter-part categorical Principal Component Analysis (catPCA), as well as Multiple Correspondence Analysis (MCA). Each of these methods allows me to visualize the hidden structures of the responses users gave. To measure reliability, I will combine approaches from both classical test theory and the more recent item response theory. I will use Cronbach's α in both its original and ordinal variant, McDonald's ω, the Latent Class Reliability Coefficient (LCRC) and Mokken Scaling Analysis. I will discuss each of these methods and their derivations. Also, I will look at which of their

characteristics we need to take into account when using them on VAA data.

Chapters 4 and **5** will focus on the scales. To begin with, I will discuss the way in which users perceive questionnaires. I do so by using a model of the response process and give certain reasons why they respond to questions the way they do. Then, I will introduce the EUVox data-set that I will use in both chapters. From EUVox, I derive three sets of scales that I will construct using different methods. Next, I will rank each of these scales on quality, unidimensionality and reliability. In **Chapter** 4 I will then focus on some of the countries and scales that performed either very well or very disappointing. I will use a combination of PCA, catPCA and MCA to uncover the underlying reasons for this.

Chapter 6 will focus on how the questionnaire can influence the match between the user and the party. I will do so by reporting on an experiment that I carried out during the 2017 elections for the House of Representatives in the Netherlands. During these elections, I launched a VAA in which several questions were either worded positive or negative. Using the method of small multiples, I will show that such small changes can influence the match between the user and the parties. Yet, as we will see, the size of this match depends not only on the party but also on the question.

Chapters 7 and **8** will discuss the various visualisations that VAA designers can use in their VAAs. In **Chapter 7** I will discuss the current theory on visualisation and how users perceive graphs. Together with this, I will provide some criteria of what makes a good visualization. Then, I will provide an overview of the visualizations that VAAs have used thus far and judge these against the criteria. In **Chapter 8**, I will discuss the results of an online experiment that focused on VAA visualisations. Here, I find users use different ways to look at the political map. Yet, more important is that they seldom seem to agree on which party is closest to them.

Finally, the concluding chapter of this thesis will provide a summary of the main findings and answer the research questions.

3 | Concepts & Methods

In this chapter, I will discuss how I will assess the quality of the scales, questionnaire and visualizations. I will define quality as consisting of validity and reliability. Thus, a VAA of high quality has high validity and reliability. Validity is the degree with which we measure what we think we are measuring (King, Keohane, and Verba 1994, p.25). While there are many types of validity, external, internal, and test validity are the most important (Carmines and Zeller 1979; Shadish, Cook, and Campbell 2002). External validity measures the extent to which any findings also apply outside of their original setting. In our case, this could mean whether a finding of one VAA also applies to other VAAs. Or if a finding in one country also applies to other countries. This kind of validity means that a finding is not limited to the original choice of persons, settings, treatments, or outcomes (Shadish, Cook, and Campbell 2002). Internal validity refers to causality. Thus, high internal validity means the relationship between the variables that interest us is causal. So, we need to prove the A precedes B, A co-varies with B, and that no other explanation for the relationship is possible (Shadish, Cook, and Campbell 2002, p.53). As one can never prove causality, this makes internal validity the hardest of all validity types. Test validity measures if a test measures what it should measure. This is the most well-known version of validity. We can break it down into three groups: criterion, content and construct (Cronbach and Meehl 1955). Criterion validity is the extent to which a measure co-varies with another measure which aims to measure the same (Carmines and Zeller 1979). For example, party positions measured with manifestos and with the help of an expert survey should co-vary. Content validity is the extent to which a measure measures all it should measure and nothing it should not measure. So, if we would measure positions on a left-right scale, our measures should only include measures that measure left-right positions, and no measures that measure other positions. Construct validity measures if the measure covers all aspects of the measure. Constructs are important in the social sciences, as we cannot measure such things as a political preference in a direct way. To do so, we have to make a construct based on surveys and interviews. The validity of this construct then influences to a large degree how valid our findings are. *Reliability* means that if we would run our study in the same way again, we would get the same result (King, Keohane, and Verba 1994). This can apply to persons, settings, time periods and other changes. In other words, if two identical persons would fill out a survey under identical conditions, the results should be identical. Another way to look at reliability is by relating it to random error. Reliability then describes the degree of random error.

So, how can we measure reliability and validity? For validity, I use the approach by Blasius and Thiessen (2001b,a, 2012) and Thiessen and Blasius (2008). They suggest using data

reduction techniques to visualize the data. This can reveal problems that would otherwise go unnoticed. For example, plotting the items in a scale allows me to see if certain items in scales co-vary or if a scale measures more than a single construct. Of the data-reduction techniques available, I will use Multiple Correspondence Analysis, Principal Component Analysis and categorical Principal Component Analysis. For reliability, I will use methods from Classical Test Theory and Item Response Theory. From the former, I will use Cronbach's α, the ordinal version of Cronbach's α and the ordinal version of the coefficient ω. From the latter, I will use the latent class reliability coefficient (LCRC) and Mokken Scaling Analysis (MSA). I will now discuss each of these methods, their derivations, and certain aspects that we should take into account when applying them on VAA data.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) linearly transforms a set of (numeric) variables into a smaller set of uncorrelated variables (Dunteman 1989, p.7). As such, it is a useful technique if we want to reduce the number of variables we have to a smaller number of underlying dimensions. Given s variables with n respondents, PCA allows us to reproduce these variables without any loss of information from s latent factors. We do so by expressing each variable v_s through the association α_{jk} with each latent factor f_k[1]:

$$\begin{aligned} v_1 =& f_1\alpha_{11} + f_2\alpha_{12} + \ldots + f_s\alpha_{1s} \\ v_2 =& f_1\alpha_{21} + f_2\alpha_{22} + \ldots + f_s\alpha_{2s} \\ & \vdots \\ v_s =& f_1\alpha_{j1} + f_2\alpha_{j2} + \ldots + f_s\alpha_{js} \end{aligned}$$

Here, v_p are the standardized variables, α_{jk} represents the correlation between jth observed and kth latent factor (known as the factor loadings), and f_k holds the values for all nth respondents on the kth latent factor. While we can retrieve all variables this way, the aim of PCA is to reduce the number of factors while retaining a sizable portion of the overall variance (later on, while discussing Correspondence Analysis, we will also call this variance the *inertia*). We perform PCA not on the original matrix of variables but on either the correlation, covariance or SSCP (sums of squares and cross products) matrix of the variables. Whether we choose a correlation or covariance (or SSCP) matrix depends on the degree to which we want the different variances among the different variables to influence the results. The covariance (or SSCP) matrix is preferable when we measure all variables in comparable units, and the difference in variance makes an important contribution to our interpretation. The correlation matrix standardizes all variables and thereby ignores any difference in the variances. In effect, both correlation and covariance matrices are matrices in which both rows and columns show the same variables and where the entries show either the correlation or the covariance between two variables. For the purposes of the next explanation, we will use

[1] In the literature, these factors are otherwise known as discriminant functions, canonical functions or variates, and principal components

the correlation matrix.

To find the factors f_k, we need to calculate the associations α_{jk}. We can find these by calculating the *eigenvectors* and *eigenvalues* of the correlation matrix, which we find by diagonalization of this matrix (Husson, Lê, and Pagès 2011, p.10). The eigenvalue of a factor represents the amount of correlation (association) captured by that principal component, while the eigenvector represents the contribution of each of the original variables to the principal components. To understand how this works, we turn to matrix algebra, assuming a correlation matrix. As such, we can rewrite the formulas above as:

$$\mathbf{Z} = \mathbf{FA}, \text{ with} \frac{1}{n}\mathbf{F}^\intercal\mathbf{F} = \mathbf{I} \tag{3.1}$$

Here, $\mathbf{Z}$ are the original variables, $\mathbf{F}$ are the factors, and $\mathbf{A}$ is the correlation between $\mathbf{Z}$ and $\mathbf{F}$. $\mathbf{F}^\intercal$ is the *transpose*[2] of the matrix $\mathbf{F}$ and $\mathbf{I}$ is the *identity matrix*. As the identity matrix $\mathbf{I}$ is a diagonal matrix with $1s$ on the diagonal and $0s$ in the off-diagonal, this means that the factors only correlate with themselves and not with each other. We call this the *orthogonality* of the factors.

With $\mathbf{Z} = \mathbf{FA}$ in mind, we now turn to the correlation matrix we construct using the variables in $\mathbf{Z}$. What we do when correlating each variable against all other vectors is the same as multiplying the vector of variables $\mathbf{Z}$ with its own transpose $\mathbf{Z}^\intercal$, taking the number of cases n into account:

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z} \tag{3.4}$$

Using the property that $(\mathbf{FA})^\intercal = \mathbf{A}^\intercal\mathbf{F}^\intercal$ (the transpose of the multiplication of $\mathbf{F}$ and $\mathbf{A}$ is equal to the transpose of $\mathbf{A}$ multiplied by the transpose of $\mathbf{F}$)[3] we can rewrite this as:

[2] In a transpose matrix, the (i, j) entry is the (j, i) entry of the original matrix:

$$\begin{bmatrix} A & B \\ C & D \\ E & F \end{bmatrix}^\intercal = \begin{bmatrix} A & C & E \\ B & D & F \end{bmatrix} \tag{3.2}$$

when we then multiply these matrices, we get:

$$\begin{bmatrix} A & B \\ C & D \\ E & F \end{bmatrix} \times \begin{bmatrix} A & C & E \\ B & D & F \end{bmatrix} = \begin{bmatrix} (A \cdot A) + (B \cdot B) & (A \cdot C) + (B \cdot D) & (A \cdot E) + (B \cdot F) \\ (C \cdot A) + (D \cdot B) & (C \cdot C) + (D \cdot D) & (C \cdot E) + (D \cdot F) \\ (E \cdot A) + (F \cdot B) & (E \cdot C) + (F \cdot D) & (E \cdot E) + (F \cdot F) \end{bmatrix} \tag{3.3}$$

Notice that on the diagonal the values are only multiplied with themselves and off-diagonal only with other values. As will we will see later on, in PCA we assume *orthogonality*, which means that the diagonal values such as $(A \cdot A) + (B \cdot B))$ equal 1 and all values off-diagonal such as $(C \cdot A) + (D \cdot B)$ equal 0. The resulting matrix is also known as $\mathbf{I}$, or the *identity matrix*.

[3] We can prove this as follows. Take $\mathbf{M} = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ and $\mathbf{N} = \left[\begin{smallmatrix} E & F \\ G & H \end{smallmatrix}\right]$. Moreover, the transpose of both matrices is $\mathbf{M}^\intercal = \left[\begin{smallmatrix} A & C \\ B & D \end{smallmatrix}\right]$ and $\mathbf{N}^\intercal = \left[\begin{smallmatrix} E & G \\ F & H \end{smallmatrix}\right]$. The product of $\mathbf{MN}$ is then:

$$(\mathbf{M} \times \mathbf{N}) = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} C & D \\ E & F \end{bmatrix} = \begin{bmatrix} (AE + BG) & (AF + BH) \\ (CE + DG) & (CF + DH) \end{bmatrix} \tag{3.5}$$

The transpose of this resulting matrix $\mathbf{MN}$ is:

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z} = \frac{1}{n}\mathbf{A}^\intercal\mathbf{F}^\intercal\mathbf{F}\mathbf{A} = \mathbf{A}^\intercal\mathbf{A} \tag{3.8}$$

In other words, we can reproduce the entire correlation matrix $\mathbf{R}$ from the multiplication of the factor loadings $\mathbf{A}$ with the transposed factor loadings $\mathbf{A}^\intercal$ - this is also known as the *fundamental theorem* in factor analysis (Yong and Pearce 2013, p. 82).

To get the matrix of factor loadings $\mathbf{A}$, we first have to calculate the eigenvalues and eigenvectors of the matrix $\mathbf{R}$. We do so by applying eigendecomposition[4] to the original correlation matrix $\mathbf{R}$[5] (Abdi and Williams 2010). This results in:

$$(\mathbf{M} \times \mathbf{N})^\intercal = \begin{bmatrix}(AE + BG) & (CE + DG)\\ (AF + BH) & (CF + DH)\end{bmatrix} \tag{3.6}$$

If we would take the transpose of $\mathbf{N}$ and $\mathbf{M}$ separately and then multiply them, we would get:

$$\mathbf{N}^\intercal \times \mathbf{M}^\intercal = \begin{bmatrix}E & G\\ F & H\end{bmatrix} \times \begin{bmatrix}A & C\\ B & D\end{bmatrix} = \begin{bmatrix}(AE + BG) & (CE + DG)\\ (AF + BH) & (CF + DH)\end{bmatrix} \tag{3.7}$$

In other words, $(\mathbf{MN})^\intercal = \mathbf{N}^\intercal \times \mathbf{M}^\intercal$.

[4] This technique is also known as spectral decomposition, eigenvalue decomposition or canonical decomposition.

[5] The procedure for eigendecomposition is as follows. First, $\mathbf{v}$ is an eigenvector of the correlation matrix $\mathbf{R}$ and $\boldsymbol{\lambda}$ the eigenvalue that belongs to it if:

$$\mathbf{R}\mathbf{v} = \boldsymbol{\lambda}\mathbf{v} \tag{3.9}$$

In other words, as soon as we find a vector which, when multiplied with the correlation matrix $\mathbf{R}$ is equal to that same vector multiplied by the eigenvalue $\boldsymbol{\lambda}$, then $\mathbf{v}$ is an eigenvector of $\mathbf{R}$. In this case, the eigenvalue is nothing more than a scalar for $\mathbf{v}$. Note that we can rewrite the above equation as:

$$\begin{aligned}\mathbf{R}\mathbf{v} - \boldsymbol{\lambda}\mathbf{v} &= 0\\ \mathbf{R}\mathbf{v} - \boldsymbol{\lambda}\mathbf{I}\mathbf{v} &= 0\\ (\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} &= 0\end{aligned} \tag{3.10}$$

We can do so as adding the identity matrix $\mathbf{I}$ does not add any new information. In effect, as multiplying the matrix $(\mathbf{R}-\boldsymbol{\lambda}\mathbf{I})$ with the eigenvector $\mathbf{v}$ equals 0, we can find an eigenvector by putting a known eigenvalue into the equation. Inversely, we can say that the values of $\boldsymbol{\lambda}$ where the equation holds are the eigenvalues of $\mathbf{R}$. As we are looking for a non-zero vector $\mathbf{v}$, in order for $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} = 0$ to be true, $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})$ must equal 0. This means that $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} = 0$ needs to be *non-invertible* or *singular*. If $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} = 0$ would not be *singular*, this would mean we could multiply it with its *inverse*:

$$\begin{aligned}(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})^{-1}(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} &= (\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})^{-1}0\\ \mathbf{v} &= 0\end{aligned} \tag{3.11}$$

Yet, as we want $\mathbf{v}$ to be non-zero, we do not want the equation to hold, hence we want $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} = 0$ to be non-singular. For $(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I})\mathbf{v} = 0$ to be non-singular, its *determinant* has to be 0. Hence, we can rewrite the equation as:

$$\det(\mathbf{R} - \boldsymbol{\lambda}\mathbf{I}) = 0 \tag{3.12}$$

We call this equation the *characteristic equation* or *characteristic polynomial* of $\mathbf{R}$, and is an n^{th} order polynomial, with n being the number of variables we construct the matrix $\mathbf{R}$ from. Note that in this equation, the *eigenvector* $\mathbf{v}$ has disappeared, leaving us with only a single unknown - the *eigenvalue* $\boldsymbol{\lambda}$, which we can find by solving the equation. We can then plug in the eigenvalues to find the respective eigenvectors.

We can place the resulting eigenvectors $\mathbf{v}$ together to form a matrix $\mathbf{B}$, while the found eigenvalues $\boldsymbol{\lambda}$ multiplied with the diagonal matrix $\mathbf{I}$ form the diagonal matrix $\boldsymbol{\Lambda}$. Using this to rewrite equation (12), we get:

$$\mathbf{R}\mathbf{B} = \mathbf{B}\boldsymbol{\Lambda} \tag{3.13}$$

Using the theorem that $\mathbf{B}\mathbf{B}^\intercal = \mathbf{I}$ and multiplying with $\mathbf{I}$ is similar to multiplying with 1, we can rewrite

$$\mathbf{R} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^\intercal \tag{3.15}$$

Here, $\mathbf{B}$ is the matrix of eigenvectors, $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues (with the values in decreasing order), and $\mathbf{B}^\intercal$ the transpose of the matrix of eigenvectors. Using the rules that a diagonal matrix ($\mathbf{\Lambda}$) is equal to its own transpose ($\mathbf{\Lambda}^\intercal$) and the associativity of multiplication, we can rewrite this to:

$$\mathbf{R} = \mathbf{B}\mathbf{\Lambda}^{1/2}(\mathbf{B}\mathbf{\Lambda}^{1/2})^\intercal \tag{3.16}$$

In turn, given that $\mathbf{R} = \mathbf{A}\mathbf{A}^\intercal$, we can rewrite the equation to:

$$\mathbf{A} = (\mathbf{B}\mathbf{\Lambda}^{1/2})^\intercal \tag{3.17}$$

which allows us to calculate the matrix $\mathbf{A}$ containing the factor loadings.

3.1.1 Properties of the Eigenvalue

These loadings, being correlations, range between -1 and 1 and are the correlations between the variable and the factor. As with regular correlations, the square of these loadings tells us the amount of variation of the variable that the factors capture. So, for each latent factor k, adding the squared association between the j^{th} observed and the k^{th} latent variable in question, over each of the s variables, the result is the eigenvalue corresponding to that factor:

$$\sum_{j=1}^{s} \alpha_{jk}^2 = \lambda_k \tag{3.18}$$

Moreover, through this link with the correlations α_{jk}, the eigenvalues λ_k represent the variance of the original variables. Thus, their sum is equal to that of the original variables:

$$\sum_{j=1}^{s} \lambda_k = \sum_{j=1}^{s} \sigma_j^2 \tag{3.19}$$

As we are working with a correlation matrix $\mathbf{R}$, this means we standardize the variables. As such, the total variance is equal to the number of variables:

$$\sum_{j=1}^{s} \lambda_k = s \tag{3.20}$$

and, given the previous equation also to the sum of eigenvalues $\sum_{j=1}^{s} \lambda_k$. Using this logic, adding the eigenvalues λ_i for the number of principal components k and dividing them by

this equation to:

$$\mathbf{R} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^\intercal \tag{3.14}$$

Thus, we have rewritten the correlation matrix $\mathbf{R}$ to a product of its matrix of eigenvectors, eigenvalues, and the transpose of the matrix of eigenvectors. For a more detailed explanation of this process and see Abdi (2007b) and Gilbert and Gilbert (1995).

the original number of variables s provides the explained variances (Dunteman 1989, p.17):

$$\sum_{i=1}^{k} \lambda_i / s \tag{3.21}$$

This number will be 1 if we use all components and will be < 1 if we use fewer components. In the same way, we can use the *communality*,

$$h_j^2 = \sum_{k=1}^{s^*} \alpha_{jk}^2 \tag{3.22}$$

to describe the proportion of the variance in variable j explained by the s^* factors we use in our analysis (Blasius and Thiessen 2012, p.37). Finally, the sum of squares of each of the associations over all variables j for any factor k equals 1:

$$\sum_{k=1}^{s} \alpha_{jk}^2 = 1 \tag{3.23}$$

We impose this constraint as otherwise, we could make the variance large by selecting large weights (Dunteman 1989, p.15).

3.1.2 Factors and Rotations

Remember that we started out by formulating the variables as the product of the factors and the association between the factor and the variables ($\mathbf{Z} = \mathbf{FA}$). As we described earlier how to calculate the matrix of factor loadings $\mathbf{A}$, we can now rewrite the equation to find the factors and their respective scores. Given that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ (the product of a matrix with its transpose leads to the identity matrix $\mathbf{I}$) and since $\mathbf{FI} = \mathbf{F}$ (because $\mathbf{F}\mathbf{F}^{\intercal} = \mathbf{I}$), we can find the factors by:

$$\begin{aligned} \mathbf{Z} &= \mathbf{FA} \\ \mathbf{Z}\mathbf{A}^{-1} &= \mathbf{F}\mathbf{A}\mathbf{A}^{-1} \\ \mathbf{Z}\mathbf{A}^{-1} &= \mathbf{FI} \\ \mathbf{Z}\mathbf{A}^{-1} &= \mathbf{F} \end{aligned} \tag{3.24}$$

In other words, we can derive the matrix of factor scores $\mathbf{F}$ by multiplying the inverse of the factor loadings $\mathbf{A}^{-1}$ with the original matrix of variables $\mathbf{Z}$. It is these factor scores that we can use to position the original data on the factors (Blasius and Thiessen 2012, p. 38).

Once we calculate the factor scores, there are two more problems to solve before we can apply PCA to a data-set: how many components to keep and whether to rotate the PCA solution and if so, using what technique?

On the number of components, an often used criterion is Kaiser's criterion, which suggests keeping all factors where $\lambda_k > 1$. The reasoning behind this is that each latent factor should at least explain more variance than a single original variable (which always have an explained variance of 1 - the correlation with itself). Yet, not only is this criterion quite arbitrary, but it also non-robust as a higher number of original variables — which leads to less explanation by a single variable — would need a factor to explain less variance than when there would have

been only a few variables. While other solutions, like the criterion by Jolliffe (2004, p.111-149) (retaining all factors with $\lambda_k > 0.7$) and parallel analysis (Horn 1965) are available, the most common method is the *scree test* advocated by Cattell (1966). This test calls for a plot with the eigenvalues on the Y-axis and the successive components on the X-axis. Using this plot, we keep all factors before the point of inflexion (where the line of the graph flattens). Finally, the number of components also depends on the aim of the analysis. Often, we use two components for aesthetic reasons, even though the criterion used would tell us otherwise.

We can use factor rotation to better interpret the factors (or principal components) that emerge from PCA. With rotation, we maximize the loadings of the variables on one factor and minimize them on the other. The rotation itself can be either *orthogonal* (e.g. varimax and quartimax) or *oblique* (e.g. oblimin, promax, and simplimax). Under the former, we rotate the factors but keep their orthogonality, while under the latter, we allow the factors to correlate. This means the factors do not need to be orthogonal. Which of the methods to choose once again depends on our goal and sometimes applying rotation might make actually make the results worse. Blasius and Thiessen (2012) for example, show that rotations can lead to erroneous conclusions about if a concept exists of two unipolar dimensions or a single bipolar one. Thus, we should always study the unrotated solution before rotating.

From here on, we will use scree plots to establish the number of factors to keep and will only apply rotation if the unrotated solution to the PCA does not allow us to draw any interesting conclusions, and when rotating makes theoretical sense (especially when using oblique rotation).

3.1.3 Principal Component Analysis vs. Factor Analysis

In the literature, PCA is often confused with factor analysis (FA). The difference between them is simple in that PCA does not incorporate an error term (Blasius and Thiessen 2012, p.41). These errors are unique variances only associated with a single variable and have nothing in common with the remaining variables (Dunteman 1989, p.55). As such, PCA is able to fully extract the variance — represented by the communality — and FA is not (Widaman 2007). Whether to use FA or PCA is thus dependent on the assumptions one can and wants to make about the data. In other words, if we have a mathematical model (assumptions) from which to estimate the factors we use FA and if we want to decompose complex data into simpler linear variates we use PCA (Field, Miles, and Field 2012, p.760). In my case, I use PCA as my goal is not to study the existence of certain expected factors but to discover these factors in the data I am studying.

3.2 Categorical Principal Component Analysis

One of the disadvantages of PCA is it treats all data as numerical. Also, it expects all the relationships between variables to be linear (Linting 2007, p.337). This is problematic when using Likert-type scales (as is common in many surveys) - as PCA forces the categories to occupy equal distances. Instead, categorical PCA (catPCA)[6] calculates the distances between

[6] This method is also known as *non-linear* or *non-metric Principal Component Analysis*. Besides, Gifi (1980, 1990) places the method in a broader framework termed "homogeneity analysis".

the categories. So, while PCA sees the values or the categories ($1, 2, 3 \ldots$ etc.) as actual values which are then standardized, in catPCA calculates new *optimal scores* or *quantification values* using a process called *optimal quantification*[7] instead of the original categories. Optimal quantification is a quantification technique that assigns numerical values to qualitative scales taking the restrictions of the measurement characteristics of the qualitative variables into account (Mori, Kuroda, and Makino 2016, p.9; Young 1990, p.358). Optimal quantification replaces the categories with category quantifications so that they account for the highest amount of variation in the original quantified variables (Linting 2007, p.338). If all variables in catPCA are numeric, no optimal quantification is needed, and the results of catPCA and PCA will be the same (Blasius and Thiessen 2012, p.133-138; Linting 2007, p.338).

Once the catPCA replaces the original scores with the optimal scores, the procedure is like PCA (Blasius and Thiessen 2012, p.41-46; Linting 2007). The only other difference is that unlike in PCA, we have to specify the number of dimensions (m) in advance. This means that the catPCA solutions for m and $m+1$ are not nested. In practice, it means that when we run a catPCA with two dimensions specified, the results for the first dimension would be different then if there would be only one dimension specified. The way in which we carry out further analysis depends on the precise method one uses[8]. The most popular method is the one developed by Gifi (1980, 1990)[9]. This method allows for nominal, ordinal and metric data, as well as non-parametric permutations of the data. There are implementations in various statistical packages (such as **homals** in **R** and **PRINCALS** in **SPSS**). It employs an iterative process in that performs the optimal quantification and PCA at the same time. The process alternates through an iterative algorithm that ends when the optimal quantifications do not change any more (Linting 2007). Here, we will give a basic overview of this approach and how we can use it to study categorical data. See for a full overview of the technique Gifi (1980, 1990), Meulman, Kooij, and Heiser (2004), Leeuw (2006), and Linting (2007).

Remember from PCA that we could write the our matrix of variables ($\mathbf{Z}$) as the product of a factor ($\mathbf{F}$) and the factor loadings ($\mathbf{A}$):

$$\mathbf{Z} = \mathbf{FA} \tag{3.25}$$

In catPCA we replace the matrix $\mathbf{Z}$ with variables z and categories z_j by a new matrix $\mathbf{Q}$ with variables q and *optimal scores* q_j (with $q_j = \phi_j(z_j)$)[10]. To get the factors $\mathbf{Z}$ and the associations $\mathbf{A}$ in PCA, we used *eigenvalue decomposition*, which through the use of eigenvectors and eigenvalues gave us the information we needed. The approach by Gifi (1980, 1990) is to replace this process by using a least squares *loss function*[11] and minimizing it (Linting 2007, p.356). This allows us to not only include metric data, but categorical data as well. For metric data, the results will be the same as during regular PCA. Because the values in the original matrix $\mathbf{Z}$ do not need to change, we only calculate the loss over the product

[7] This process is also known as *optimal scaling* or *optimal scoring*.

[8] Methods other than the one discussed here are the "distance" approach by Meulman (1986). See also Leeuw (2006) for an algorithm based on the *principle of majorization*.

[9] "Albert Gifi" is the *nom de plume* of various members of the Department of Data Theory of the Universiteit Leiden (Jolliffe 2004, p. 374)

[10] ϕ_j indicates the transform to the original data of z_j

[11] A loss function is a function that tells us how much information we have "lost" when going through a certain procedure.

of the matrices **FA**. We can give this loss function as (Leeuw 2006):

$$\sigma(\mathbf{F}, \mathbf{A}) = \sum_{i=1}^{n} \sum_{j=1}^{s*} (z_{ij} - \mathbf{f}_i^\intercal \mathbf{a}_j)^2 \tag{3.26}$$

Note that we sum over the number of variables $s*$ and not s as our aim is to reduce the number of variables. When we perform catPCA we additionally have to take into account certain transformations in $\mathbf{Z}$, which in catPCA additionally becomes the matrix $\mathbf{Q}$ with the quantifications, leading to:

$$\sigma(\mathbf{F}, \mathbf{A}, \mathbf{Q}) = \sum_{i=1}^{n} \sum_{j=1}^{s*} (q_{ij} - \mathbf{f}_i^\intercal \mathbf{a}_j)^2 \tag{3.27}$$

To arrive at values for $\mathbf{Q}$, $\mathbf{A}$, and $\mathbf{F}$ we use an *alternating least squares* method. This iterative method updates one of the three values, while keeping the other two fixed. As we know the (starting) values for $\mathbf{Q}$ but not those for $\mathbf{A}$ or $\mathbf{F}$, most algorithms choose values for $\mathbf{F}$ at random at the start. We continue this process until the improvement in loss values $\sigma(\mathbf{F}, \mathbf{A}, \mathbf{Q})$ becomes lower than a set threshold (Linting 2007, p.357). In effect, solving this loss function means that we combine both the quantification task and the PCA model in a single iterative process[12].

We can restrict this loss function in several ways to allow for nominal, ordinal and numeric data as well as various transformations of the data (Leeuw 2006). For *nominal* data, there is no restriction on the values in $\mathbf{Q}$), and the solution, for them, would be the same as when we would apply Multiple Correspondence Analysis (MCA). Thus, we place a category so that it is at the centre of those variables with which it is most related. For *ordinal* data, we restrict the first column of $\mathbf{Q}$) to be either increasing or decreasing, while the rest is free. This in effect makes the rank[13] of our matrix $\mathbf{Q}$) equal to 1. When visualized, this forces the category points of $\mathbf{Q}$ to lie on a vector that goes through the origin (Meulman 1998). In the case of *metric* data, the first column in $\mathbf{Q}$) is fixed and linear, while the other categories are free (Leeuw and Mair 2009).

The main visualization applied with catPCA is the biplot. This two-dimensional plot shows two axes given by the principal components and the analysed variables given as vectors going through the origin and the point with the component loadings of the variable on both principal components as its coordinates. The closer two vectors are to each other, the closer is the association of their corresponding variables Blasius and Thiessen (2012, p.42). On these vectors, the biplot positions the category points by multiplying the category quantifications with the corresponding factor loadings on both the principal components. The order of the category points is then similar to the order of the quantifications, and the origin represents the mean of the quantified variable (Linting 2007; Linting et al. 2007). As with PCA, we can rotate the biplots in catPCA if necessary.

[12] Note that we can expand this loss function even further to allow for missing values or different person weights - for more on this, see Linting (2007, p.356-358)

[13] The rank of a matrix refers to the number of different dimensions of that matrix. If we see a matrix as being a build-up of a set of columns, the rank of a matrix is the number of unique columns. Unique in this sense means that a column cannot be a combination or transformation of other columns. As such, when the rank equals 1, this means that any extra columns in the matrix are nothing more than combinations of the first column

3.3 Multiple Correspondence Analysis

Thus far, we spoke about using PCA when we have metric data and catPCA when we have non-metric data such as ordinal or nominal data. While it is possible to analyse nominal data using the framework by Gifi (1980, 1990) which we discussed above, here we follow Blasius and Thiessen (2012) and use the *geometrical* approach developed by Benzécri (1973a,b)[14]. This approach rests on three key ideas: the *geometric modelling* of both variables and individuals in low-dimensional spaces called "maps"; the *formal approach* which bases itself on linear algebra; and the principle of *description*, based on Benzécri's motto that "the model should follow the data, not the reverse" (Le Roux and Rouanet 2010, p.1-2; Blasius and Greenacre 2006, p.6). For categorical data, this approach offers Correspondence Analysis (CA) for two-way frequency tables and Multiple Correspondence Analysis (MCA) for Individuals X Variables tables. Thus, MCA is the extension of CA for multiple variables.

While the algorithm used for both CA and MCA is the same, in CA we place our input in a contingency or stacked table, while in MCA we use an indicator or Burt matrix[15]. Here, we will use the matrix $\mathbf{N}$ with rows I and columns J to denote either the Burt matrix $\mathbf{B}$, the indicator matrix $\mathbf{I}$, the contingency table $\mathbf{C}$ or the stacked table $\mathbf{S}$[16]. First, we divide $\mathbf{N}$ with elements n_{ij} by the sample size n to get the correspondence matrix $\mathbf{P}$ with elements p_{ij}. From this matrix we calculate two new vectors: a vector with the row marginals and a vector with the column marginals[17]. We calculate these by:

$$r_i = \sum_{j=1}^{J} p_{ij} \qquad c_j = \sum_{i=1}^{I} p_{ij} \tag{3.28}$$

Which in matrix notation becomes:

$$\mathbf{r} = \mathbf{P1} \qquad \mathbf{c} = \mathbf{P}^\intercal\mathbf{1} \tag{3.29}$$

We can then rewrite these two vectors to two new diagonal matrices with the values of $\mathbf{r}$ or $\mathbf{c}$ on the diagonal and 0's on the off-diagonal:

[14] Apart from the "Dutch" school of Gifi (1980, 1990) and the "French" school of Benzécri (1973a,b), a third, "Japanese" school exists which bases itself on a similar system called *quantification of qualitative data* developed by Hayashi (1950, 1952, 1953). I do not discuss this method here, but will refer to Nishisato (1980, 2007) for a good overview of the technique

[15] A contingency table or cross tabulation is a table that gives the frequency distribution between two variables, with one variable in the rows and the other in the columns. A stacked table is a table in which the values for all the groups are in a single column, with a second column indicating to which group the value belongs. An indicator matrix (also known as a disjunctive table) is a matrix which codes all responses as dummy variables. Here, the number of variables is equal to the number of categories and is 1 when an individual chose that category and 0 when not. A Burt matrix is a matrix which contains the cross-tabulation between all the variables (and their categories) analysed. This matrix, denoted as $\mathbf{B}$ relates to the indicator matrix $\mathbf{Z}$ by: $\mathbf{B} = \mathbf{Z}^\intercal\mathbf{Z}$.

[16] Note that while we base the matrices $\mathbf{B}$ and $\mathbf{I}$ on the same data, they produce different results as we construct the table in a different way.

[17] Marginals are the rows or columns that contain the sum of all columns or rows in a matrix

$$\mathbf{D}_r = diag(\mathbf{r}) \qquad \mathbf{D}_c = diag(\mathbf{c}) \tag{3.30}$$

Keeping in mind that what we want to visualize is either the differences between individuals (in rows) or the variables (in columns), we need some measure on how different or similar two columns or rows are. In the *geometric* method, we can do this by calculating the χ^2 distances between two values from either the diagonal matrix of row masses $\mathbf{D}_r$ or the diagonal matrix of column masses $\mathbf{D}_c$. The formula for this χ^2 statistic is:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{3.31}$$

with n_{ij} as the observed values and $\hat{n}_{ij}$ as the expected values[18]. If we divide the χ^2 statistic by the sample size n we get a measure of the "total inertia", with a minimum of 0 and a maximum of $min((I, J) - 1)$. Inertia[19] is a central concept in MCA (and CA) and is a synonym for variance (Le Roux and Rouanet 2010, p.18). What it tells us is how far a profile (row or column) is from the expected or average profile. When it is farther away, the χ^2 statistic will be higher and thus also the total variance. If all profiles are the same as the average, the value of χ^2 will be close to 0, with a low variance as a result. In statistics, this notion of inertia is also known as the *mean-square contingency coefficient* and is equal to the square of the *phi coefficient* or ϕ (Greenacre 2017, p.28):

$$\phi^2 = \frac{\chi^2}{n} \qquad \text{and} \qquad \phi = \sqrt{\frac{\chi^2}{n}} \tag{3.32}$$

Note that when we talk about the *phi coefficient* later on, we refer to ϕ and not its square ϕ^2. As the correspondence matrix $\mathbf{P}$ is nothing more than the original matrix $\mathbf{N}$ divided by the sample size n, we can replace the values in the above formula with those mentioned earlier. So, we replace the observed values n_{ij} with the values p_{ij} and the expected values $\hat{n}_{ij}$ with $r_i c_j$ (which is the product of the row totals and the column totals). This gives us:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} s_{ij}^2 = \frac{\chi^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{3.33}$$

with s_{ij}^2 being the square root of the *standardized residuals*. Taking away the square root, we can write s as:

$$s_{ij} = \frac{(p_{ij} - r_i c_j)}{\sqrt{r_i c_j}} \tag{3.34}$$

[18] Expected values are those values that we would expect to appear when there is no relation between the two variables. We can calculate them thus: $(\text{row total} \times \text{column total})/n$, with n being the total number of observations in the table. Here, we use or $r_i c_j$ to denote the expected values as we already divided the correspondence matrix $\mathbf{P}$ by n.

[19] This is an obvious reference to the concept of inertia in physics, which states that the farther a mass is away from the centre of an object, the harder it is to change that object - in other words, the object becomes more inert.

or, in matrix form:

$$\mathbf{S} = \mathbf{\Delta}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\intercal)\mathbf{\Delta}_c^{-1/2} \tag{3.35}$$

with $\mathbf{S}$ being the new matrix of standardized residuals, and $\mathbf{\Delta}_r^{-1/2}$ and $\mathbf{\Delta}_c^{-1/2}$ representing the roots of the row and column totals. Keeping the link to the χ^2 statistic in mind, we see the standardized residuals are nothing more than a measure of how strong the difference between the observed and expected values is. They can thus tell us how strong each cell is contributing to our χ^2 value.

Another way to look at the matrix $\mathbf{S}$ is in terms of similarities and dissimilarities. This way, we can say that each element in the matrix $\mathbf{S}$ tells us how similar or dissimilar two categories of our variables are. If they are very dissimilar, the observed value will be close to the expected value, as we expect no relation between the two. If they are very similar, the observed value will be higher than the expected value as we would expect both categories to occur together more often. As such, the matrix $\mathbf{S}$ contains the dissimilarities between the categories of our variables in the same way as the correlation matrix $\mathbf{R}$ contains these dissimilarities between the variables (Blasius and Thiessen 2012, p.48).

To find the matrix of factor loadings $\mathbf{A}$ and the matrix of factor scores $\mathbf{F}$ in PCA we used eigenvalue decomposition (EVD). We can use a similar method for the matrix $\mathbf{S}$, this time called *singular value decomposition* (SVD). The strength of SVD is we can apply it to any arbitrary matrix, while EVD only works for symmetric matrices. Moreover, when we apply SVD to symmetric matrices (like a correlation matrix) the result would be the same as under EVD[20]. So, for our matrix $\mathbf{S}$ the SVD[21] is:

[20] We can prove this as follows. First, we take a symmetric matrix $\mathbf{A}$ and perform SVD:

$$\mathbf{A} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^\intercal \tag{3.36}$$

.

Because $\mathbf{A}$ is symmetrical, $\mathbf{U} = \mathbf{V}$. So, we can rewrite the above equation to:

$$\mathbf{A} = \mathbf{U}\mathbf{\Delta}\mathbf{U}^\intercal \tag{3.37}$$

which is similar as the result from the EVD mentioned earlier (there written as $\mathbf{R} = \mathbf{B}\mathbf{\Lambda}\mathbf{B}^\intercal$). Note that because we need symmetry, this does not work for all matrices on which EVD is possible.

[21] Given its relation to the EVD, calculating the SVD is in many ways like calculating the EVD. This is because like in EVD, we want to rewrite the matrix $\mathbf{S}$ as a product of its eigenvalues and its eigenvectors (remember $\mathbf{R} = \boldsymbol{\lambda}\mathbf{v}\mathbf{v}^\intercal$). Yet, as $\mathbf{A}$ need not be a symmetrical matrix we need to find a way to make it symmetrical. We do so but applying EVD not on $\mathbf{A}$ but on $\mathbf{A}\mathbf{A}^\intercal$ and $\mathbf{A}^\intercal\mathbf{A}$ (as a matrix multiplied by its transpose is always symmetric). We can do so in the following way. First, let's define the SVD for both the matrix $\mathbf{A}$ and its transpose $\mathbf{A}^\intercal$:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Delta}\mathbf{V}^\intercal \\ \mathbf{A}^\intercal &= \mathbf{V}\mathbf{\Delta}\mathbf{U}^\intercal \end{aligned} \tag{3.38}$$

Next, we need to remember the following:

$$\begin{aligned} \mathbf{U}^\intercal\mathbf{U} &= \mathbf{I} \\ \mathbf{V}^\intercal\mathbf{V} &= \mathbf{I} \end{aligned} \tag{3.39}$$

Moreover, remember that an $\mathbf{I}$ matrix is a matrix with $1s$ on the diagonal and $0s$ on the off-diagonal, so leaving it out does not change anything here. If we combine the above, we get:

$$\begin{aligned} \mathbf{A}\mathbf{A}^\intercal &= \mathbf{U}\mathbf{\Delta}\mathbf{V}^\intercal\mathbf{V}\mathbf{\Delta}\mathbf{U}^\intercal = \mathbf{U}\mathbf{\Delta}^2\mathbf{U}^\intercal = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal \\ \mathbf{A}^\intercal\mathbf{A} &= \mathbf{V}\mathbf{\Delta}\mathbf{U}^\intercal\mathbf{U}\mathbf{\Delta}\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Delta}^2\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\intercal \end{aligned} \tag{3.40}$$

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\intercal \qquad \text{and} \qquad \mathbf{U}^\intercal\mathbf{U} = \mathbf{V}^\intercal\mathbf{V} = \mathbf{I} \tag{3.41}$$

with $\boldsymbol{\Delta}$ being a *diagonal* matrix with (positive) singular values in descending order, the matrix $\mathbf{U}$ with the *left-singular values* and the matrix $\mathbf{V}$ with the *right singular values*[22]. Using these, we can calculate the *standard* and *principal* coordinates of the rows and columns. The difference between the *standard* and *principal* coordinates is that we normalize the *standard* coordinates, while we weight the *principal* coordinates as well. As such, we can only interpret *standard* coordinates on the dimension they relate to, while we can compare *principal* coordinates to more dimensions as they are now on comparable scales (Greenacre 2017, p.62-63). The *principal* coordinates thus give us the position of the rows and columns on the dimensions (akin to the factor scores on the underlying principal components). The standard coordinates $\boldsymbol{\phi}$ for the rows (given by $\mathbf{U}$) are[23]:

$$\boldsymbol{\phi} = \boldsymbol{\Delta}_r^{-1/2}\mathbf{U} \tag{3.42}$$

Similar, the standard coordinates for the columns $\boldsymbol{\Gamma}$ given by $\mathbf{V}$ are:

$$\boldsymbol{\Gamma} = \boldsymbol{\Delta}_c^{-1/2}\mathbf{V} \tag{3.43}$$

The principal coordinates are then nothing more than the standard coordinates weighted by the row and column masses (given by $\boldsymbol{\Delta}$). For the principal coordinates of the rows $\mathbf{F}$ this leads to:

$$\mathbf{F} = \boldsymbol{\Delta}_r^{-1/2}\mathbf{U}\boldsymbol{\Delta} = \boldsymbol{\phi}\boldsymbol{\Delta} \tag{3.44}$$

And for the principal coordinates of the columns $\mathbf{G}$:

$$\mathbf{G} = \boldsymbol{\Delta}_r^{-1/2}\mathbf{V}\boldsymbol{\Delta} = \boldsymbol{\Gamma}\boldsymbol{\Delta} \tag{3.45}$$

Moreover, as with the other methods the square of the singular values is equal to the *principal inertia* or the *eigenvalue* of a particular axis:

$$\lambda_k = \alpha_k^2 \quad \text{with} \quad k = 1, 2(\ldots), K \quad \text{where} \quad K = \min\{I - 1, J - 1\} \tag{3.46}$$

For a further overview of the exact derivation and proof of these formula's, see Greenacre

In other words, we can see the SVD of $\mathbf{A}$ as a combination of two EVD's. Moreover, we see that the matrix with singular values $\boldsymbol{\Delta}$ is the square root of $\boldsymbol{\Lambda}$, that $\mathbf{U}$ contains the eigenvectors of $\mathbf{A}\mathbf{A}^\intercal$ and $\mathbf{V}$ the eigenvectors of $\mathbf{A}^\intercal\mathbf{A}$. We can now compute the eigenvectors and eigenvalues of the matrices $\mathbf{A}^\intercal\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\intercal$ using EVD and use the obtained information to complete our SVD. See for a more detailed description Abdi (2007a) and Greenacre (2017, p.243-244).

[22] Intuitively, we can see the *right-singular values* as vectors that rotate our data, the singular values as a scalar that stretches the data, and the *left-singular values* as vectors that rotate the data again to ensure the data align with the orthogonal axes. See for a graphical overview Lay, Lay, and McDonald (2016)

[23] As with the rest of this chapter, I have option to use the matrix notation for MCA as favoured by Greenacre (2017) and Blasius and Thiessen (2012). This notation is different from the notation by Le Roux and Rouanet (2004, 2010) and the original notation by Benzécri (1973a,b)who favoured a notation based on linear algebra.

(2017) and Greenacre and Blasius (2006) and Blasius and Thiessen (2012). See also Le Roux and Rouanet (2004, 2010) and Husson, Lê, and Pagès (2011) for a non-matrix approach to the same problem.

3.3.1 The Guttman Effect

Before we move on, it might be good to discuss a feature of MCA known as the "horseshoe" or "Guttman" effect. This is the effect that causes the points of the MCA to form a "horseshoe" shape. The orientation of the horseshoe is most often vertical. This is the effect that causes the points of the MCA to form a "horseshoe" shape. Yet, it can also be horizontal, meaning that the vertical dimension is the first dimension. This is because the first dimension orders the categories by values, while the second orders them by response style. Categories with a low value start at the one end of the horizontal dimension and go to a high value on the other. Categories with an "average" response are close to the centre, while categories with an "extreme" response styles are farther away. The result of this is a horseshoe-like shape. Note that this is not a problem - in this case, the term "effect" might be something of a misnomer. Most of the times it is exactly what we want. That is, the choice of categories handles the most variance in the data, followed by the responses styles of the users. If the horseshoe is horizontal, this means response style was more important than the categories, which is a problem. No horseshoe effect at all means that the underlying data has even more problems, as we will see later.

3.4 Classical Test Theory

The basic idea of classical test theory is that the score of an individual on a series of items, X, is equal to the true score that user would have obtained if they had made no errors, T, plus a term representing the errors that they did make, E:

$$X = T + E \tag{3.47}$$

This formula forms the heart of CTT and is relevant for one person on a single occasion. Yet, often we are dealing with more than one person. In these cases, we are not talking about absolute scores, but about the variances of these scores. We can then rewrite 3.47 as follows:

$$\sigma_X^2 = \sigma_{T+E}^2 = \sigma_T^2 + 2\sigma_{TE} + \sigma_E^2 \tag{3.48}$$

We can further simplify this by noting that CTT assumes the following:

As σ_{TE} is nothing more than the covariance between T and E, we can thus simplify the formula to:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{3.49}$$

To get to a measure of reliability, we have to consider we want σ_E^2 to be as small as possible, and σ_T^2 and σ_X^2 as similar as possible. A measurement for reliability of X as a measurement for T can then be:

a. $\rho_{TE} = 0$ The correlation (and co-variance) between true and error scores is 0

b. $\rho_{T_1E_2} = 0$ The correlation (and co-variance) between the true score of one measurement and the error on another is 0

c. $\rho_{E_1E_2} = 0$ The correlation (and co-variance) between the error of one measurement and the other is 0

d. $\mu_E = 0$ The mean error size is 0

$$\rho_X = \frac{\sigma_T^2}{\sigma_X^2} \tag{3.50}$$

As ρ_X nears 1, this indicates that σ_T^2 and σ_X^2 are very similar, and that σ_X^2 is thus a good measure for σ_T^2.

Yet, as it is impossible to measure T directly, it is impossible to measure σ_T^2 as well. But there are various ways to get a good estimate of what T might be. The most common of them is with help of the test-retest method. This method is related to the most common idea of reliability and sees whether a series of items — on repetition — would have yielded the same result. If possible, we would repeatedly administer the series of items and compare the results - if the results are similar the reliability is high, when the results are very dissimilar the reliability is low. Yet, not only is it often unpractical or impossible to run a retest, but there are also problems with users remembering the items or having changed their mind in the meanwhile - all leading to them to provide different answers (Carmines and Zeller 1979, pp.37-40).

A more useful method is that of parallel measurements. Parallel measurements are measurements that have identical scores T and equal variances (Carmines and Zeller 1979, p.32). Thus, a measure X is paralel to X' when:

$$\begin{aligned} X &= T + E \\ X' &= T + E' \end{aligned} \tag{3.51}$$

with

$$\sigma_E^2 = \sigma_{E'}^2 \text{ and } T = T \tag{3.52}$$

Like 3.50, we can make a measurement for reliability by observing the degree to which X and X' are similar - or in other words: to which degree they correlate. So, we can say:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_{T+E}\sigma_{T+E'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_T^2 + \sigma_{TE} + \sigma_{TE'} + \sigma_{EE'}}{\sigma_X \sigma_{X'}} \tag{3.53}$$

However, as earlier saw that CTT assumes that errors are uncorrelated with other errors and true scores, this means that $\sigma_{TE} + \sigma_{TE'} + \sigma_{EE'} = 0$. Moreover, as the same rules go for T as for T' and for X as for X', We can thus rewrite 3.53 as:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T'}^2}{\sigma_{X'}^2} \tag{3.54}$$

Note that this allows us to find the until now unmeasurable σ_T^2, as we know the two other terms of the equation: $\sigma_T^2 = \sigma_X^2 \rho_{XX'}$. This is important, as when we use this in 3.50 ($\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$), we get:

$$\rho_{XT} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\rho_{XX'}\sigma_X^2}{\sigma_X^2} = \rho_{XX'} \tag{3.55}$$

as the two terms of σ_X^2 delete each other. The simple implication of this is we only need the correlation between two items $\rho_{XX'}$ to say something about ρ_{XT}, which, as noted is the reliability of X as a measurement for T. The two measures that follow are all based on this idea.

3.4.1 Cronbach's α

Note that I said that parallel measures are a better alternative to test-retest methods as it foregoes the need of having to replicate the study. While true, parallel measures have their own problem: parallel items are rare. No two items are exactly the same and one could even argue that including two similar items is rather wasteful. Moreover, on closer inspection, a questionnaire with many parallel measures is nothing more as a test-retest within a single test.

As an answer to this, we can look at measures that correlate the items within a test without these items needing to be parallel. We call this *internal consistency*. This rests on the idea that the higher the correlation between the items, the more reliable the scale is. The most well-known of *internal consistency* measures is Cronbach's α. To see how the coefficient works, suppose we have a scale with N items constructed as $X = X_1 + X_2 + \cdot + X_N$, with σ_X^2 being the variance of the full set of items, and $\sigma_{X_i X_j}$ the covariance between items X_i and X_j. Then, Cronbach's α for two items is:

$$\alpha = \frac{N}{N-1} \times \frac{\sum\sum_{i \neq j} \sigma_{X_i X_j}}{\sigma_X^2} \tag{3.56}$$

For more than two items, we can replace $\sum\sum_{i \neq j} \sigma_{X_i X_j}$ by $\sum_{i=1}^{N} \sigma_{X_i}^2$, which is the sum of all the variances of each of the items[24].

Cronbach's α thus looks at the contribution of the total variance of the items to the variance of the scale. The higher this ratio, the more the items have in common (as they share a higher degree of variance) and the more reliable the scale.

Note that we do not have to work with the covariances, but can use the correlations as well. As we define correlation as $\rho_{X,Y} = \frac{\sigma_{(X,Y)}}{\sigma_X \sigma_Y}$, and $\bar{p}$ is the mean of all the correlations between the items, we can redefine Cronbach's α as:

[24] Remember that we can do this, as the variance of a single item is the same as the covariance for two similar items: $\sigma_{X_x X_x} = \sigma_{X_x}^2$.

$$\alpha = \frac{N\bar{\rho}}{1 + \bar{\rho}(N-1)} \tag{3.57}$$

Here, we can see the relation between $\bar{\rho}$ and α: as soon as $\bar{\rho}$ increases, so does α. In other words, α is related to the correlation between the items. If there is high correlation α will be high, and for low correlation, α will be low. Besides, increasing the number of items N also increases α, though given that N occurs both in the nominator and the denominator, this effect will lessen as N increases.

Note that the inter-item correlation matrix can be consisting out of values of Spearman's ρ, or Pearson's r. In the latter case, we would refer to this coefficient as α, in the FORMER as the ordinal version of α. Especially in cases where we use Likert-type scales, we should use the ordinal version of α.

3.4.2 McDonald's ω

While popular, there are several problems with Cronbach's α. Most of these problems have to do with the underlying assumptions of CTT: a. that CTT does not distinguish between the "difficulties" of the different items, b. that CTT assumes that every item measures the underlying latent variable as well as any other, and c. that the error of each of the items is the same. In other words, CTT assumes that all items are similar and have an equal effect on the "true" score, T, we are trying to measure. This is neither realistic nor practical. Another way to think about this is by using the *Spearman Single-Factor Model*. The logic of this model is that it shifts the attention from the individual items to the underlying latent variable we try to measure.

In the Single-Factor Model, we have five variables of interest. Most important is F, which is the position of a person on the underlying latent variables we are trying to measure. Also, we have X_j, which is the score for an item (like the one in CTT discussed earlier), E_j, which represents the unique error related to that item, and μ_j, which is the overall mean for the item among all persons. Most different from CTT is that the Single-Factor Model also includes λ_j, which is the factor loading of the item. This factor represents the importance of that item in measuring the underlying latent variable model. Taken together, this becomes:

$$X_j = \mu_j + \lambda_j F + E_j \tag{3.58}$$

with F being a measure for a person on a certain item. In other words, the score on a specific item is dependent on the average score on that item, the persons score on that item times the loading of that item on the underlying factor and a certain amount of error. To see how this relates to CTT, note that we can replace $\lambda_j F$ for T if we assume that the loadings of all the items are the same - that is, each of the items measures the underlying latent variable equally well.

Let us now look at how this would influence the total score on all the items together. Remember that the total score on a series of items in CTT is $X = \sum_j X_j$ - i.e. the total score for each of the items X_j added up. If we then see that the Single-Factor Model states that $X_j = \mu_j + \lambda_j F + E_j$, we can rewrite 3.49 into: $X = \sum_j \mu_j + \sum_j \lambda_j F + \sum_j Ej$. As $\sum_j \lambda_j F$ is the part of X that all items share due to a common factor, and E_j represents

the differences between the different items, the model is similar to the $X = T + E$ presented earlier. Taking in mind that $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$ and $\rho_X = \frac{\sigma_T^2}{\sigma_X^2}$, we can then define a new reliability coefficient, this time taking the Single-Factor Model in mind.

This coefficient is known as the coefficient ω, as define by McDonald (1999):

$$\omega = \frac{{\sigma_{T_j}}^2}{\sigma\sigma_X^2} = \frac{\sum[\lambda_j]^2}{\sigma_X^2} = \frac{[\sum \lambda_j]^2}{[\sum \lambda_j]^2 + \sum \psi_j^2} \tag{3.59}$$

where psi_j^2 represents the variance not accounted for by the latent factors - i.e. the errors. In other words, ω is the ratio of the variance of the true score T to the total variance of the total score X.

The advantage of ω over α is that ω does not assume that all items are the same, as Cronbach's α does[25]. Also, ω does not assume that the errors are uncorrelated and thus allows for many factors being able to "cause" the variance for each of the items. In other words, ω is another way of specifying α if we are unsure whether the items belong to a common latent variable. Moreover, as with α, we can calculate ω from a matrix consisting of values of Pearson's r or Spearman's ρ.

3.5 Item Response Theory

Earlier on, we discussed the contributions of Classical Test Theory. While useful, CTT has several limitations that have led to the development of a new branch of thinking, known as Item Response Theory (IRT). Remember that the goal of CTT was to reduce the errors E to such a degree that X is equal to T. In other words, the score of the test estimates the "true" score that we want to get at. IRT takes this idea further by fulling focusing on this "true" score - in the terminology here known as the *latent* variable.

To do so, IRT introduces the *item response function* (IRF). This function takes both the characteristics of a single item and those of the user into account (Schuur 2003). As such, it describes the probability that a user will give a certain response to an item based on their true position on an underlying factor. Users who thus score high on a latent factor are thus expected to score high on an individual item as well (Linden and Hambleton 1997).

By taking both the individuality of the user and the item into account, IRT is more flexible than CTT and allows for more variability. Also, as it centres around the latent variable, it allows for an easier building of scales. Here, we will focus on two applications of IRT: a reliability measure known as the Latent Class Reliability Coefficient (LCRC) and a technique to build scales known as Mokken Scaling Analysis (MSA).

3.5.1 The Latent Class Reliability Coefficient

The Latent Class Reliability Coefficient (LCRC) is a new measure of reliability based on latent class models (Ark, Palm, and Sijtsma 2011). These models relate a set of observed categorical variables (or items) to a set of latent unobserved latent categorical variables or items - which is the same as the goal of IRT.

[25] This is also known as τ-equivalence.

To see how this works, we take $\pi_{x(i)}$ to be the probability that a user answers item i correct. Note that a correct answer in an IRT context means that the user answers the item as expected based on their position on the latent dimension we are trying to measure. For example, given an item on leaving the European Union, the correct answer for someone who is a proponent of leaving the European Union would be to *agree* with the statement. An incorrect answer would be when they *disagree*. Going back to the formula, we also take $\pi_{x(j)}$ to be the probability that a user answers item j correct. So, $\pi_{x(ij)}$ is the probability that a user answers both items i and j correct, and $\pi_{x(ii)}$ the probability that they answer item i correct in two independent repetitions - the quantity we are after but do not know as we did not repeat the questionnaire. Furthermore, we let σ_X^2 denote the variance of an item, and $\rho_{XX'}$ the correlation between the original response (X) and the response in a possible re-test (X'). Note that $\rho_{XX'}$ is exactly the same measure as we discussed earlier with classical test theory. Applied to IRT, Molenaar and Sijtsma (1988) show that we can write it as:

$$\rho_{XX'} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sum_i \sum_x \sum_y \pi_{x(i),y,(i)} - \pi_{x_i}\pi_{y_i}}{\sigma_X^2} \tag{3.60}$$

where the nominator is the same as the term σ_T^2 in CTT. The denominator in both the left and right-hand part of the equation represents — in CTT terms — the variance of the "true" score that we are trying to measure. Yet, while we can measure the left-hand side of the equation, the right-hand side is problematic, as $\pi_{x_i}\pi_{y_i}$ represents the probability of obtaining sore x and y on two independent instances of the same item for the same respondent - which, as we saw, is problematic. To get around this, we can estimate the right-hand side of the equation using latent class models. These latent class models are "unrestrictive", in the sense that they do not assume any single underlying latent factor. The formula then becomes:

$$\begin{aligned} LCRC = & \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} \\ & + \frac{\sum_i \sum_x \sum_y [\sum_{u=x}^{m} \sum_{v=y}^{m} \sum_{k=1}^{K} P(\zeta = k) P(X_i = u | \zeta = k) P(X_i = v | \zeta = k) - \pi_{x(i)}\pi_{y(i)}]}{\sigma_X^2} \end{aligned} \tag{3.61}$$

where ζ is the underlying number of latent classes of the series of items. We thus have to define the number of latent classes first, which we can do with confirmatory factor analysis.

As the LCRC does not assume a single underlying factor, as α does and allows for differences in item difficulties, which ω lacks, Ark, Palm, and Sijtsma (2011) find that the LCRC is the least biased and should come closest to the "true" reliability of the model. Moreover, the LCRC does not assume monotonicity and non-intersecting item response functions. While we have not discussed this so far, these characteristics are interesting in the view of Mokken scales, which I will discuss now.

3.5.2 Mokken Scaling Analysis

Mokken Scaling Analysis (MSA) is a method for dimensionality investigation and a measurement model in one technique (Wismeijer et al. 2008, p.325; Schuur 2003). It can discover whether the positions respondents take on a set of items form a single dimension and test to which degree this dimension holds as a scale. It was developed by Mokken (1971) and Mokken and Lewis (1982) to address certain problems arising from the use of Rasch scaling (Schuur 2003). While in the first instance the technique only used dichotomous responses (like Rasch scaling) later extensions allowed more than two (polytomous) response options (Molenaar 1991, 1997).

Mokken scales come in two different forms based on two different underlying models: the Double Monotonicity Model (DMM) and the Monotone Homogeneity Model (MHM). The main difference between the two models is that the former assumes that the difficulty ordering of the items is the same irrespective of the latent variable while the latter does not do so (Stochl, Jones, and Croudace 2012). This is also known as Ordinal Specific Objectivity, Invariant Item Ordering (Schuur 2003, p.144) or the assumption of non-intersecting item response functions (Stochl, Jones, and Croudace 2012). While it has several desirable properties — such as that the order of difficulty of the items is the same for all users — these assumptions are often too severe for use in VAA research. Thus, the MHM is most often used (Germann and Mendez 2016).

Aside from the assumption of non-intersection - which is unique to the DMM, both the DMM and MHM share three other assumptions:

Unidimensionality All items in the scale θ measure a single latent trait. On this scale, all users s have a scale value θ_s and all items i have a scale value δ_i.

Monotonicity The IRF is monotonically non-decreasing. This means that the higher a user is placed on the latent trait, the more likely it is that they get higher scores on the items measuring that same trait.

Local independence The response of a user to an item is not influenced by their responses to other items in the scale. As such, we expect users to start with each item with a "clean mind" with no memory of the previous question.

To test the assumption of *unidimensionality*, we can perform a test of homogeneity based on Loevinger's H coefficient, which measures whether the responses of the items in the scale are consistent (Wheatley 2016). Loevinger's H coefficient equals 1 when the items form a perfect Mokken scale, and 0 when there is no association between the answers of the users on the items (Stochl, Jones, and Croudace 2012). For different items X_i and X_j we define the item scalability coefficient H_i as:

$$H_i = \frac{\sum_{j \neq i} Cov(X_i, X_j)}{\sum_{j \neq i} Cov_{max}(X_i, X_j)} \tag{3.62}$$

while the coefficient H for all k items is:

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Cov(X_i, X_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Cov_{max}(X_i, X_j)} \tag{3.63}$$

Generally, items form a Mokken scale if the coefficient H is larger than 0.3 and each item-specific H_i is also larger than 0.3 (Germann and Mendez 2016). H values between 0.30 and 0.40 form a weak scale, values between 0.40 and 0.50 a medium scale, and all values ≥ 0.50 a strong scale (Mokken 1971, p.185).

To test the assumption of *monotonicity*, we use a criterion called **crit** (Schuur 2003, p.53-54). This criterion takes into account many violations of monotonicity and returns a value which tells us to what degree they are violated. Here, we consider values above 80 to be problematic. See for more information on this criterion Molenaar and Sijtsma (2000).

We can run MSA both exploratory and confirmative. While in the former the aim is to establish Mokken scales from a set of items, in the latter we investigate an existing scale. In their application of MSA to VAA data, Germann and Mendez (2016) first use confirmatory MSA to test whether the ex-ante derived scales hold, and after finding they do not, run an exploratory MSA to establish new scales. Both confirmatory MSA and exploratory MSA (which comes in two different forms) are available in the **mokken** package for **R** (Ark 2007, 2012).

4 | The Quality of Scales

In this chapter, I will focus on the first part of the first research question: what is the quality of the scales Voting Advice Applications use? To find this quality, I will use three measures. The first focuses on the questions that make up the scale. I will call this the response quality of the scale. This measures if users understood the questions and the response categories. The second is the dimensionality of the scale. This measures if the scale is unidimensional or multidimensional. This is important as multidimensional scales are difficult to interpret and can lead to confusion. The third is a measure of reliability. This measures the degree of consistency of the measure. I will apply these three measures to various types of scales I derive from the EUVox VAA.

4.1 The Response Process

Before we can dive into the construction of the scales and their quality, we first have to understand how users reply to the questionnaire. As with all questionnaires, the basic principles are simple: there are a series of questions to which the users provide a series of answers. A valid questionnaire has valid statements and valid answers, without the one implying the other. Here, I will focus on one aspect of this: the way in which the question influences the user. In other words: how do the questions in the questionnaire influence the user? Answering this allows me to answer some other questions as well. Like, why do users with clear preferences sometimes deviate from what we would expect them to reply? For example, why does someone who disagrees with government intervention into the economy also disagree with privatization? Also, why do users use the neutral or no opinion options? Does it mean they are truly neutral on the matter or do they not know what they think? Moreover, is there a logic into which response options users take? Are there some users who only use the "completely agree" or "completely disagree" options and leave out the rest? Or do all users consider all the options as possibilities at all times? And finally, how does the wording of the statement influence the user. Does it matter whether we talk about climate change or global warming? Or does it not?

To understand how these disconnected questions connect, we first have to understand how a user responds to a question. In other words, what happens when a user sees a question on their screen? A practical way to understand this is by thinking of responding as a process (Tourangeau, Rips, and Rasinski 2000, p.7-14). As with any process, it sees the user go (subconscious) through several steps while responding to the question. Figure 4.1 shows what such a response process looks like.

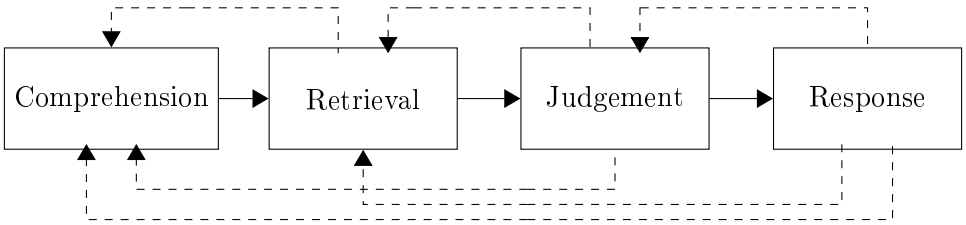


Figure 4.1: The Response Process, adapted from Groves et al. (2004, p.202).

Seen this way, the process might look rather complex. Yet, it becomes clearer if we see the process as four separate steps, at each of which the user can go back to any of the previous steps. The first step is the one of *comprehension*. This refers to the way in which the user gets acquainted with the question and the way in which they choose to react to it. In the second step, users *retrieve* the relevant information for the answering of the question. Especially for attitude questions, users rarely have a clear pre-established view what their position on a question is. Even though they might have established ideas, the precise wording of the question might throw them off-balance. Wilson and Hodges (1992) offer two different ideas of how we can understand this behaviour. Or users have pre-existing evaluations that they need to look up, or they have general attitudes that they use to answer each specific question. As a mid-way Tourangeau, Rips, and Rasinski (2000) offers the idea that while users do have pre-existing evaluations, at any point we ask questions about them, users update them with new information they have collected in the meantime. In the third step, the *judgement* stage, users judge the information retrieved during the "retrieval" stage to reach a conclusion. Finally, during the fourth stage, the users attempt to map their answer to the response options. Also, they might make last-minute edits to their answers. This can be to prevent embarrassment on personal questions, or to give a socially desirable answer. So, to give a valid answer, a user has to *comprehend* the question, *retrieve* the right information, *judge and synthesize* this information correctly, and *respond* in a way that is closest to their judgement (Blasius and Thiessen 2012, p.3). It is not hard to see why this rarely happens. When questions become more difficult or personal, users might not make the correct judgement or edit their final response. Alternatively, when users lack interest in the topic of the question, they might well skim through the process and provide a sub-optimal result.

Whether this happens and the degree to which this happens depends on the behaviour of the user during the survey. In the context of Figure 4.1, we can see this as the environment in which the user goes through the process. We want this response behaviour to be such that the user considers each question with the same interest and responds as well as possible. Yet, we often find they revert to certain response styles. They do this to simplify the response process (Van Vaerenbergh and Thomas 2013; Baumgartner and Steenkamp 2001). The most well-known of these are the Acquiescence and Disacquiescence Response Style (ARS/DARS). Users with ARS tend to agree with the questions, while users with a DARS will disagree. The problem is that they do so regardless of the actual content of the question. Designers of VAAs address this by reversing some questions. So, if a user would have to agree with a question to have to VAA position them as left-wing, there will also be a question where they would have to disagree. Another response style is the Midpoint Response Style (MRS) in which the users only use the middle category of the scale. Such options represent a safe

option for users who lack an opinion or have insufficient knowledge to respond to the question. Thus, the user's non-substantial response appears as a substantial one (Blasius and Thiessen 2001a). Another version of this style is the Extreme Response Style or the Mild Response Style (ERS/MLRS). Users with ERS tend only to use the highest and lowest categories of the scale, while users with MLRS tend to avoid such categories. These two styles identify users who have either very intense or very mild feelings (Moors 2008). Another type of response style is the Response Range (RR), also known as Limited Response Differentiation (LRD). This style has users select a narrower range of response options to consider. The way in which users limit this range depends on individual users, and we often cannot predict it in advance. For example, users might consider only the middle three options of a five-point response scale. Or they might simplify a seven-point response scale to its extreme options. In the latter case, the result is like ERS. Finally, there is the Noncontingent Response Style (NCRS). Here, users provide careless or arbitrary answers. Such behaviour is typical in VAAs for users who want to discover what the VAA looks like and which parties are available. Not all these response tendencies are the same, and they do not always cause a problem for the data. They are only a problem as soon as there is an association between them and any of the independent variables (Blasius and Thiessen 2012). The strength of this association depends on the difficulty of the questions and the characteristics of the user (Leeper 2014; Knowles and Condon 1999).

4.1.1 Question Difficulty

Question difficulty itself depends on the formulation of the question and the type of response options. For the questions, one of the most frequent problems is that questions are too lengthy. Users often have to read such questions twice to avoid getting lost in them. Often they come with *qualifications*. These are extra — but not necessary — pieces of information(Camp, Lefevere, and Walgrave 2014, p.17). In a question like "external loans from institutions like the IMF are a good solution to crisis situations", the IMF is a qualifier. Often the designers' intention of the qualifier is to aid the user. Yet, often it leads to the user only considering the qualifier and not the possible other options (Gemenis 2013b, p.273). Another problem occurs when questions contain jargon, technical terms and acronyms. They are a problem if users are not acquainted with them and do not know what they mean. In the EUVox for Cyprus, the question "The privatization of AHK and CYTA should be completed" assumes that users know that CYTA is a telecommunications provider and AHK a public electricity company. If they are unaware of this, they might choose to skip the question or provide a random guess (Oppenheim 1992, p.128-129).

More problems can occur when the question contains a negation. The question "EU citizens living in Germany should not have access to the same social benefits as German citizens" creates difficulties for users because those who agree that EU citizens living in Germany should have access have to choose the "disagree" option to provide the correct answer (Schriesheim, Eisenbach, and Hill 1991). Related to this are the *double-barrelled questions*. These are questions that need a single answer for what are in fact two different questions. An example from EUVox is the question "There should be a common EU foreign policy even if this limits the capacity of Country X to act independently". This question

contains two questions: a) whether there should be a common foreign policy and b) whether the capacity of countries to act independently should be limited. While a user might agree with (a) they might disagree with (b) or vice versa. For such a user, selecting the "agree" or "disagree" option does not correspond with their opinion. Also, given that different users might read the question in a different way and either assign more weight to part (a) than to part (b), users with a similar opinion might respond in a different way (Oppenheim 1992, pp. 126–128; Bradburn, Sudman, and Wansink 2004, 142–145). Also problematic are the *quantifications*. We can find these in questions that quantify a certain object in a sentence, for example by adding words like "more", "lower", or "better". A typical example is the EUVox question "The number of public-sector employees should be reduced". Not only is it unclear what "reduced" means, but including such a quantifier also makes the statement more difficult (Krosnick and Presser 2010, pp. 265–266). Because what does it mean when a user disagrees with this question? Does it mean that they do not want to reduce the number of public-sector employees, or that they want to increase (the opposite of reducing) their number? Besides, the user should be aware of the current status quo before agreeing or disagreeing on either an agreement or disagreement. As such, the interpretation of what "reduce" means will differ between different users (Schaeffer 1991). Moreover, its meaning will differ between different questions as well. For example, "reduce" in "reduce the number of employees" means something different from "reduce" in "reduce the number of nuclear weapons" (Windschitl and Wells 1996). This places an even higher burden on the user if such questions were to follow each other in the survey. Finally, we need to make sure that the formulation of the questions makes sense to the user. While the formulation might make sense to the designers, it might not make sense to someone who is alien to either politics or the topic of the question. Blasius and Thiessen (2012, p.5) call this the "cultural distance bias hypothesis". This means that the larger the distance between those that create a question and those that respond to it, the more problems will arise.

With regard to the response options, there can be problems as well. To start with, it matters how many response options are available. This is because as users answer different if confronted with a large or small number of options (Gaskell, O'Muircheartaigh, and Wright 1994). We should thus find a balance. To find this balance, we have to make sure that users can map their opinion to the response options as precise as possible. Also, the response options must map the entire measurement continuum and appear to be ordinal. This means that if it is possible to completely disagree with a question, this option should be available (Krosnick and Presser 2010, pp. 238–275). Combining these requirements makes the choice for the number of response options a trade-off. Many options make the responses more precise but make the question more difficult (Couper et al. 2006). Fewer options provide less precision while making the task of the user easier. In the case of VAAs, Likert scales are most often used with either 3 (Stemwijzer) or 5 (Kieskompas) response options. Other variants, like a thermometer with 101 response options, exist but are rare. This is because most users tend to simplify such scales to numbers divisible by 10 (Kroh 2007). Another issue is how we label the response options. For VAAs the most frequent occurring options are either the 3-point scale with agree/no opinion/disagree labels, or the 5-point scale with completely disagree/tend to disagree/neutral/tend to agree/completely agree labels. While the intention of these labels is to provide more clarity for the user which option to choose, Gemenis (2013b) points out

that such labelling confounds direction (agree/disagree) with intensity (completely/tend to). As such, it forces users to think along many dimensions, complicating their task. Besides, the order in which designers list these responses can influence the responses as well (Liu and Keusch 2017; Yan and Keusch 2015). Thus, a scale running from *disagree* to *agree* is different from one running from *agree* to *disagree*. Related to this is the use of the Neutral and Do Not Know options. Designers include neutral options in most VAAs as "neutral", "neither agree/disagree", or "open-minded". The advantage of this is that it allows ambivalent users to express their position. Yet, it also allows them to hide unpopular opinions (Johns 2005). Also, they may choose this option when they have no opinion. This can be because a no opinion response was not offered, or to make a good impression (Blasius and Thiessen 2001a). Yet, including a no opinion option can also provide another way for users to hide their opinion. In VAAs, the option is often included, either as "no opinion" or as a "skip" option. Finally, there is the issue of any *reference points*. These refer to any information in the response scales towards what the current status quo is. Currently, I know of no VAAs that include such reference points in their scales. Still, Gschwend and Proksch (2010) show that including such an option reduces non-response. This is because reference points make it easier for the user to grasp the scale. Besides, it makes it easier for the users to understand the directional aspect of the scale as it is clearer what "tend to" or "completely" refer to.

4.1.2 Personal Characteristics

There are many types of personal characteristics. Most important are the level of education of the user, their age and income, their personality, and their culture. For culture, Weech-Maldonado et al. (2008) finds evidence that Hispanics tend to favour ERS, while Baron-Epel et al. (2010) report higher levels of ERS and MRS for Jews than Arabs in Israel. Also, Van Vaerenbergh and Thomas (2013, p.204-205) find differences between countries for levels of IQ and corruption as well as geographical location. While for most VAAs this is unimportant, for a trans-national VAA like EUVox this means that we should adjust any results for the response styles. Another part of the personal characteristics is the personality of the user. Studies like those by Austin, Deary, and Egan (2006) show that use of ERS relates to certain personality traits. Still, most of these studies use rating scales which themselves are subject to response tendencies (Van Vaerenbergh and Thomas 2013, p.204). Besides, most of the effects found are marginal (Blasius and Thiessen 2012). Other studies on personal characteristics show similar results. For the level of education, there is an inverse relationship between the use of response tendencies and level of education. But, as with the personal characteristics, the precise strength of this relationship differs between studies and is most often small (Van Vaerenbergh and Thomas 2013; Blasius and Thiessen 2012). For the user's age, there is even more uncertainty, though Jong et al. (2008) finds that both younger and elderly users show signs of ERS, indicating a curvilinear relationship. The only other tendency for age reported, ARS also has inconsistent findings. Some scholars find an effect, while others report no effect (Van Vaerenbergh and Thomas 2013, p.202). For gender, results are also confusing. Austin, Deary, and Egan (2006) finds evidence of an effect, while Marin, Gamba, and Marin (1992) fail to do so. The results for income are more consistent. Here, there is agreement being that both ARS and ERS are higher when income is lower (Greenleaf 1992). The main reason that

there is so much disagreement is that scholars have to deal with impression management. This refers to the idea that individuals at any time manage the impression they want others to have of them. In the case of surveys, this occurs when users provide socially desirable responses. Or when their response is different depending on the social status of the interviewer. While it might seem that impression management is less important in the case of online surveys, the contrary seems to be the case (Mueller et al. 2014). Dodou and Winter (2014) find impression management strategies are similar in offline, online, and paper surveys.

To summarize, we get responses of high quality when the question is not too difficult and matches the users' characteristics. This means a difficult question is not a problem for a user with a high level of education but can be a problem for a user with a low level of education. As soon as this match is wrong, and the question is too difficult, users will engage in *task simplification*. This task simplification is part of a broader idea of *satisfycing* (Krosnick 1991; Simon 1957). Here, users do not choose the perfect response, but the one which is "good enough". In the worst case, this might go so far that the user refuses to answer the question. For simple questions like "In which part of the country do you live?" simplification is not likely. For more difficult questions, such as "to maintain public order, governments should be able to restrict demonstrations" it is more likely. Unfortunately, most of the questions in VAAs are difficult as a result of their political nature. Political issues are rarely clear-cut and sometimes require some knowledge from the user. Moreover, designers of VAAs might want to use double-barrelled questions to illustrate a policy trade-off, include quantifiers to provide the necessary context to the question, or compress many ideas into a single question to reduce the number of questions (Gemenis 2013b).

For the VAA data in this and the following chapter, I thus expect the following. First, that long and more complicated questions will lead to more task simplification by the users. We can see this simplification when they only use certain response categories or often skip the question. Second, I expect that the higher the level of education of the user, the lower the amount of simplification is. Third, I expect there will be differences in the degree of simplification between various countries.

4.2 The Data-Set

EUVox[1] was operating during the 2014 elections for the European Parliament (Mendez and Manavopoulos 2018). It had versions for 28 countries, allowed users to position themselves on 30 different questions, and showed their position on both political maps and by the use of matching percentages. The questionnaire used a Likert scale with possible responses consisting of "completely agree", "agree", "neither agree nor disagree", "disagree", "completely disagree" and "no opinion". For each version, users could choose between English or the main language spoken in their country. Of the 30 questions, 21 were common to all countries (20 in the case of France). Seven (six in the case of France) of the core questions concerned powers of the EU, seven handled economic issues, and another seven handled cultural issues. The 21 core issues were selected to cover the key issues found by the Chapel Hill Expert Survey data (Polk et al. 2017). The remaining issues (9 for most countries and 10 in the case of France)

[1] `http://www.euvox2014.eu/`

were selected by country experts to capture salient topics specific for their countries.

Code	Question
EU1	Country X should exit the Euro (Eurozone countries)/ never adopt the Euro (non-Eurozone countries)
EU2	A single member state should be able to block a treaty change, even if all the other member states agree to it
EU3	The right of EU citizens to work in Country X should be restricted
EU4	There should be a common EU foreign policy even if this limits the capacity of Country X to act independently
EU5	The EU should redistribute resources from richer to poorer EU regions
EU6	Overall, EU membership has been a bad thing for the Country X
EU7	EU treaties should be decided by [name of national parliament] rather than by citizens in a referendum
EC1	Free market competition makes the health care system function better
EC2	The number of public-sector employees should be reduced
EC3	The state should intervene as little as possible in the economy
EC4	Wealth should be redistributed from the richest people to the poorest
EC5	Cutting government spending is a good way to solve the economic crisis
EC6	It should be easy for companies to fire people
EC7	External loans from institutions such as the IMF are a good solution to crisis situations
CU1	Immigrants must adapt to the values and culture of Country X
CU2	Restrictions on citizen privacy are acceptable in order to combat crime
CU3	To maintain public order, governments should be able to restrict demonstrations
CU4	Less serious crimes should be punished with community service, not imprisonment
CU5	Same-sex couples should enjoy the same rights as heterosexual couples (FR)/ to marry (remaining countries)
CU6	Women should be free to decide on matters of abortion
CU7	The recreational use of cannabis should be legal

Table 4.1: EUVox - Common Questions

Table 4.1 shows the common set of questions. Seven of them are about the EU. These questions focus on issues relating to the Euro, common foreign policy and EU membership. Another seven questions cover economic issues. These questions deal with state intervention in the economy, free market, and distribution of wealth. A final set of seven questions deal with cultural issues such as immigration, same-sex rights, and abortion. The VAA designers

chose these three issues to cover similar topics as those found by Chapel Hill Survey data (Polk et al. 2017). The designers constructed the other questions in cooperation with the country teams. They did this to capture specific salient topics at that moment (Wheatley 2016). These questions could belong to one of the three scales.

Country	Raw dataset	Clean dataset	% Clean
Austria	10,669	7,527	71%
Croatia	7,666	5,281	69%
Czech Republic	28,630	24,084	84%
Denmark	126,261	92,633	73%
Estonia	18,172	12,267	68%
Finland	8,274	6,729	81%
France	8,704	6,656	76%
Germany	9,658	7,208	75%
Greece	63,687	46,098	72%
Hungary	6,711	5,536	82%
Ireland	9,523	6,198	65%
Italy	36,614	26,235	72%
Lithuania	9,072	7,050	78%
Poland	73,521	58,429	79%
Portugal	54,165	42,199	78%
Slovakia	7,238	5,905	82%
United Kingdom*	100,897	77,403	77%

*Includes only England

Table 4.2: Number of users in the EUVox data-set.

Of the 28 available countries, I exclude 11 from this analysis. I exclude The Netherlands and Sweden because the VAAs used different questionnaires. Also, I exclude Belgium, Latvia, Luxembourg, and Malta because of the limited number of users. Finally, I exclude Bulgaria, Cyprus, Romania, Slovenia and Spain as no information on the certain scales was available. Also, as the data-set for England was larger than those for Northern Ireland, Scotland and Wales I exclude the latter three.

Table 4.2 shows the number of users for the remaining 22 countries. Here, the second column shows the number of entries in the data-set for each country. The third column shows the entries remaining after I cleaned the data-set. The fourth column shows the percentage of clean data compared to the raw data. I have to clean the data because as with other online questionnaires users fill out VAAs unsupervised. Thus, they can fill out the VAA many times, click through the questions without reading them or use the same response category for each question. To clean the data, I follow the steps suggested by Andreadis (2014) and Mendez, Gemenis, and Djouvas (2014). So, I remove all entries that: 1) users filled out using a mobile phone, 2) were filled out by returning users (based on an identifier based on their IP-address), 3) where the time taken to complete all the 30 questions was less than 120 seconds, 4) where the time taken to complete all the 30 questions was more than 5400 seconds (90 minutes), 5) where the time taken to respond to any of the questions was less than 2 seconds, 6) where the time taken to respond to three or more issues was 3 seconds or less, 6) where users responded to more than 10 issues in the same way, and 7) where users skipped more than 10 questions

using the "No Opinion" response. The number of entries I retained varies between countries and ranges between 65% (for Ireland) and 84% (for the Czech Republic).

Country	Original	Number of Users with:								Final
		Mobile	Return	Ans. < 2 sec.	> 3 Ans. $< 3s$	$> 5399s$	$< 121s$	> 10 Similar	> 10 NA	
AT	10,669	1,538	951	398	215	6	22	5	7	7,527
BG	7,214	364	636	287	63	19	10	31	28	5,776
CY	5,176	671	613	225	87	6	3	20	9	3,542
CZ	28,630	1,059	1,345	849	1,098	48	72	33	42	24,084
DE	9,658	1,201	379	408	370	18	45	10	19	7,208
DK	126,261	23,558	5,552	2,936	832	145	124	72	409	92,633
EE	18,172	577	628	3,432	1,168	29	0	19	52	12,267
ES	159,552	41,716	16,855	3,478	1,981	124	159	139	396	94,704
FI	8,274	535	411	296	225	12	24	8	34	6,729
FR	8,704	736	690	346	229	16	9	15	7	6,656
GR	63,687	7,020	7,907	1,525	698	81	91	230	37	46,098
HR	7,666	863	948	285	238	7	17	19	8	5,281
HU	6,711	209	580	227	119	6	15	5	14	5,536
IE	9,523	1,894	458	369	517	11	63	5	8	6,198
IT	36,614	6,034	2,691	849	600	27	89	56	33	26,235
LT	9,072	643	900	265	161	20	17	7	9	7,050
PL	73,521	4,201	6,845	1,951	1,621	94	162	141	77	58,429
PT	54,165	4,461	4,925	1,404	817	96	101	73	89	42,199
RO	9,508	556	616	310	53	13	7	11	20	7,922
SI	3,842	236	443	181	168	5	11	5	10	2,783
SK	7,238	525	389	247	129	11	12	16	4	5,905
UK*	100,897	6,730	5,425	3,782	6,406	75	803	117	156	77,403

*Includes only England

Table 4.3: Number of users in the EUVox data-set. Each column shows the number of users with the specific characteristics mentioned.

Table 4.3 shows the cleaning procedure in more detail. I removed most users from the data-set because they used the version of the VAA on their mobile phone or because they were returning users. Especially in the case of Spain, I had to drop the data of 58,571 users, or 36.71% of the original users because of these two considerations. The reason I removed the mobile-phone users is that the user had to swipe the statement to the left to respond with a "No Opinion". As this is different from the obvious button that the website had, it might not have been clear for all users that no opinion was possible. To avoid this confusion, I dropped the mobile phone users from the sample. The returning users are less of a problem. On average, I only dropped 7,83% of the users. For the cleaning based on the responses, I removed most entries because there were answers under 2 seconds. Of note here is the case of Estonia, where I removed 18.91% of the entries whereas the average of all the other countries is 3.35%. Given that Estonia also loses most entries when I remove those that have more than 3 answers under 3 seconds (6.43%) indicates that users from Estonia replied quickly to the statements. This might be an early indicator of the problematic nature of the data from this country, as will I show later on.

4.3 Scales in EUVox

I now turn to the scales that EUVox uses. Until now, VAA designers have assigned the questions to the scales based on where they think they belong. This approach is simple and fast, but not always correct. For example, Louwerse and Otjes (2012) and Gemenis (2013a) found the scales for the 2009 EU Profiler lacking in unidimensionality. This makes the scales hard to interpret for the user and possibility meaningless. To address this issue, scholars made various attempts to improve on these *original* scales. One way is by using a combination of Exploratory and Confirmatory Factor Analysis (EFA and CFA) (Wheatley 2015a,b). The advantage of this method is it is well understood and produces many goodness-of-fit indexes that we can use to assess the quality of the scale. Yet, the problem with EFA and CFA is that both methods assume metric data, while VAA data is often of an ordered-categorical kind. Using EFA and CFA on such data also may lead to over-dimensionalisation. This means that there is likely to be a dimension measuring not the dimension of interest, but the difficulty of the questions (Eijk and Rose 2015). An alternative is to use IRT models like MSA. This is the approach taken up by Katsanidou and Otjes (2016), Mendez and Wheatley (2014), Wheatley (2016), Wheatley, Carman, et al. (2014), and Wheatley and Mendez (2018). The advantage of MSA is that it does not assume all questions in the VAA have the same difficulty, which EFA and CFA do. This is helpful, as users of VAAs are often more likely to agree to some questions than to others. So, we can use MSA to arrive at unidimensional scales of hierarchically ordered questions that measure a single underlying dimension[2]. Besides, scholars sometimes combine MSA with other methods. Gemenis (2013a) uses a combination of factor analysis, MSA and item-rest scores to test for a correct degree of monotonicity (which indicates a higher score on

[2] Note that this claim is not uncontroversial. For example, Smits, Timmerman, and Meijer (2012) show that what MSA does is create Mokken scales: scales that are as long as possible with a given minimal strength to each of the questions (based on Loevinger's H values). This would mean that a Mokken scale can be multidimensional, or that many Mokken scales can refer to a single dimension. Yet, this is most likely only the case when the values of H and H_i are small. While I will use MSA to establish dimensionality I keep this in mind when dealing with scales with low values of H and H_i.

the question goes together with a higher score on the underlying latent dimension). Shikano (2013) meanwhile, uses Bayesian Markov-Chain-Monte-Carlo inspired models to estimate dimensions based on the positions of the parties in the VAA, while König and Waldvogel (2018) and König and Nyhuis (2018) use Multidimensional Scaling (MDS) to transform VAA related data into a map that reveals the political space of France during the presidential elections of 2017. Finally, Germann, Mendez, et al. (2015) and Germann and Mendez (2016) uses a combination of MSA and several measures of reliability. They call this approach "dynamic scale validation" (DSV). The advantage of it is that we can use it while the VAA is still in operation. This is because we do not need to use all users, but only a set number of early users (the first $2000 - 5000$ entries).

During its operation, EUVox used two kinds of scales. The first are the original scales that were there when the VAA launched. The second are the scales that resulted from the dynamic scale validation that took place sometime after the launch. Besides these two, I will also use a third kind of scales based on a quasi-inductive method (Wheatley 2015a). I construct these scales in a similar fashion as the scales created by the DSV. The difference is that in DSV I restrain the MSA to the questions assigned to it in the original scale, while here I remove that restraint. Thus, I allow MSA to use all the questions. While this often leads to longer scales, it also means that sometimes only a single emerges.

	EU							EC							CU							AD									
Country	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
Austria	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		⋆		●	⋆	⋆	○	○		
Croatia	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●	●	○	○	○	○			
Czech Republic	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●		⋆	⋆	⋆	○			
Denmark	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆	⋆	●	⋆	⋆	⋆			
Estonia	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆	○		○	●				
Finland	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●										
France	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		○		●	⋆	⋆	○	○	●	●
Germany	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●				⋆	⋆					
Greece	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●		○	○	●	●	●	⋆		○	
Hungary	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	⋆						●	●	
Ireland	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○		○	●	⋆	⋆	○	○	○	
Italy	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	○	○	○	○				
Lithuania	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	○		⋆	○		●			
Poland	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	●	●	●	⋆	○	●				
Portugal	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	⋆	○	○	○	○	○	○	●	○	
Slovakia	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	●		⋆	⋆			⋆	○	
United Kingdom	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○	○	●	●	●	●	●	●	●	○	⋆	●	●	●	⋆	⋆	○	○	

Table 4.4: Items included in the original EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

Table 4.4 shows the *original* scales. In this table, questions with a $\star$ belong to EU scale, questions with a $\circ$ to economic scale, and questions with a $\bullet$ to cultural scale. Clear from this overview is that in case of the cultural and economic issues, the scales used all questions, while in the EU case, they only used only six of the seven questions. The question left out is the $EU7$ question relating to referenda on EU treaties. While it is unclear why the designers left this question out, it is most likely because the topic did not belong to an EU dimension. Whether this is the case, is something I will look into later when I construct the other scales. Furthermore, the inclusion of the additional questions differed per country. Where Finland uses none of these questions, other countries sometimes use almost all, most often for the economic scale. In the case of Portugal, the designers used 7 of the 9 for the economic scale, while both Croatia and Italy use 4 of them. The EU and cultural scale also claim some questions. Here, Denmark with six on the EU scale and Poland with four on the cultural scale are the most noticeable. Most of the additional questions point to specific problems of that country. The designers thus assign the questions to those scales that the country teams think they best belong to. For example, in the case of the United Kingdom, three of these questions ($AD3$, $AD4$, and $AD5$) dealing with Islam, asylum seekers and immigration are placed in the cultural scale. The expectation here is that voters place these issues with the other cultural issues such as abortion and restriction of privacy, and not with issues on the economy or EU. Still, we have to see whether this is actually the case. For one, it would seem not controversial to argue that immigration in the United Kingdom is also very much an EU issue. If this would emerge later on, it would also be logical for question $CU1$ to be in that scale. Similarly, for Denmark, as 6 of the additional questions were used for the EU scale, it will be interesting to see whether the EU scale does span such a wide range or if these questions belong to other scales.

	EU							EC							CU							AD									
Country	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
Austria	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●				●	⋆		○			
Croatia	⋆	⋆	⋆	⋆		⋆			○	○			○		●				●	●	●	○	●	●							
Czechia	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○		○		●				●			○	●	●	⋆	⋆	⋆	○			
Denmark	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●			⋆	⋆	⋆	●	⋆		⋆	○		
Estonia	⋆	⋆	⋆	⋆					○	○		○	○						●	●	●										
Finland	⋆	⋆	⋆	⋆		⋆	⋆	○	○	○	○	○	○		●				●	●										●	
France	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●		●	●			⋆	○			●
Germany	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○	○	○		●	●	●		●	●	●				⋆						
Greece	⋆	⋆	⋆	⋆		⋆		○	○	○	○		○	○	●				●	●	●		○	○	●	●	●	⋆			
Hungary	⋆	⋆	⋆	⋆		⋆		○	○	○				○	●	●			●	●	●	⋆	⋆	⋆							
Ireland	⋆	⋆	⋆	⋆		⋆		○	○	○	○	○	○	○		●	●		●	●	●			○					○		
Italy	⋆	⋆	⋆	⋆		⋆	⋆	○	○	○		○	○		●		●		●	●	●	●	●								
Lithuania	⋆	⋆	⋆	⋆		⋆		○		○	○		○		●		●		●	●	●	⋆	○		⋆						
Poland	⋆	⋆	⋆	⋆	⋆	⋆		○	○	○	○		○		●				●	●		●	●	●	⋆			●			
Portugal	⋆	⋆		⋆		⋆		○	○	○	○	○	○	○	●	●	●		●	●	●	⋆	○	○				○		○	
Slovakia	⋆	⋆	⋆	⋆		⋆		○	○	○			○		●				●	●	●	○	●		⋆			⋆	⋆		
United Kingdom	⋆	⋆	⋆	⋆	⋆	⋆			○	○	○	○	○		●			●	●			○	⋆	●	●	●		⋆			

Table 4.5: Items included in the Final EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

To perform dynamic scale validation, we have to go through several steps. First, we use the **mokken** package (Ark 2007, 2012) to calculate Loevinger's H values for homogeneity and the crit values for monotonicity[3]. Then, we calculate Cronbach's α, the ordinal version of Cronbach's α, the ordinal version of ω, and the Latent Class Reliability Coefficient (LCRC)[4]. This shows us which scales perform as expected and which do not. For the second step, we use the automated item selection procedure (aisp) in the **mokken** package to build new scales. To do so, we first add the reverse scores of each of the questions. This is to allow for the fact that the direction of the questions can differ. As a result, each question will occur twice in the data-set. We then run the **aisp** function to look for Mokken scales with $H_i > 0.30$ using the genetic algorithm (**ga**). This algorithm has the advantage that it checks all the possible combinations, though it does so at the cost of a longer computing time (Straat, Ark, and Sijtsma 2013). Of the scales thus generated, we remove all the scales that are mere duplicates of each other and scales that only consist of two questions or less. We then check the resulting scales for monotonicity and remove any question with a **crit** value that is > 80 [5]. We do so inductively, that is running the **check.monotonicity** function after the removal of each question, as the removal of one question influences the **crit** scores of the other questions. Finally, we calculate the reliability measures for each of the scales. Appendix C shows the results the dynamic scale validation for all countries.

Table 4.5 shows the scales after dynamic scale validation. While almost all scales lost questions, the precise implications of were different for each of the scales. The EU dimension lost the least number of questions, with questions $EU1$, $EU2$, and $EU4$ remaining for all countries. This is a first sign that these questions (on the Euro, treaty change and common foreign policy) capture the underlying EU dimension well. This was not the case for question $EU5$ — related to redistribution — which DSV dropped in most cases. This also happened to question $CU4$ on community service, which DSV dropped in all countries except the United Kingdom. Meanwhile, question $CU5$ DSV retained in all cases, as well as question $CU1$ (except for Estonia and Ireland). In all cases, the DSV scales are shorter than the original ones. Estonia is the most noticeable, with 4 questions for the EU and economic scales, and 3 for the cultural scale. To get an idea why this is the case, we can look at the individual H values (see Appendix C). For Estonia, we see that the H score for the EU, economic and cultural scales are 0.20, 0.15, and 0.13 - all well below the 0.30 mark. Also, only two questions passed the 0.30 mark and most of the crit values were high. Yet, even when removing several questions, the H values for the scales were 0.33, 0.26 and 0.23, thus barely passing, or still under, the 0.30 mark. Besides, most of the H_i values for the individual questions are below 0.30, while the crit values are high in several cases as well. Most problematic is the cultural scale, with only three questions, none of which reach the 0.30 mark and two of which have crit values > 80. This means that even after DSV, we cannot view the cultural scale as representing a single underlying latent dimension. An altogether different case is Hungary, where the EU scale scored high values of H and H_i for the original scale and even higher ones

[3] For the monotonicity calculations, the minimum size of the rest score group was set to 100.

[4] I calculated the regular Cronbach's α and the LCRC using the **mokken** package, while I calculated the ordinal version of Cronbach's α and the ordinal version of ω using the **psych** package.

[5] If questions have similar **crit** values, we remove the one with the lower H_i. Note that after doing so, we have to recalculate the H_i values for all the other questions as well, as the value of H_i depends upon the questions included in the scale.

for the DSV scale. Here, DSV improved the original scale by removing the worst performing question $EU5$ and adding another.

	EU							EC							CU							AD									
Country	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10
Austria	⋆	⋆	⋆	⋆	⋆	⋆		○		○	○	○	○		⋆	●	●		⋆		●	⋆	⋆	⋆	⋆			○			
Croatia	⋆	⋆	●	⋆		⋆			○				○		●				●	●	●	○	●	●							
Czech Republic	⋆	⋆		⋆	⋆	⋆		○	○	○	○		○	⋆	⋆							○		⋆	⋆	⋆	⋆				
Denmark	⋆	⋆	○	⋆	○	⋆		○	○	○	○	○			⋆							⋆	⋆	⋆	⋆	⋆		⋆	○		
Estonia	⋆	⋆		⋆	○	⋆	⋆				○				●								●			○	●				
Finland	⋆	⋆	⋆	⋆	○	⋆	⋆	○	○	○	○	○	○		⋆	●	●		⋆		●	○		○						○	
France	⋆	⋆	○	⋆	○	⋆		○	○	○	○	○	○		○		○		⋆	⋆				⋆	○		⋆	○	○	○	○
Germany	⋆	⋆	⋆	⋆	⋆	⋆	●	○	○	○	○	○	○	●	⋆				⋆			⋆		⋆	⋆	●			●	●	
Greece	⋆		●			⋆	⋆	⋆	⋆	⋆		⋆	⋆	⋆		⋆	⋆		●	●	●	⋆	⋆	⋆	●	●	●			⋆	
Hungary	⋆	⋆	⋆	⋆		⋆		⋆		⋆				⋆	⋆				⋆	⋆		⋆	⋆	⋆							
Ireland	⋆	⋆	●	⋆	○	⋆	⋆	○	○	○	○	○	○		●					●	●	⋆	⋆	○	●						
Italy	⋆		⋆			⋆		○	○	○		○	○		⋆		●		⋆	●/⋆	●	⋆	⋆			⋆					
Lithuania	⋆					⋆	⋆	○		○			○						⋆			⋆	⋆	⋆	⋆		⋆				
Poland	⋆	⋆		⋆	○	⋆		○	○	○	○		○		⋆				⋆	⋆		⋆	⋆					○			
Portugal	⋆					⋆		⋆	⋆	⋆			⋆	⋆			⋆		●	●	●		⋆	⋆				⋆			
Slovakia	⋆		⋆	⋆		⋆		○	○	○			○		⋆				⋆	●	●	○	⋆	⋆	⋆					●	
United Kingdom	⋆	⋆	⋆	⋆	⋆	⋆			⋆	⋆	⋆	⋆	⋆		⋆				⋆			⋆	⋆	⋆	⋆	⋆		⋆		⋆	

Table 4.6: Items included in the Quasi-Inductive EUVox Scales. Questions with a ⋆ belong to EU scales, questions with a ○ to economic scales, and questions with a ● to cultural scales.

Restraining the analysis to the questions that the designers assigned to the scale in the original version has the advantage that we end up with a similar number of scales. This is important for VAAs, as we need at least two scales to construct the political map. Yet, we also saw that this restriction leads to some unsatisfactory or very short scales. Another way to improve upon the original scales is thus by removing this restriction. We call this method the quasi-inductive approach. Inductive because it works bottom-up and quasi because the scales are still constrained by the total number of questions in the VAA. Table 4.6 shows the results from this. From a first glance, we see some interesting differences. Most clear is that this approach did not maintain all scales for each country. In the case of the United Kingdom and Hungary, only a single EU scale remains, while most other countries have either two or three scales. The latter is not strange, given that the VAA focused on the European Parliament elections and thus has a focus on EU related issues. For Estonia, which had troublesome scales in both its original and DSV form, we find that while the approach found three scales, they are still both short and low in H values. We can say the same for Lithuania, which lost a cultural scale, but whose economic scale is short (containing the minimum of 3 questions). The question $CU4$, already dropped in all countries but the United Kingdom, now disappears completely from any scale. Questions $EU1$ and $EU6$ - on the Euro and EU membership, however, are included in each EU-scale, again underlining their importance.

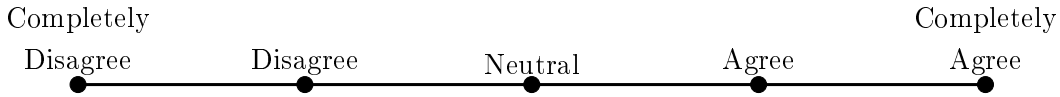
4.4 Comparing the Scales

After a first impression, I will now turn to a deeper assessment of the scales. I will do so with three different indicators. The first is *quality*, as measured by the "dirty data index". This index, introduced by Blasius and Thiessen (2012), assesses whether the users perceived the scale as ordinal or metric. The second measure is *unidimensionality*, which I will measure using Loevinger's H. The third measure is *reliability*, which we measure using the Latent Class Reliability Coefficient.

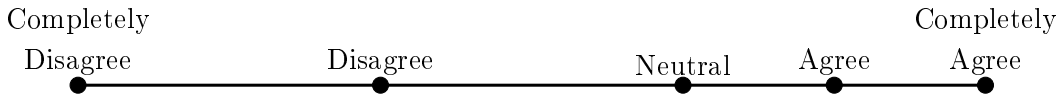
4.4.1 Quality

A scale of high quality is a scale to which users respond as intended. This means that they take careful consideration of the questions and the response options and find the proper connection between them. Yet, users often simplify by skimming over the questions or using only a few of the response categories (Blasius and Thiessen 2012, 2015). To measure quality, Blasius and Thiessen (2012) propose a measure known as the "dirty data index". This index tests whether the ordered five-point Likert scale that measures the users' position on a certain topic are able to keep their order in the latent space. This indicates if the users perceived the response categories as being metric and used the categories as such. If this is not the case, this can be a sign of satisficing. To understand the logic behind this, let us first review the differences between metric, ordinal and nominal data. If we take a question - for example, question $EU6$, "Overall, EU membership has been a bad thing for the Country X" - we can visualize the position of a user on that question by means of an underlying line:

A user can thus have *any* position on the question. They can be *slightly* in favour of the EU, or *really* against. Yet, for measurement, it is often not practical to simply let them write this down. It is easier to provide the user with pre-made categories and have them choose which one fits. Also, we order these categories so that the underlying line has a certain logic to it. In our case, we provided the user with five categories and ordered them as such:

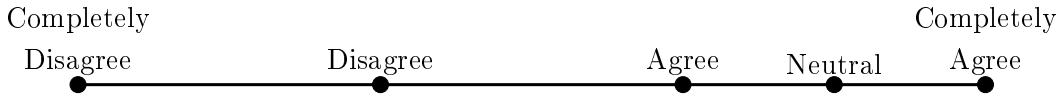


This configuration of categories runs from "Completely Disagree" to "Completely Agree" with a neutral category in the middle. We call this distribution of categories *metric*. As we can see, the categories are in the correct order *and* the distances between them are similar. This type of data is arithmetically interesting as it means that the distance between completely disagree and neutral is twice the distance between disagree and neutral. For the sake of simplicity, most scholars assume this type of metric data as it is a requirement for running a PCA. Yet, it is more common that the data are ordinal:

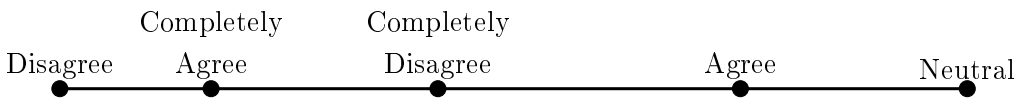


Here, we see that while the categories are still in the correct order, the distances between them vary. In this case, the distance between disagree and neutral is twice as large as the distance between neutral and agree. Besides, the neutral category is not in the centre, but more towards the agree side. This composition of categories indicates that users who are neutral on the underlying dimension (represented by the black line) will choose categories *between* disagree and neutral, and not neutral as one would expect. To deal with this type of data, we have to use Categorical PCA to get correct results.

It would be even more problematic if the categories themselves are also not in the right order. For example:



Here, we see that the neutral and agree categories are in the wrong order. Such switching of positions is also called a "tie" in the literature. Looking at the underlying latent variable, this would mean that users choosing the "neutral" category, actually agree more with the statement than those users who choose the "agree" category. While this is still a mild case with only a single category switching, in the more extreme case, both the distances are different and the order of the categories is meaningless:



This is the case when we talk about *nominal* data. In this case, users did not observe a common underlying variable but treated the categories without relation to the others. In

these cases, we cannot use either PCA of Categorical PCA but have to turn to other methods like MCA.

As we can see, while there is only one way for the data to be metric, the degree of how ordinal or nominal the data is can vary. Data can be very ordinal - as in the distances between the categories vary heavily — or very metric — the distances between the categories are nearly similar. Calculating this degree of ordinality is the basis of the DDI. To do so, the DDI looks at the data from two perspectives. The first time it looks at it as if it were metric, applies a PCA, and calculates where the categories should be. The second time, it looks at the data as if it were categorical and applies a Categorical PCA and calculates where the categories actually are on the underlying dimension. These positions are the quantification values. If the difference between the CatPCA and the PCA solution is similar (or close to 0), we can say the data is metric, while the farther the deviation is from 0, the more ordinal it becomes.

To understand how this works in practice, take a look at Figure 4.2. This figure shows a normal distribution with several quantification values indicated by vertical lines. In this case, the example is for a single question with five (Likert-style) categories. The first category holds 28.05% of the cases, and its corresponding z-value is thus -0.58. This value, however, is in a sense the "border" of the category, where the first category goes to the second. If we want to compare the PCA categories with the catPCA categories, we have to find the "centre" of the category, as this is what we will have catPCA calculate. Thus, we take the "mid-point" of the PCA category. This is where 14.025% of the cases are to the left and 14.025% of the cases are to the right, which is at the value of -1.08. The catPCA value for this category is then calculated and found to be at -1.5085. This is the *quantification* or *optimal value*. The area under the curve to the left of this value is the *quantification area* and is 0.0657. The difference between the PCA and catPCA solution for this category is then found by calculating the area between the mid-point of the PCA solution and the quantification value. We can do this by subtracting the area left from the quantification value (the quantification area which is 0.0657) from the area left from the mid-point (0.14025). This gives us $0.14025 - 0.0657 = 0.07455$, which is shown by the light shaded area. We then calculate these difference areas for each of the five categories and standardize them by an upper bound based on the number of categories minus 1. The resulting score is the DDI for that particular question.

As I will calculate the DDI for a considerable number of questions, I will use a more computational approach. Here, we start with calculating the mid-points of the question(s). To do so, we take the mass of the first category and divide it by 2:

$$g_{1k} = m_{1k}/2 \quad (\text{for } j = 1) \tag{4.1}$$

Here, g_jk is the (first) mid-point of category 1 for question k and m_1 the mass of category 1 (being the percentage of cases in question 1) for question k. The masses for each category are given by $m_j = f_{jk}/N$, where m_j is the mass for category j, f_{jk} is the percentage of cases for category j in question k, and N is the total number of cases for all the questions (which is thus the same for all the categories).

For the second mid-point, we add m_{1k} to the half the mass of the second category ($m_{2k}/2$):

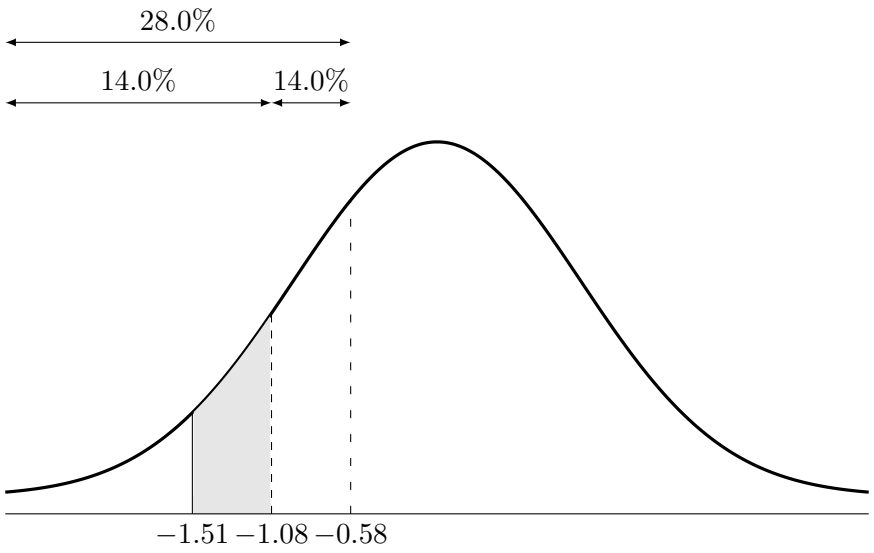


Figure 4.2: Standard Normal Distribution with Quantification value for the first category. The grey area indicates the differences between the midpoint of PCA solution and the Cat-PCA solution. Figure adapted from Blasius and Thiessen (2012, p.135).

$$g_{2k} = g_{1k} + (m_{2k}/2) \quad (\text{for } j = 1) \tag{4.2}$$

We repeat this procedure for each of the masses of the categories, so that in effect we add the first of the (J_{K-1}) masses $(c_{(1,j-1)})$ plus half the mass of the last category, with the number of thresholds being the same as the number of categories J_K. Here, $c_{(1,j)k}$ represents the cumulative masses of the categories j of question k, which is calculated by $c_{(1,j)} = m_j + c_{(1,j-1)}$ if $(j = 1, c_{(1,j-1)} = 0)$. Thus, for $j = 1$ to J_k (for each question k):

$$g_{jk} = (g_{jk} - 1) + (m_{jk}/2) \quad (\text{with } g_0 = 0) \tag{4.3}$$

Here J_K are the number of categories in each question, j the specific category, K the total number of questions and k the specific question. This gives us all the mid-points needed. Next, we calculate the area-under-the-curve to the left of the quantification value $q_j k$[6] (the quantification of category j of question k) by finding the area that corresponds to that value on the standard normal function $N(\mu, \sigma^2)$. This gives us the quantification area q_{jk} for category j of question k. We then calculate the difference areas $d_j k$ between the quantification area and the thresholds for each category j and add them to get the total of the areas of difference d_k:

$$d_k = \sum_{j=1}^{j} g_j - q_j \tag{4.4}$$

Subsequently, we standardize the value d_k by an upper bound, which is $l/(l-1)$ with l = number of categories. This gives us the DDI for a single question k. We then repeat

[6] To calculate the quantification values we used the CATPCA package provided in SPSS (Meulman, Heiser, and Inc. 2004). Other options to calculate these values are the **homals** or **gifi** packages in **R** (Leeuw and Mair 2009; Mair and Leeuw 2017).

this procedure for all the other questions k and finally add the standardized values for all the questions K and divide them by the total number of questions K to get the DDI for the total scale.

We can use the DDI not only to assess whether the users perceive the scale as ordinal or metric, but also as a proxy to how well they understood the question. The reasoning here is that when the users understood the question well, they were able to fully grasp the meaning of the categories, their placement of them on the underlying dimension of the question, and give equal weight to each of the categories. As a result, the question will be treated as metric. If they misunderstood the question, or whether there were different opinions between users of what the categories meant, the question will become more ordinal. Thus, the DDI can tell us how well the users understood the questions. The DDI itself is standardized between 0 and 1, with a value close to 0 indicating that all users understood the questions and were able to comprehend the questions and a value close to 1 indicating there were systematic and severe violations in the data, while random data produces values around 0.50. Blasius and Thiessen (2012) advise to interpret values smaller than 0.30 and 0.15 as indicating data of respectively "good" and "exceptional" quality, and values exceeding 0.50 as indicating data of "bad" quality.

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.22	0.38	0.26	0.16	0.25	0.21	0.13	0.23	0.17
Croatia	0.47	0.16	0.37	0.14	0.27	0.22	0.12	0.22	0.22
Czech Republic	0.25	0.26	0.44	0.20	0.26	0.11	0.18	0.26	–
Denmark	0.11	0.18	0.18	0.17	0.17	0.17	0.19	0.19	–
Estonia	0.36	0.40	0.52	0.14	0.19	0.60	0.07	0.33	0.39
Finland	0.13	0.20	0.21	0.12	0.17	0.22	0.10	0.17	0.21
France	0.20	0.47	0.16	0.15	0.41	0.17	0.18	0.38	–
Germany	0.26	0.33	0.16	0.29	0.17	0.17	0.29	0.19	0.31
Greece	0.29	0.27	0.12	0.28	0.29	0.17	–	0.26	0.14
Hungary	0.24	0.20	0.34	0.18	0.26	0.32	–	0.24	–
Ireland	0.30	0.39	0.27	0.25	0.26	0.22	0.20	0.28	0.24
Italy	0.32	0.17	0.26	0.13	0.19	0.15	0.13	0.17	0.12
Lithuania	0.42	0.16	0.41	0.27	0.24	0.16	0.31	0.15	–
Poland	0.49	0.37	0.33	0.44	0.37	0.18	0.42	0.23	–
Portugal	0.28	0.23	0.32	0.21	0.23	0.14	–	0.17	0.20
Slovakia	0.28	0.32	0.35	0.17	0.32	0.18	0.17	0.22	0.30
United Kingdom*	0.25	0.32	0.22	0.17	0.29	0.22	–	0.19	–

* Only includes England

Table 4.7: DDI Scores for the Original, DSV and Quasi-Inductive Scales.

I run the procedure for the DDI three times: for the data according to the original scales, the DSV scales, and the quasi-inductive scales. Table 4.7 gives the results, while Figure 4.3 visualizes them.

In the figure, squares represent the DDI values for the economic scales, circles for the EU scales, and diamonds for the cultural scales. Moreover, open symbols represent the values for

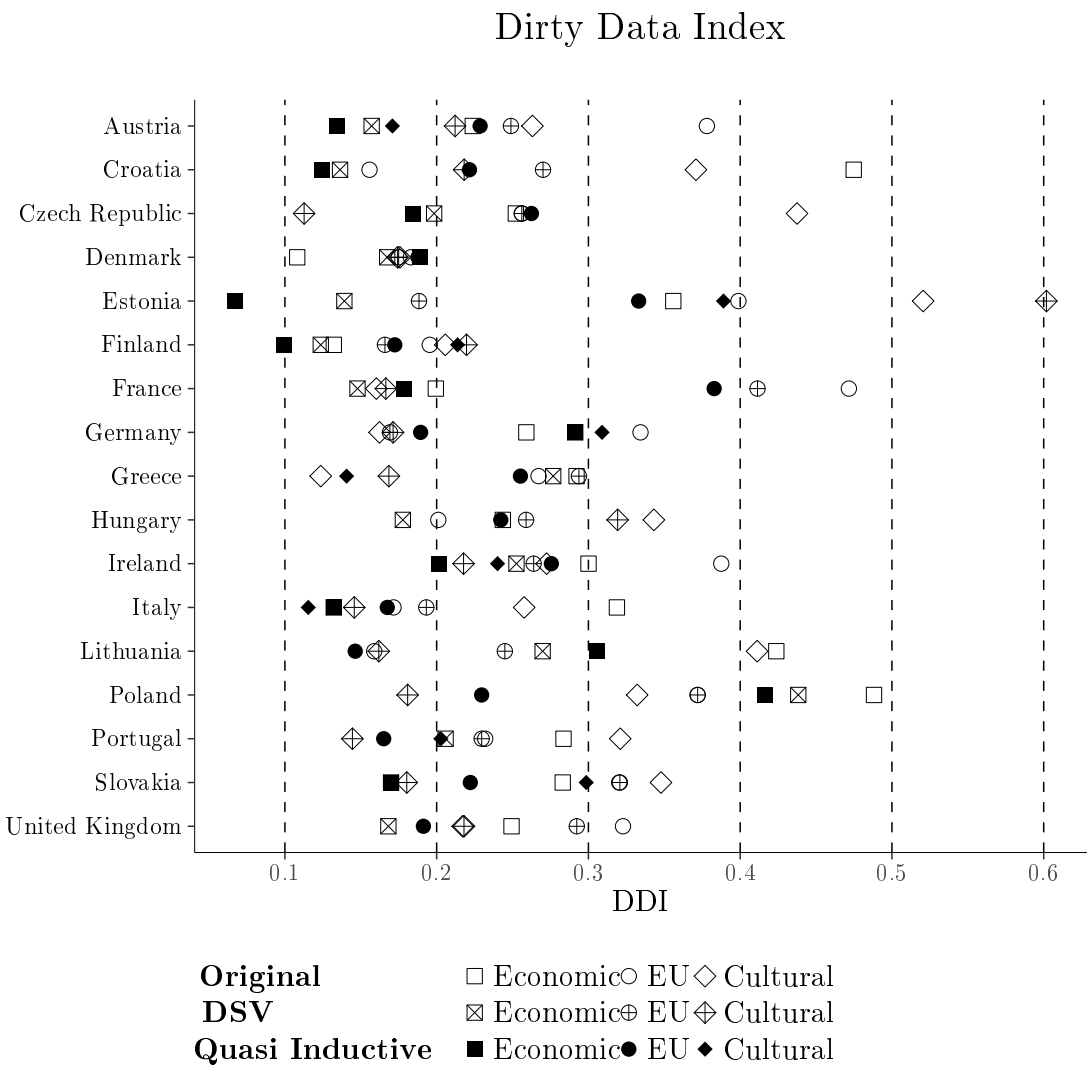


Figure 4.3: Overview of the DDI Scores for the Original, DSV and Quasi-Inductive Scales.

the original scale, closed ones for the DSV scale, and crossed ones for the quasi-inductive scale if these scales were present. We can see that most countries have DDI values between 0.10 and 0.30, indicating that the data were relatively metric. Most of the values above 0.30 belong to the original scales, and those values belonging to the DSV and quasi-inductive scales are in all cases lower than the scores for the original scales. Only two values are higher than the 0.50 mark, both belonging to Estonia. Interesting about this country is that it has both the highest and lowest values: its DSV cultural scale has a value of 0.52, while its quasi-inductive economic scale has a value of 0.07. In other words, while the first scale is ordinal, the latter can is near metric. In any case, the original values for Estonia are all problematic as they are all above the 0.30 mark, as are two of the quasi-inductive scales. This indicates that even after several procedures, these two scales remain problematic. For further purposes, it would thus be best to ignore the other scales and only use the quasi-inductive economic scale.

Two other interesting cases are Ireland and Hungary. In the case of Ireland, though the scales differ on quite some questions, the differences between the DDI scores are very small. Interestingly, the H values for the economic scale increased from 0.24 to 0.34 between the

original and quasi-inductive scales, while the DDI remained quite similar at 0.25 and 0.20. In the case of Hungary, while the scores for the cultural and economic scales improved between the original and the DSV scales, they got worse for the EU scale. A more common case is Lithuania. Here, the original economic scale scored 0.42 but was improved upon in both its DSV and quasi-inductive version (0.27 and 0.31). The same occurred, but then even stronger, for its cultural and EU scales. Some countries score consistently higher than others. This is especially the case for Denmark and Finland, where most of the values were between 0.1 and 0.2. Other countries were more spread out. One such case is France, where most of the scales perform well, except the one dealing with the EU. This scale starts at 0.47 and in its DSV version "improves" to 0.41, with even the quasi-inductive version only reaching 0.38. In the case of the Czech Republic, the original cultural scale was an outlier at 0.44 but improved to 0.11. Yet, the scale did not emerge from the quasi-inductive method.

What can we conclude from this? To begin with is that the quality of the scales differs not only per country but also per scale. Moreover, the DDI scores do not always improve between the scales. In other words, the scalability of the scale may be better, but the understanding of the scale did not improve. To get a better idea of why this is, we will now turn to the Loevinger's H values for each of the scales.

4.4.2 Unidimensionality

Scales allow us to measure in an indirect way that which we cannot measure in a direct way. Thus, while we cannot measure the "leftness" of someone in a direct way, we can measure their position on various topics related to it. If we then add together all the responses they gave in a scale, we can arrive at a good indicator of their "leftness". How good this indicator is, depends to a large degree on whether the scale measures what it should measure. In other words, does the scale indeed measure "leftness" or does it also measure anything else? If so, we call such a scale multidimensional. Multidimensional scales are problematic as they allow for more than a single interpretation. Imagine, for example, that our scale not only measures "leftness", but also "progressiveness". What then, does a single unit of such a scale mean? Does it mean that someone gets more left and also more progressive? Or only more left? And how do we interpret the scale in the first place?

Because of this, most social scientists and VAA designers strive to work with unidimensional scales. These scales measure a single dimension and only this dimension (Gerbing and Anderson 1988). To arrive at unidimensional scales, designers most often depend on theory. Based on experience, they assign questions to those scales they think they belong to. Yet, as we saw with the original scales earlier, this is no guarantee for success. Thus, designers have turned to various ways to assess the dimensionality of a scale. Most popular are EFA and CFA, Structural Equation Modelling (SEM), and MSA. Of these methods, both EFA and CFA are the most well-known. The key difference between them is that while EFA is dependent on the data, CFA bases itself on previous research. Thus, designers can use CFA to test whether their scale is unidimensional, and they can use EFA to find new unidimensional scales. Yet, while popular and easy in use, these methods have some problems, especially in relation to data based on VAAs. The main problems with EFA are it does not allow for correlated errors, that it depends on arbitrary eigenvalues, and does not allow for estima-

tions of model fit (Bollen 1989). CFA then has as its main problem that we need to know which dimensions there are. Also, both methods deal with numeric data. Yet, the data from ordered categorical survey questions like Likert questions are categorical. Thus, methods of factor analysis often provide more dimensions than there are, with some dimensions occurring because of response patterns (Eijk and Rose 2015). One solution for this is to use SEM. This is as SEM can use categorical data and various authors have applied it with success as such (Bostic, McGartland Rubio, and Hood 2000; Yu and Hsu 2013; Yu, Lin, and Hsu 2013). Another solution is to use MSA. I already discussed MSA earlier when talking about the DSV. MSA has as its main advantage that it cannot only help us to estimate the unidimensionality but also allow us to construct new scales with ease. Thus, for the remainder, I will focus on MSA and go a bit deeper in discussing what it does and can do.

MSA uses two concepts to assess unidimensionality: homogeneity and monotonicity. To measure homogeneity, Mokken (1971) proposed to use Loevinger's H. This coefficient comes in three types: a) a value H which shows how accurate the scale can order the users on the underlying dimension (Mokken, Lewis, and Sijtsma 1986), b) a value H_ij, which indicates how well questions i and j co-vary (Loevinger 1947, 1948), c) a value H_i that tells us how well a question covaries with the other questions in the scale. For scale construction, H and H_i are what interest us. Values of both are restricted between 0 and 1 (Hemker, Sijtsma, and Molenaar 1995), with 0 indicating there is no relation between the questions at all and 1 that the questions are related and ordered. For unidimensionality, scales with $H < 0.3$ lack unidimensionality, while scales between $H = 0.3$ and $H = 0.4$ are weak, scales with $0.4 < H < 0.5$ are medium and scales with $H > 0.5$ are strong. For a Mokken scale, the H of the scale needs to be at least 0.3, as do the individual H_i values for each of the questions. To measure monotonicity, we make use of the *restscore*. This is the score that remains when we subtract the score for the question we look at from the sum score of all questions. If monotonicity holds, this means that the higher the rest score of the user, the more likely it is that the user obtained a higher score on a question. To measure this with more precision, Sijtsma and Molenaar (2002) developed the *crit*-value which takes into account the violations when monotonicity does not hold as well as several other values of the scale to generate a statistic that allows for easy interpretation. Here, I take 80 as a cut-off point above which I deem a question to violate monotonicity[7]. Using these two measures, we can use MSA to make unidimensional scales using either the standard *automated question search procedure* (aisp) or a *genetic algorithm* (ga). The aisp starts with those two questions that have the highest H_ij value. Then, it adds the question with the highest H_ij value relative to the original two questions until H drops below a criterion (often $H = 0.3$). The procedure is then repeated for the remaining questions until aisp can find no more scales (Ark 2012). While the advantage of this method is it is quick and simple to execute, it does not lead to an optimal partition of the questions into scales. The main reason for this is that as soon as a question is only a little under the lower bound for the criterion, aisp does not consider it. Moreover, as the method works hierarchical, much depends on which questions aisp chooses as its starting questions. To address this, Straat, Ark, and Sijtsma (2013) developed the genetic algorithm. This algorithm, instead of working hierarchical, considers all possible

[7] I calculated both the crit values and values of Loevinger's H using the **mokken** package in R (Ark 2007, 2012).

combinations of scales and selects those scales that are the longest. While such scales are often preferred as they cover more of the questions, the genetic algorithm has a drawback that it becomes time-consuming when the number of questions increases (Straat, Ark, and Sijtsma 2013).

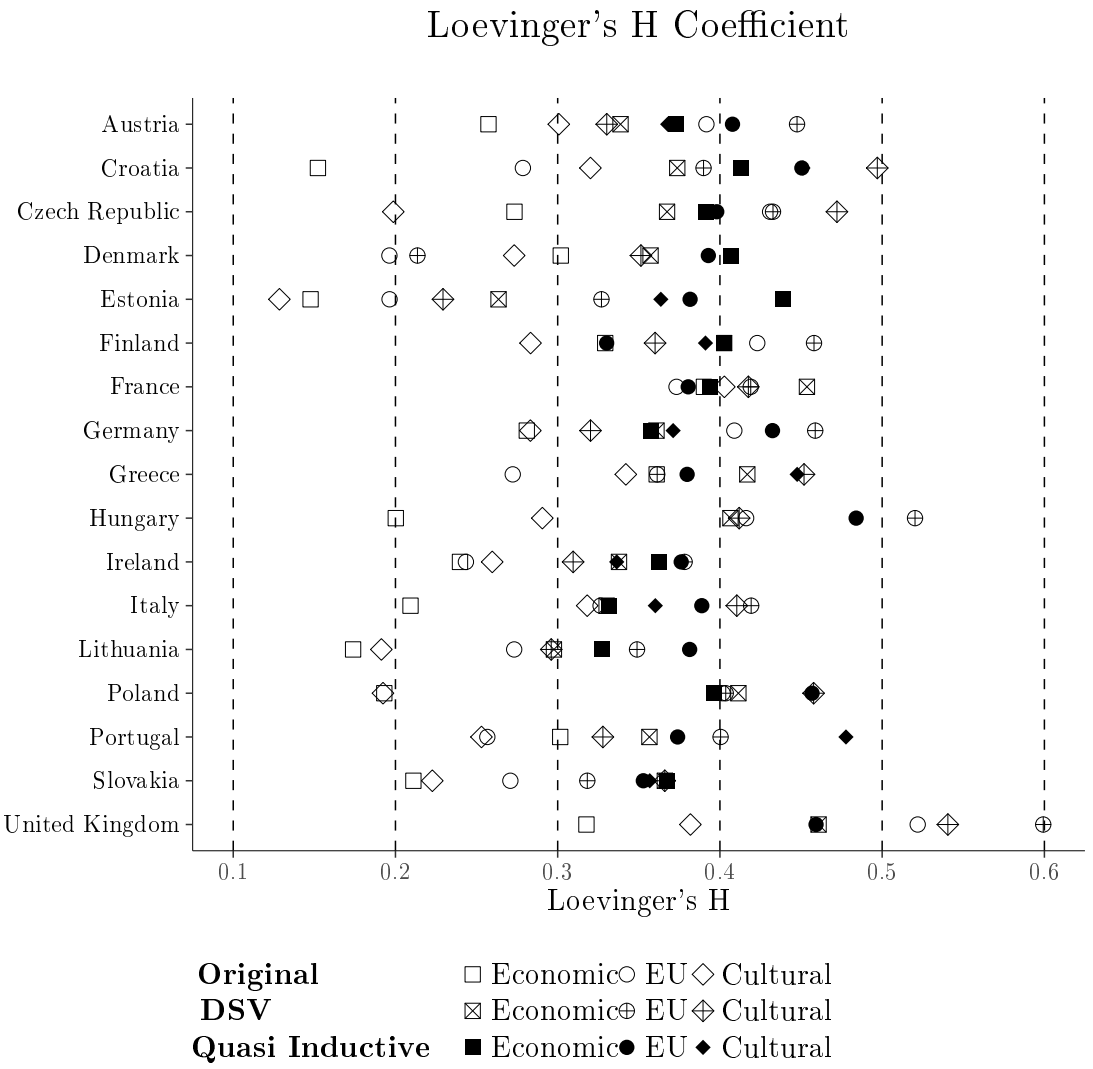


Figure 4.4: Overview of the values for Loevinger's H Coefficient for the Original, DSV and Quasi-Inductive Scales.

Let us now return to the scales. Table 4.8 shows the values for the scales. Figure 4.4 visualizes them. Note that, opposite to the DDI, we want our values of H to be as high as possible. The first thing we note is that most values for the DSV and quasi-inductive scales in are above 0.3. The highest values are in the United Kingdom, where the EU scale in its DSV form reached 0.60. In its quasi-inductive form, a new EU scale emerged, which included also the other two scales and reached a position of 0.46. Note that this value is obscured in the plot by a similar position of the EU DSV scale position. In the same way, in Hungary, the original economic scale improved from 0.20 to 0.52 but was also subsumed by an EU scale in its quasi-inductive form. The degree to which the value of H increases is different for each country and each scale. In the case of Poland, both the EC and CU original scales had

a value of 0.19 and increased to 0.41 and 0.46 after the removal of several questions. In the case of the EC scale, these comprise the questions relating to cutting government spending (EC5) and loans from external institutions (EC7) as well as a question about the protection of the environment versus economic growth (AD5). Each of the questions correlated lowly with the other questions, with the loans from external institutions question even correlating negatively with the other ($H_i = -0.204$). This means that for that question we observe more errors than expected under statistical independence. This is most likely caused because the question is not in the correct direction, meaning users who agreed with the other questions of the scale disagreed with this question. Other countries in which the H improved between the original and DSV version of the scale was Croatia, where the EC scale increased from 0.15 to 0.37 and the Czech Republic, where the CU scale increased from 0.20 to 0.47. In other countries, like Ireland, Portugal and Denmark, the improvement (in this case for the CU scale) was low and helped little to reach the 0.30 mark. Overall, the scales for EC and CU improved the most between the original and DSV scales, while the EU scale showed less improvement.

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.26	0.39	0.30	0.34	0.45	0.33	0.37	0.41	0.37
Croatia	0.15	0.28	0.32	0.37	0.39	0.50	0.41	0.45	0.45
Czech Republic	0.27	0.43	0.20	0.37	0.43	0.47	0.39	0.40	–
Denmark	0.30	0.20	0.27	0.36	0.21	0.35	0.41	0.39	–
Estonia	0.15	0.20	0.13	0.26	0.33	0.23	0.44	0.38	0.36
Finland	0.33	0.42	0.28	0.40	0.46	0.36	0.40	0.33	0.39
France	0.39	0.37	0.40	0.45	0.42	0.42	0.39	0.38	–
Germany	0.28	0.41	0.28	0.36	0.46	0.32	0.36	0.43	0.37
Greece	0.36	0.27	0.34	0.42	0.36	0.45	–	0.38	0.45
Hungary	0.20	0.42	0.29	0.41	0.52	0.41	–	0.48	–
Ireland	0.24	0.24	0.26	0.34	0.38	0.31	0.36	0.38	0.34
Italy	0.21	0.33	0.32	0.33	0.42	0.41	0.33	0.39	0.36
Lithuania	0.17	0.27	0.19	0.30	0.35	0.30	0.33	0.38	–
Poland	0.19	0.40	0.19	0.41	0.40	0.46	0.40	0.46	–
Portugal	0.30	0.26	0.25	0.36	0.40	0.33	–	0.37	0.48
Slovakia	0.21	0.27	0.22	0.37	0.32	0.37	0.37	0.35	0.36
United Kingdom*	0.32	0.52	0.38	0.46	0.60	0.54	–	0.46	–

* Only includes England

Table 4.8: H Values based on the original and DSV scales for EUVox

The lowest scores are for Estonia and Denmark. Despite changing the scales, even the DSV scales for the EU and cultural dimensions in Estonia were well below the 0.3 mark, and the value for the economic scale just above it at 0.33. The quasi-inductive scales fare better but are small. Both the economic and cultural scales consist of the minimum of 3 questions, while the EU scale has 5 questions. To better understand why this is the case, we can have a look at how the scales are actually made up. Table 4.9 shows this for the cultural scale.

As is clear, for both of the original and DSV scales, any of the individual H values were

Question	Original H_i	DSV H_i	Quasi-Inductive H_i
CU1	0.16	–	0.34
CU2	0.13	–	–
CU3	0.17	–	–
CU4	0.10	–	–
CU5	0.15	0.26	–
CU6	0.08	0.24	–
CU7	0.15	0.19	–
AD2	–	–	0.39
AD6	0.10	–	0.36
H	0.13	0.23	0.36
DDI	0.52	0.60	0.39
LCRC	0.54	0.45	0.55

Table 4.9: Coefficients for the Questions on the Cultural Scale in Estonia. Question AD2 reads *The EU should impose economic sanctions on Russia, even if this jeopardizes gas supplies to EU countries*, while Question AD6 reads *Obtaining Estonian citizenship is unfairly difficult for certain groups in society, such as the elderly.*

higher than 0.30. As such, there was little that held these questions together as a scale. Using only three questions improves the situation somewhat, but it is only when we allow the algorithm to include any question possible that we find a sufficient scale. Still, even here the H values are not particularly high. This is because the cultural dimension has three issues dealing with citizenship and immigration, which are often problematic topics. Also, while the scalability increases between the DSV and the quasi-inductive version, the DDI decreases. So, while the questions form a stronger scale, users had more difficulty understanding what the scale was about. In the case of Denmark, something interesting happened in that the score for the DSV economic scale is lower than the original score. This is most likely because of different cleaning procedures between the data that were used to establish the DSV scales and the data used here. Another point of interest is the differences between the scales. In the case of France, each of the scales scores around 0.4. In other words, each of the scales was equally strong. In most other countries, we can draw a similar conclusion, though the distances vary more.

4.4.3 Reliability

We finish by taking a look at the reliability. Remember that reliable means that the scale is precise in what it measures. In other words, the true score of the user is responsible for as large a degree of the total variation as possible (Carmines and Zeller 1979). As we cannot observe the true score, we have to estimate it. To do so, the most established technique is that of internal consistency. Here we measure this with the Latent Class Reliability Coefficient (LCRC). We can interpret the LCRC analogous to Cronbach's α. Thus, it scales between $0 - 1$, and we can consider 0.9 to be a lower bound (Sijtsma 2009).

Figure 4.5 visualizes the values for the LCRC. Table 4.10 shows them. Most of the values lie between 0.50 and 0.90, with only two countries (Hungary and the United Kingdom) showing values above 0.90, in both cases for the EU scale. This means that for most cases

	Original			DSV			Quasi-Inductive		
Country	EC	EU	CU	EC	EU	CU	EC	EU	CU
Austria	0.74	0.83	0.77	0.76	0.83	0.77	0.75	0.89	0.58
Croatia	0.67	0.70	0.82	0.67	0.74	0.85	0.72	0.85	0.67
Czech Republic	0.78	0.87	0.68	0.79	0.87	0.74	0.77	0.88	–
Denmark	0.76	0.81	0.73	0.80	0.81	0.72	0.80	0.89	–
Estonia	0.63	0.65	0.54	0.56	0.60	0.50	0.67	0.69	0.55
Finland	0.77	0.80	0.72	0.78	0.82	0.65	0.83	0.83	0.60
France	0.84	0.82	0.85	0.85	0.83	0.85	0.90	0.83	–
Germany	0.73	0.85	0.71	0.75	0.85	0.73	0.75	0.88	0.74
Greece	0.85	0.74	0.84	0.84	0.76	0.84	0.84	0.89	–
Hungary	0.67	0.86	0.76	0.72	0.89	0.76	–	0.92	–
Ireland	0.78	0.70	0.73	0.80	0.71	0.71	0.78	0.81	0.71
Italy	0.73	0.76	0.81	0.69	0.80	0.83	0.68	0.85	0.61
Lithuania	0.63	0.73	0.63	0.63	0.77	0.66	0.54	0.82	–
Poland	0.69	0.81	0.77	0.75	0.81	0.85	0.81	0.89	–
Portugal	0.85	0.72	0.72	0.85	0.75	0.73	–	0.85	0.69
Slovakia	0.69	0.75	0.72	0.72	0.77	0.76	0.71	0.81	0.60
United Kingdom*	0.83	0.91	0.85	0.82	0.92	0.84	–	0.94	–

* Only includes England

Table 4.10: LCRC Values based on the original and DSV scales for EUVox

the standard of 0.90 is not reached. Moreover, in a considerable number of cases, the value of the LCRC is well below the 0.90 mark, with low values occurring for each of the types of scales. The lowest values are once more found in Estonia, where the DSV cultural scale scored only 0.50, and none of the other scales reached higher than 0.70. Combined with the low values of H for the country, we conclude it is not possible to generate valid scales for Estonia. We can find another case of a low LCRC can for the EC scale for Lithuania in its quasi-inductive form with a score of 0.54. Given that the H score of this scale is 0.33, we can also question the usefulness of this scale. We find the best results, as with the values for H, in the United Kingdom. Here, all the scales scored higher than 0.80, with the quasi-inductive scale on the EU reaching 0.94. The EU scale scores as well for the other countries in the case of the quasi-inductive scales. Here it has an average of 0.85 over all countries, while the EC and CU scale scored 0.75 and 0.64. For the original and DSV scales, the average differences were less pronounced. The original scales had averages of 0.74, 0.78 and 0.74 for the EC, EU and CU scales, while the values for the DSV scales were 0.75, 0.80 and 0.75. The quasi-inductive scales, while performing very well for the EU scale, perform the same as the DSV scales in case of the EC scale, but worse in the case of the CU scale. This is because the cultural scales in their quasi-inductive form often were small. In the case of Austria, Estonia, Finland, France, Italy, Portugal, and Slovakia they contained only three questions. Keeping the analogy with α, it is likely that the LCRC decreased as a higher number of questions leads to higher reliability of the scale (Nunnally 1967).

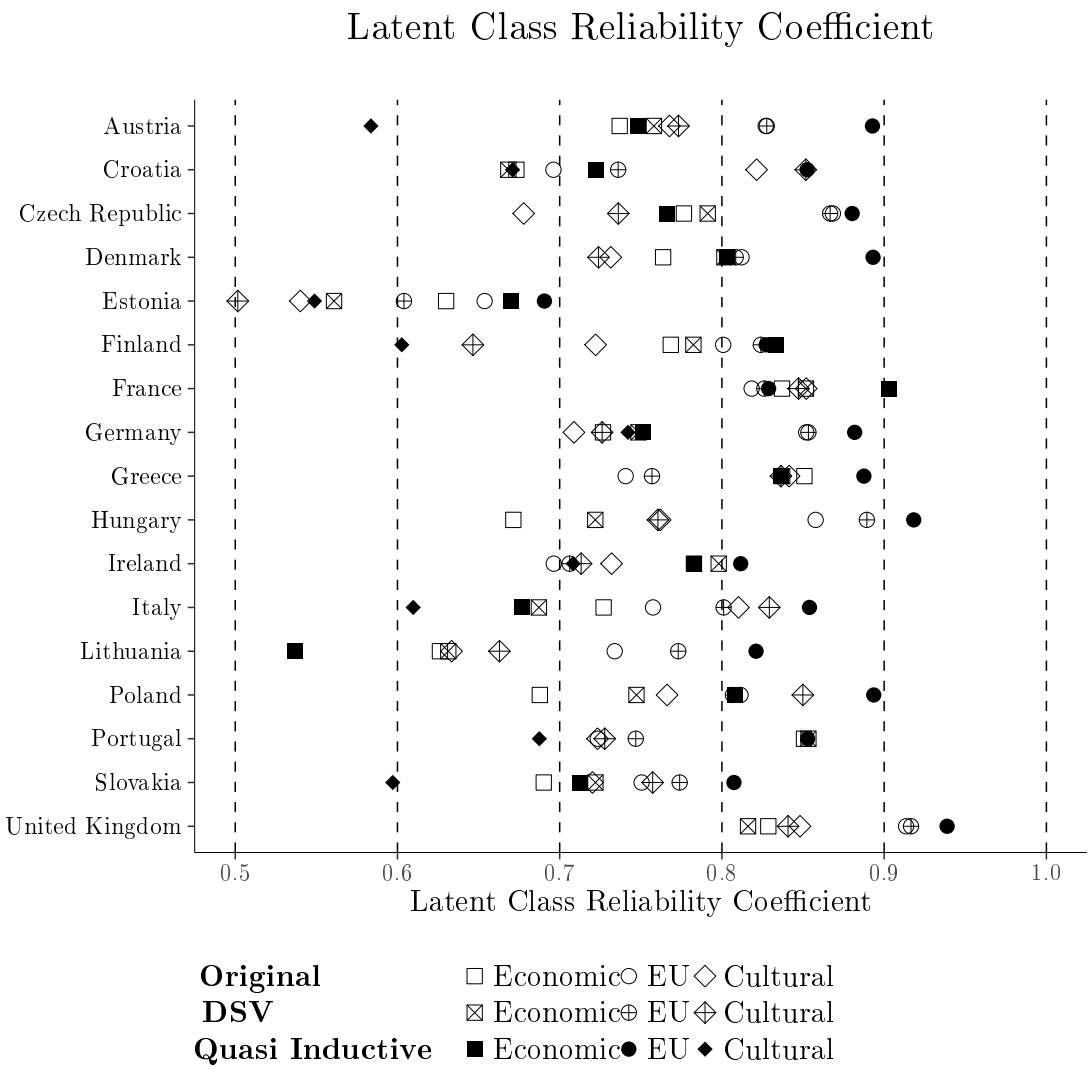


Figure 4.5: Overview of the Latent Class Reliability Coefficient for the Original, DSV and Quasi-Inductive Scales.

4.4.4 Relationships between the Measures

As I noted earlier, there is a relationship between Loevinger's H and the LCRC. This is as the aim of both is to measure the inter-item co-variances between the questions. They either do so in a direct way (Loevinger's H) or by means of an estimator (LCRC). This is because the LCRC sees reliability as internal consistency. This measures the degree to which each of the questions in the scale measures the same underlying variable. If all questions do, the scales are reliable. This is also what Loevinger's H does: finding those questions that co-vary to find an underlying dimension. It is thus interesting to look whether this relationship is also present here. To do so, we calculate the Pearson's r correlations over all the countries for each of the scales in each of their three iterations. Figure 4.6 visualizes these correlations together with their 95% confidence intervals. Figure 4.7 shows the dot-plots of the correlations.

Starting with the original scales, we find high and significant correlations between the LCRC and H values as we would expect. Besides, the confidence intervals are small. For the

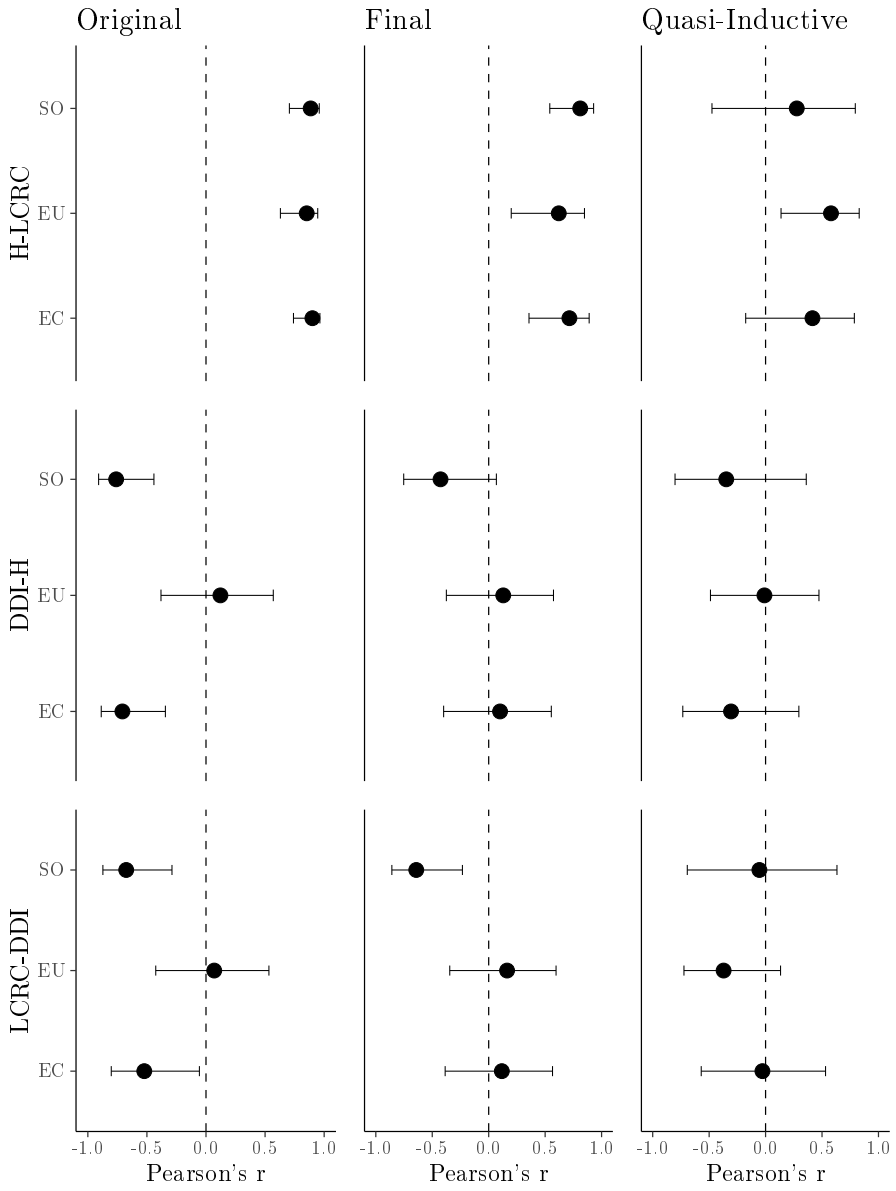


Figure 4.6: Pearson's r correlations between the DDI, Loevinger's H, and the Latent Class Reliability Coefficient with 95% confidence intervals

correlations between the DDI and H values, we find conflicting results. While the SO and EC scales are significant but negative, the EU scale is not significant. We can make a similar evaluation for the correlation between the LCRC and the DDI, though here the confidence intervals of the SO and EC scales are more extended. This means that for the economic and cultural scales an increase in the LCRC or H value is related to a decrease in the DDI. This is what we would want: increased reliability goes together with an increased understanding of the scale. Yet, in the case of the EU, the two do not seem to be related.

For the DSV scales, the relations between the LCRC and H values are still all positive, though the degree of the correlation is lower and the confidence intervals are wider. For the relationship between the DDI and H, none of the values is significant, with both the coefficient for the EU and EC scale being very close to 0. For the LCRC-DDI relation, both

the EU and EC scale are not significant, with only the SO scale indicating a significant negative correlation, like the one seen at the original scales.

For the quasi-inductive scales, none of the correlations is significant, except for the EU scales at the relation between H-LCRC. Here, the other two scales, while positive, are not significantly different from 0 and have rather large confidence intervals. For the relation between the DDI and H, the correlation for the EU scale is 0 ($r = -0.01$) while the other two scales are negative but also not significantly different from 0. This is additionally the case for the relationship between the LCRC and the DDI, with both the SO and EC scales being 0 (with $r = -0.03$ and $r = -0.05$), and the EU scale being negative but insignificantly different from 0.

This brief comparison shows that the relationship between the different measures is not so clear-cut as it would seem. As the quasi-inductive scale shows, there does not even have to be a relationship between the values of H and the LCRC of the scale, even though there is a connection between the two measures. Moreover, there seems to be no correlation between the DDI and both of the other measures. It even has a negative correlation to the values of H at the original scales. This means that designing a scale to be optimal under one measure does not guarantee that it will also be optimal under any of the other measures. It is thus well possible to design a scale with high reliability as indicated by the LCRC, but a low value for H or the DDI. Which of the measures to prefer is than up to the purposes and intents of the designer.

4.5 Other Indicators

Until now, I assessed the scales as generated by EUVox on three different measures. This begs the question if there are more measures with which we could assess our scales. Here, I will focus on one, given that it seems most relevant with regard to VAAs. This is the idea of comparability. Comparability means that the scale would function the same and thus mean the same to different users in different environments. Inadequate comparability between surveys is often a problem in large-scale surveys done in many countries. This is because users in different cultures might have different ideas on what certain concepts mean. In a simple example, the idea of what "left" means in a political left-right scale, might be different between Poland and Spain. Thus, assuming they mean the same may lead to incorrect conclusions.

While I will not explore the consequences of comparability here, comparability itself is an issue for VAA-related data. This is especially so when used in an international context. VAAs such as *euandi* (Trechsel, Garzia, and De Sio 2015, 2014), *EU Profiler* (Trechsel and Mair 2011) and *EUVox* (Mendez and Manavopoulos 2018) were all designed for the elections for the European Parliament and each had common statements that appeared in each of the country's questionnaires. Yet, though the designers did translate the statements and sometimes edited them to better fit into a country's context, this is by no means a guarantee that the statement means the same to the user. While this is not a problem for each VAA on its own, it does become problematic when comparing the data of the VAAs against each other and assuming the statements mean the same. This might lead to erroneous ideas of the true position of a country's users on a certain statement. To address this issue, scholar shave

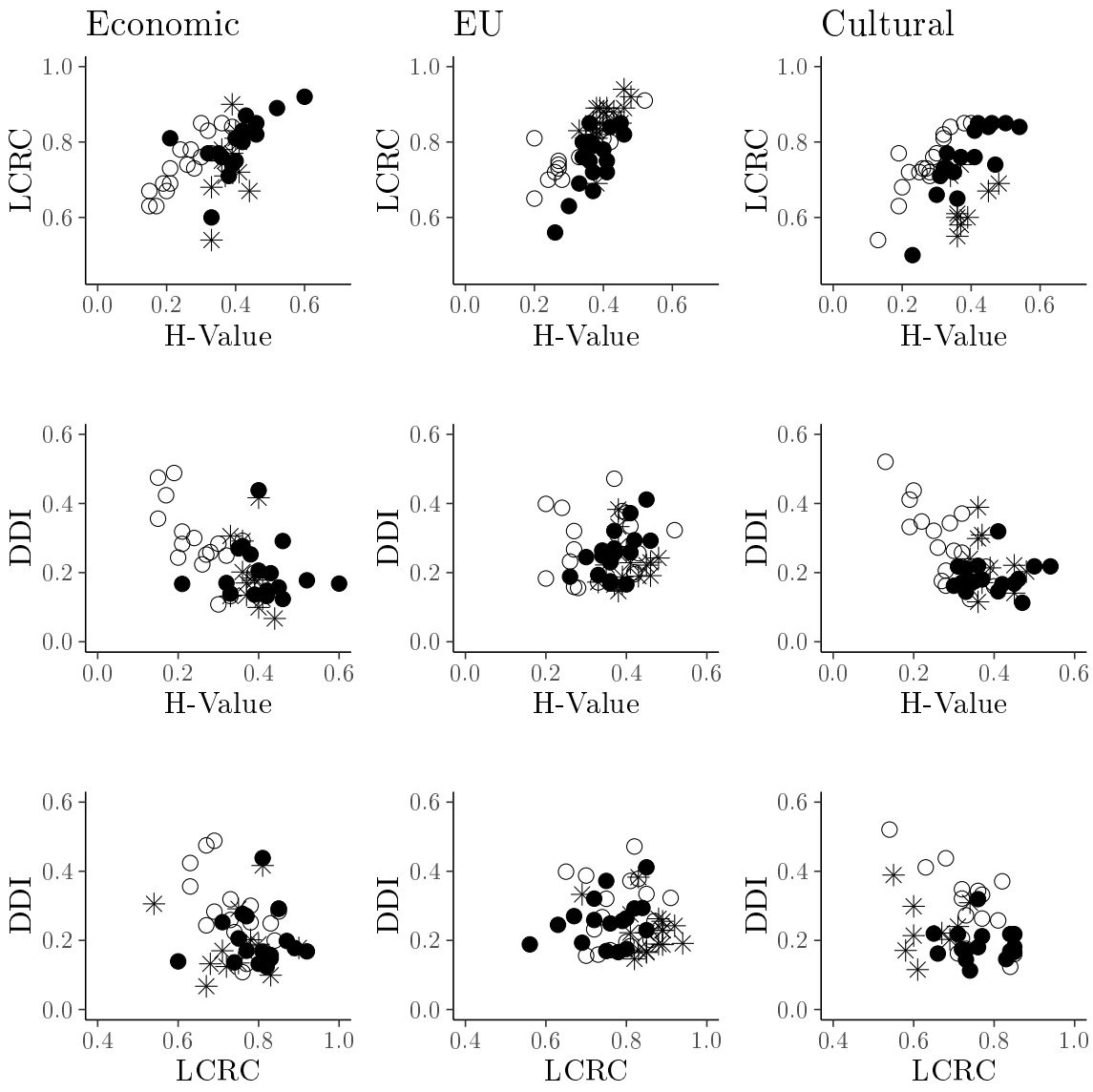


Figure 4.7: Correlation between DDI scores, Loevinger's H scores, and the Latent Class Reliability Coefficient, with the white dots show the *original* values, the black dots the *Quasi-Inductive* values, and the stars the *DSV* values. Note the different scales of the axes for each of the rows.

proposed several techniques. The most popular are: a) adapting the design, b) applying statistical corrections and c) exploit various techniques of statistical modelling (He et al. 2017).

Scholars adapting the design often opt to include *vignettes* in their surveys (King, Murray, et al. 2004). These vignettes are short stories meant to measure the degree of comparability by measuring the opinions of the user on several hypothetical stories. As the stories are the same for each user, in theory, the responses of the users should be the same. Designers can then use the degree of variability between users to correct the scale by means of re-coding or by including the information of the vignettes into the statistical model used to analyse the data. To work, vignettes need vignette equivalence and response consistency.

The former means that all users interpret and respond to the vignette in the same way, while the latter implies they respond in a similar way to the vignette as to the other questions in the survey. Even so, these assumptions do not always hold and often depend on background characteristics like age and sex (Jürges and Winter 2013). Thus, to use vignettes, scholars should validate the vignettes themselves first and the survey results and should correct for background characteristics if necessary.

Scholars can make statistical corrections to the scores. This often involves either score standardization or parcelling. Standardization involves using the mean and standard deviation of the scores to control for differences in the use of the scale. Parcelling combines individual questions into combined questions, whose score is then used. Scholars used both solutions with conflicting results, often because both methods can obscure genuine differences in the survey. Finally, techniques of statistical modelling, which often involve treating Likert-scale responses as categorical instead of continuous, are still rare and are not widely used. He et al. (2017) suggest to use categorical Multigroup Confirmatory Factor Analysis (MGCFA) and find they work better than their continuous counterparts, given that model identification is not a problem.

4.6 Conclusion

So, what is the quality of the scales Voting Advice Applications use? The short answer is that it depends. The long answer is that only a few of the scales I looked at here hold up to the standards of unidimensionality, reliability and quality. Also, I find that unidimensionality, reliability, and quality do not always improve between the original and the DSV scales or the DSV and the quasi-inductive scales. This, despite the fact that the idea of both the quasi-inductive and DSV scales is to improve on the original scales. The only measure that increases from the original to the DSV and the DSV to the quasi-inductive version of the scale is Loevinger's H. And this is because in the original scales the H values were unimportant, so their value could be below 0.30, while in the DSV and quasi-inductive scales these values were always above 0.30. The LCRC also increases in most cases, but not always. But because the LCRC is a new measure, it is premature to say whether this is because of an actual increase in the reliability of the scale or because the number of questions increased the LCRC as is the case with Cronbach's α. The DDI then showed the clearest improvement over the scales, with the original scales showing the worst performance except in some cases like Estonia. Moreover, scales with a high LCRC often also have a higher Loevinger's H coefficient for the whole scale. This is not unexpected as there is an established theoretical link between reliability and unidimensionality. Yet, this link differs and could most is most clear for the original and DSV scales. Also, there is no clear relationship between the DDI and either the LCRC or Loevinger's H. This is not surprising, as the DDI focuses on the relationship of the response categories in the question, while the reliability and the unidimensionality focus on the relationship between the questions.

What does this mean for VAA designers? To begin with, that the validation and construction of scales is very much a craft. There is not one correct way to build scales and ensure they score their unidimensionality, reliability and quality. As we have seen, unidimensional scales can still have low reliability or low quality. So, the construction of scales

is an iterative process in which the designer has to decide which aspects of the scale are the most important. So, including more questions can increase reliability, but also decrease unidimensionality. This is because the extra questions might correlate low with the other questions. Whether to prefer unidimensionality to reliability or the other way around is up to the designer. Besides, designers should decide how many scales they want to produce. For most political maps, at least two dimensions are necessary. This might be a problem when, as in the United Kingdom and Hungary, there is only a single dimension remaining after MSA. In that case, it might be necessary to overlook the higher values of LCRC and Loevinger's H for the new scale and settle with fewer optimal scales.

Something else it means to VAA designers is that they have more than one way to validate their scales. The point is what they want to achieve and which method they want to use. They can use the LCRC and Loevinger's H during the operation of the VAA (as in dynamic scale validation) to focus on the scale as a whole. They can use the DDI after they have collected the data and want to focus on the questions. They can use it for two purposes. First, when they want to check if their data is suitable for analysis. This is the kind of use that Blasius and Thiessen (2012) and Blasius, Nenadić, and Thiessen (2017) propose. Second, when they want to see if the users did understand the questions as intended. This can help them to identify difficult questions which they can then reframe for next time.

While I looked in this chapter what the quality of the scales was, I stopped short of discussing why it is what it is. In other words, how can we explain the quality of these scales? Why is the DDI is so high for some countries and scales, and so low for others? Why do some countries have scales with such low H values whilst others were high? To address these questions, I will apply MCA and CatPCA techniques to four of the countries we looked at here. These countries — Estonia, Lithuania, Ireland and Hungary — all showed particular features of interest. Thus, these countries will take centre stage in the next chapter.

5 | The Structure of Scales

In the previous chapter, I focused on the first part of the first research question: what is the quality of the scales Voting Advice Applications use? There, we saw that there are often considerable differences between the quality of the scales. In this chapter, I thus focus on the second part of the first research question: how can we explain the quality of these scales? I will do so by looking at five countries that exhibited interesting patterns in the DDI, LCRC and H values.

To begin with, I look at the economic scale in Lithuania, which behaved very much as expected. The original scale had high DDI scores and low H and LCRC scores, while in the DSV scale the DDI decreased and the LCRC and H values increased. Also, I will use Lithuania to show how we can use other variables to look at otherwise invisible patterns in the data-set. Then, I will look at the EU dimension in Ireland. Here, I focus on how the EU dimension - which is the most important one in EUVox, changed between the original and DSV version. After that, I will turn to the deviant cases. To begin with, I look at the EU dimension in Hungary. This dimension scored high H and LCRC values and low DDI values in its original version and even better ones its DSV version. Moreover, in its quasi-inductive version, only the EU dimension remained. I will look here what makes this dimension so strong. Next, I turn to Estonia, which was the worst performer on all measures in the previous chapter. I will look at its cultural scale, which scored a DDI value of 0.60 in its DSV version and should thus showcase fundamental problems. Finally, I look at the opposite case for the United Kingdom and look at its especially strong EU dimension. I end with a series of recommendations on how a valid scale in a VAA should look like and which tools designers can use to achieve this.

5.1 Lithuania

Before I use the catPCA and MCA methods, I first look at Lithuania using traditional PCA methods. The purpose of this is to show the problems that might occur when assuming the data is metric, while the DDI warned us that this was far from the case. Table 5.1 shows the rotated[1] and unrotated solution for the first and second dimensions (or components)[2]. Together, the first two dimensions account for 25.59%+12.99% = 38.68%, with eigenvalues of 2.30 and 1.17. The third and fourth dimensions account for another 21.39% of the variation ($\lambda_3 = 1.01$, $\lambda_4 = 0.91$). By the Kaiser eigenvalue criterion, the solution is thus three-dimensional (as three of the eigenvalues are above the value of 1). Yet, the difference between

[1] All PCA rotations in this chapter use Varimax rotation.

[2] In this table and all those that follow, an $*$ means that the question has been reversed.

the first and the second eigenvalues is large compared to the differences between the second, third and fourth eigenvalues. By this definition, the solution is one-dimensional, though it would mean that the main dimension only explains 25.59% of the variation in the data.

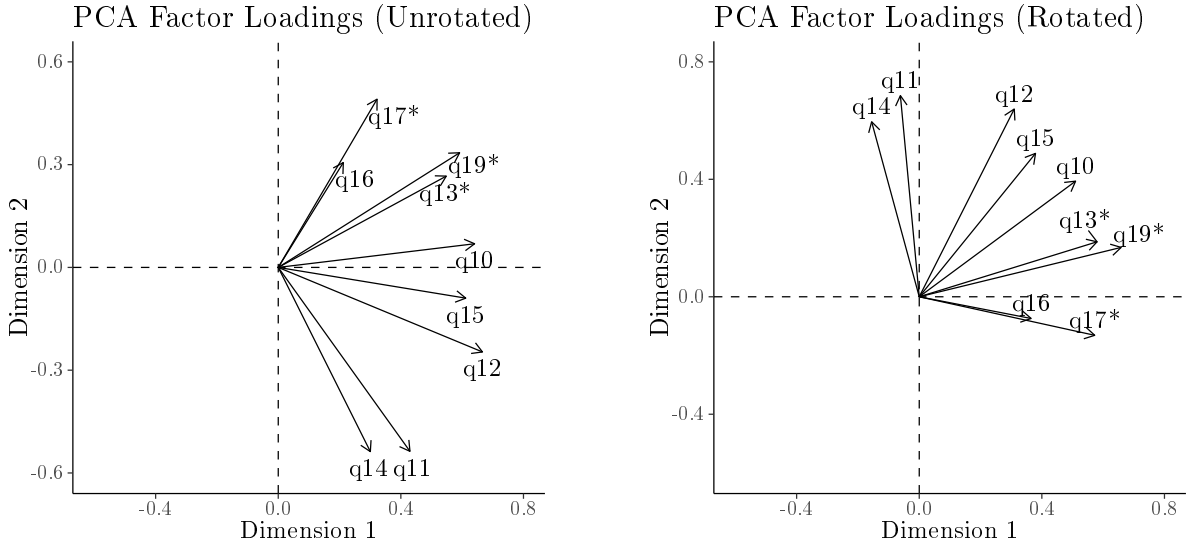


Figure 5.1: Unrotated and rotated solution of the PCA for the original scales, with the rotation being done using Varimax.

	Topic	Unrotated		Varimax Rotation	
		Component 1	Component 2	Component 1	Component 2
Q10	Free market	0.64	0.07	0.51	0.39
Q11	Public sector	0.43	−0.54	−0.06	0.69
Q12	Intervention	0.67	−0.25	0.31	0.64
Q13	Redistribution*	0.55	0.27	0.58	0.19
Q14	Spending	0.30	−0.54	−0.16	0.60
Q15	Firing people	0.61	−0.09	0.38	0.49
Q16	External loans	0.21	0.31	0.37	−0.07
Q17	Environment*	0.32	0.49	0.57	−0.11
Q19	Sale of land*	0.59	0.34	0.66	0.17
Eigenvalue		2.30	1.17	1.76	1.71
Variance		25.59%	12.99%	19.55%	19.03%

Table 5.1: Lithuania - Unrotated and Rotated PCA (Original Economic Scale)

Looking at the factor loadings, I find that except for the questions 14, 16, and 17, all questions have high enough loadings on the first component. I could thus interpret this dimension as a dimension measuring economic issues. Yet, there are two problems with this. First, is that some reversed questions load positive on the factor, while they should load negative. Second, some questions also load high on the second dimension in either a positive or negative fashion, except for questions 10 and 15. This means that the second dimension is also capturing a different underlying dimension.

Applying varimax rotation does not help. But, as varimax only changes the coordinates of the components to such degree to maximize the sum of the variances of the squared

loadings for each of the factors I did not expect this to be relevant if the underlying data is problematic. The rotated solution ensures that now both factors are equally important. Thus, with questions 10, 13, 17, and 19 load high on the first factor, and 11, 12 and 14, load high on the second factor. Looking at the topic of the questions, it is hard to find a relation between these questions and the reason why they are separated. Meanwhile, question 15 loads on both dimensions, while question 16 has a weak loading on the first dimension.

The picture becomes clearer when we look at Figure 5.1, which visualizes the loadings. In the unrotated version, all loadings are between the positive and negative part of the second dimension and the positive part of the first dimension. Besides, they are spread out, and apart from questions 10 and 15, few of the dimensions seem to correspond well to the underlying dimension. In the rotated solution, questions 11 and 14 line up well with the positive side of the second dimension, while questions 16 and 17 line op well with the first dimension. The other questions are located between the two and are thus best explained by both dimensions, with questions 10 and 15 being the best example of this.

		Unrotated		Varimax Rotation	
	Topic	Component 1	Component 2	Component 1	Component 2
Q10	Free market	0.66	0.07	0.67	0.06
Q11	Public sector	0.49	−0.42	0.49	−0.43
Q12	Intervention	0.71	−0.15	0.71	−0.16
Q13	Redistribution*	0.54	0.30	0.54	0.29
Q14	Spending	0.40	−0.51	0.39	−0.51
Q15	Firing people	0.62	−0.03	0.62	−0.04
Q16	External loans	0.03	0.59	0.04	0.59
Q17	Environment*	0.17	0.61	0.18	0.60
Q19	Sale of land*	0.55	0.37	0.56	0.36
Eigenvalue		2.35	1.40	2.35	1.40
Variance		26.14%	15.51%	26.14%	15.52%

Table 5.2: Lithuania - Unrotated and Rotated catPCA (Original Economic Scale)

Using PCA on the data-set thus presents us with a confusing picture. So, I turn now to the catPCA solution shown in Table 5.2. To run the catPCA algorithm, I have to provide the number of expected dimensions beforehand. Here, I accepted the standard options and choose for 2 dimensions. Here, I see that in both cases the eigenvalues and explained variances have improved from 25.56% to 26.14% on the first dimension and from 12.99% to 15.51% on the second dimension. This is as expected, as catPCA does not restrict the data to be metric, as PCA does. Note that this is related to the idea of the DDI and also explained why the DDI is high.

Comparing the component loadings with those in Table 5.1 shows that the loadings for questions 10, 11, 12, 13, 14, 15, and 19 are similar or have even improved. The loadings for 16 and 17 on the first dimension have been reduced, but both now load higher on the second dimension. For the other questions, a similar story goes as for the first dimension, in that most of the loadings remain the same, except for questions 16, 17 and 11, whose loadings have decreased. Rotating the solution, in this case, did not make any difference.

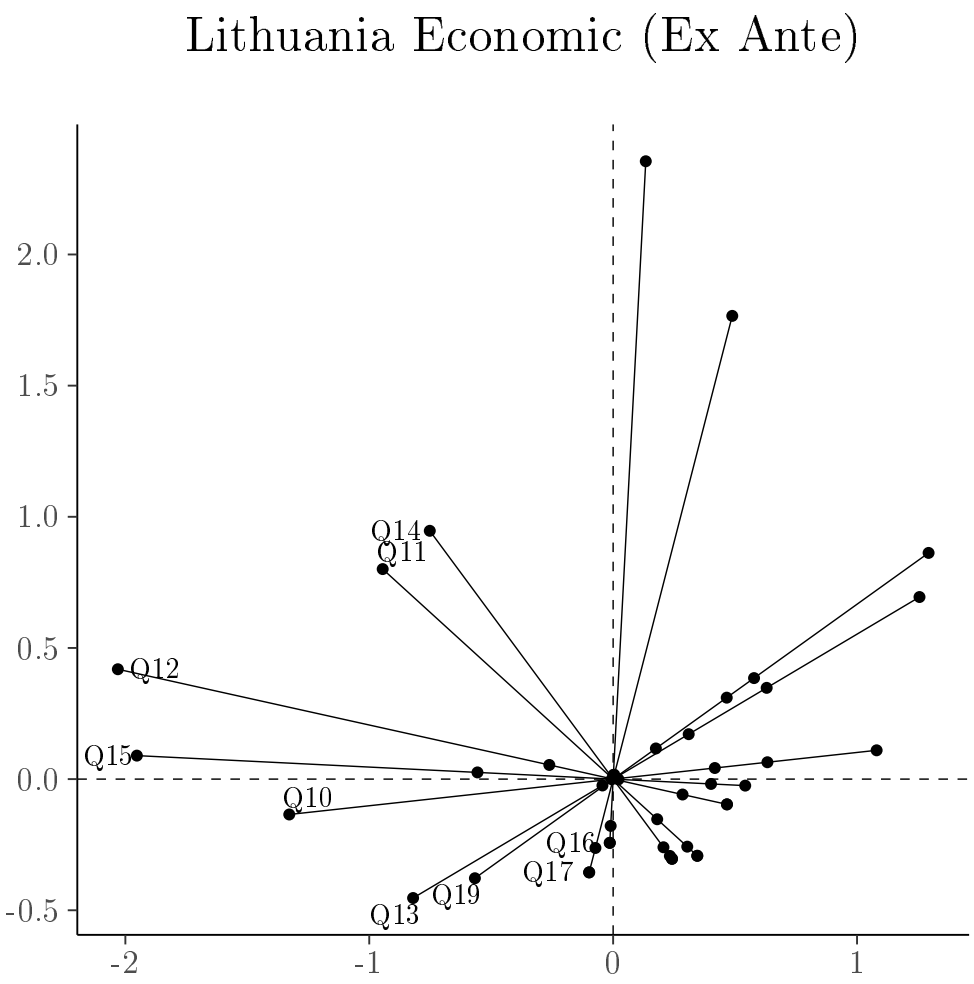


Figure 5.2: Biplot axes of the catPCA solution for the Original Economic Dimension in Lithuania

The biplot is similar to the graph showing the component loadings, but also indicates the position of the categories. Figure 5.2 shows this biplot while Table 5.3 shows the categories. In a biplot, the distance between two different axes is a measure for the association, and we can interpret them like as any other system of coordinates (Blasius and Gower 2005). Figure 5.2 shows a biplot of the nine economic questions, with the labels fixed to the "completely agree" end and points indicating the other category points. Looking at both the Figure and the Table, I can identify four problems with this scale. First, if the questions had been related, I would have only observed a single bundle of axes. But in this case, I can identify at least four. One bundle contains the questions 11 and 14 which deal with classical economical questions like the reduction of the public sector and cutting government spending. Questions 10, 12, and 15 are more concerned with the free market and related to the ease of firing employees, support for the free market and government intervention in the market. Questions 13 and 19 deal with redistribution and the sale of land, which do seem unrelated, while questions 16 and 17, which are almost perpendicular to questions 11 and 14, deal with the reduction of the public sector and cutting government spending - questions that relate to the financial

crisis. These four bundles show that the original scale didn't measure a single underlying dimension, but many.

	Topic	CA	A	N	D	CD
Q10	Free market	−2.00	−0.01	0.63	0.95	1.63
Q11	Reduction public sector	−1.92	0.37	0.62	0.70	0.70
Q12	Intervention	−2.87	−0.37	0.40	0.66	0.66
Q13	Redistribution*	−1.52	−0.08	0.58	1.17	2.33
Q14	Spending	−1.87	0.51	0.58	0.60	0.60
Q15	Firing people	−3.17	−0.90	0.04	0.65	0.88
Q16	External loans	−0.41	−0.41	−0.30	0.03	4.00
Q17	Environment*	−0.59	−0.59	−0.43	0.03	2.92
Q19	Sale of land*	−1.02	0.32	0.84	1.04	2.34

Table 5.3: catPCA Quantifications for the Original Economic scale in Lithuania. Questions with an asterisk (*) were reversed in the scale. The response options are *completely agree* (CA), *agree* (A), *neutral* (N), *agree* (A), and *completely disagree* (CD).

Besides, the biplot shows that the categories are not equally distributed - explaining the high DDI score. In the case of questions 11 and 14, the "disagree" and "completely disagree" categories are very close together, indicating that users did not distinguish between the two. Moreover, in the case of question 11, the neutral category is not in the centre of the plot as we would expect, but actually closer to the disagree categories. This indicates that users who select the neutral category actually disagreed with the question. We can see similar problems for question 16. Here, the "completely disagree" category is far away from the origin of the graph and from the other categories. Looking at the values in 5.3, I find the CD category has a quantification of 4.00, while the next category — "disagree" — has a value of 0.03. This indicates the "disagree" category is actually the neutral midpoint, while the "neutral" category indicates agreement with the question. Moreover, the "completely agree" and "agree" category have the same value and are thus "tied" - in other words, they measure the same opinion and users do not distinguish between them. It thus seems that users simplified the responses to question 16 to a binary option: either one Completely Disagreed, or they were Neutral/Agreeing on the question. This behaviour is most likely caused because users possessed little knowledge about the question - "External loans from institutions like the IMF are a good solution to crisis situations" and thus simplified their answers. In any event, the question is far from metric, and we can best consider it to be binary.

Other ties occur in questions 11, 12, 14 and 17 - in the first three cases between the "completely disagree" and "disagree" categories, and in the fourth case between the "completely agree" and "agree" categories. Moreover, for questions 11, 14 and 19, the "neutral" category does not represent the true middle but is actually on the disagreeing side of the scale. It thus seems that users used the "neutral" category to express disagreement - I will return to this in the MCA later. This is most pronounced in question 14 - "Cutting government spending is a good way to solve the economic crisis". Here, even the "agree" category is in fact on the negative side, while the CA category is far from the centre on the positive side. Moreover, while not strictly ties, also the "neutral" and "agree" categories are rather close together and close to the "disagree" and "completely disagree" categories. As such, as for question 16, it

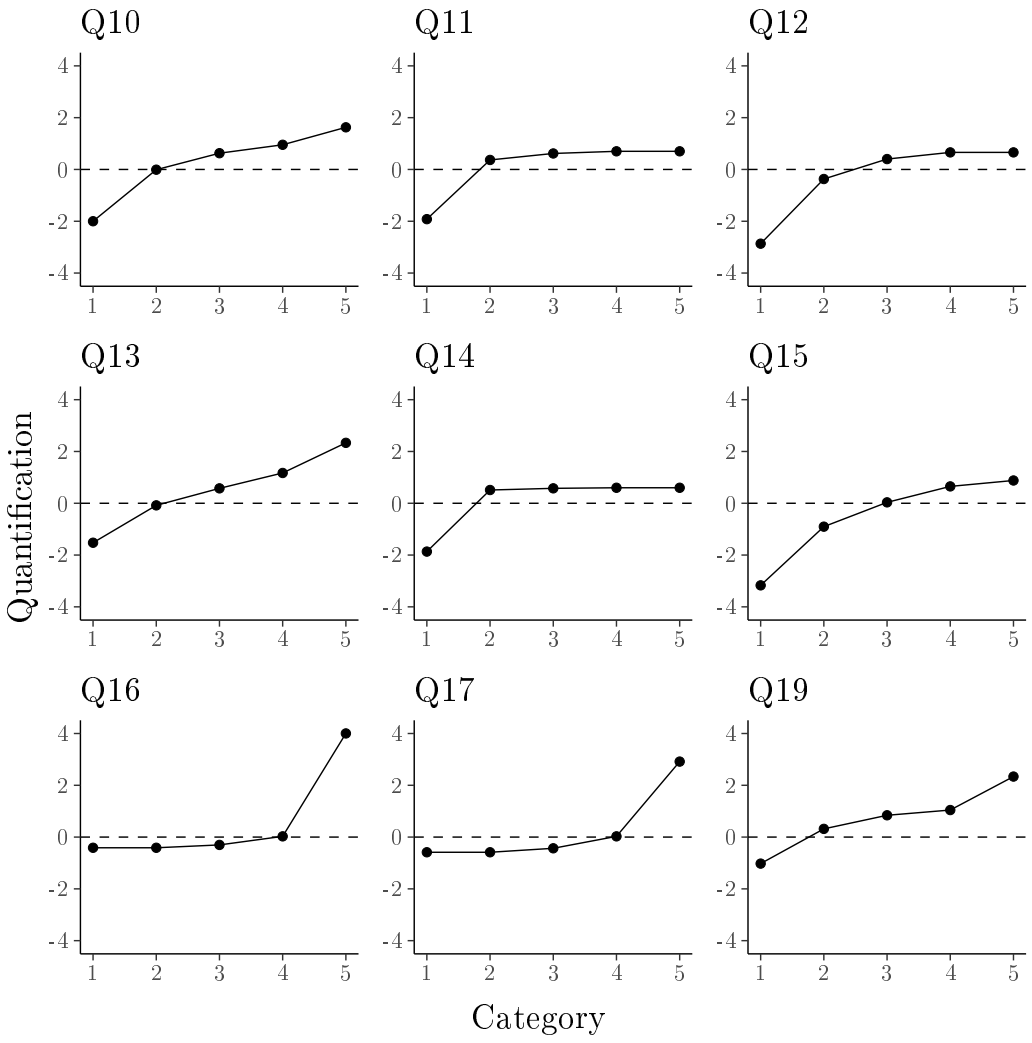


Figure 5.3: Line plot showing the relation between the quantifications and the original categories

seems that users simplified this question into a binary form. Either they completely agreed with cutting government spending, or they did not.

The result of this is that for no question the distances between the categories are equal. When we remember that the DDI measures the difference between the PCA assumed categories (which have equal distances) and the catPCA categories, this explains why the DDI was so high. Take question 10: the difference between the CA and A categories is $-0.01 - (-2.00) = 1.99$, which is more than three times larger than the difference between the A and N categories: $-0.01 - (0.63) = 0.64$. The distance between the CA and the A categories was thus much larger in the eyes of the respondents than the distance between the A and the N categories.

Once again, I can visualize all this to get a more coherent view of what is going on. Figure 5.3 shows both the original categories and their respective quantifications. In this figure, the 0 line indicates where we expect the "middle" of the latent variable to be. If the question is metric, the dots and the line connecting them form a diagonal line crossing the 0 line. At this point, the category is 3 and both the extreme points have the same absolute values. Figure

5.3 shows that this is rarely the case for any of these questions. Best performers, as we saw earlier, are the questions 10, 13, and 19, though all three cross the 0 line already around or before the second category. Questions 11, 12, 14 and 15 show a similar shape with the line crossing the 0 line, only to flatten afterwards. We find an opposite effect for questions 16 and 17, where the categories remain under the 0 until the fourth category, only to cross it afterwards. Once again, these plots show that the data is far from metric.

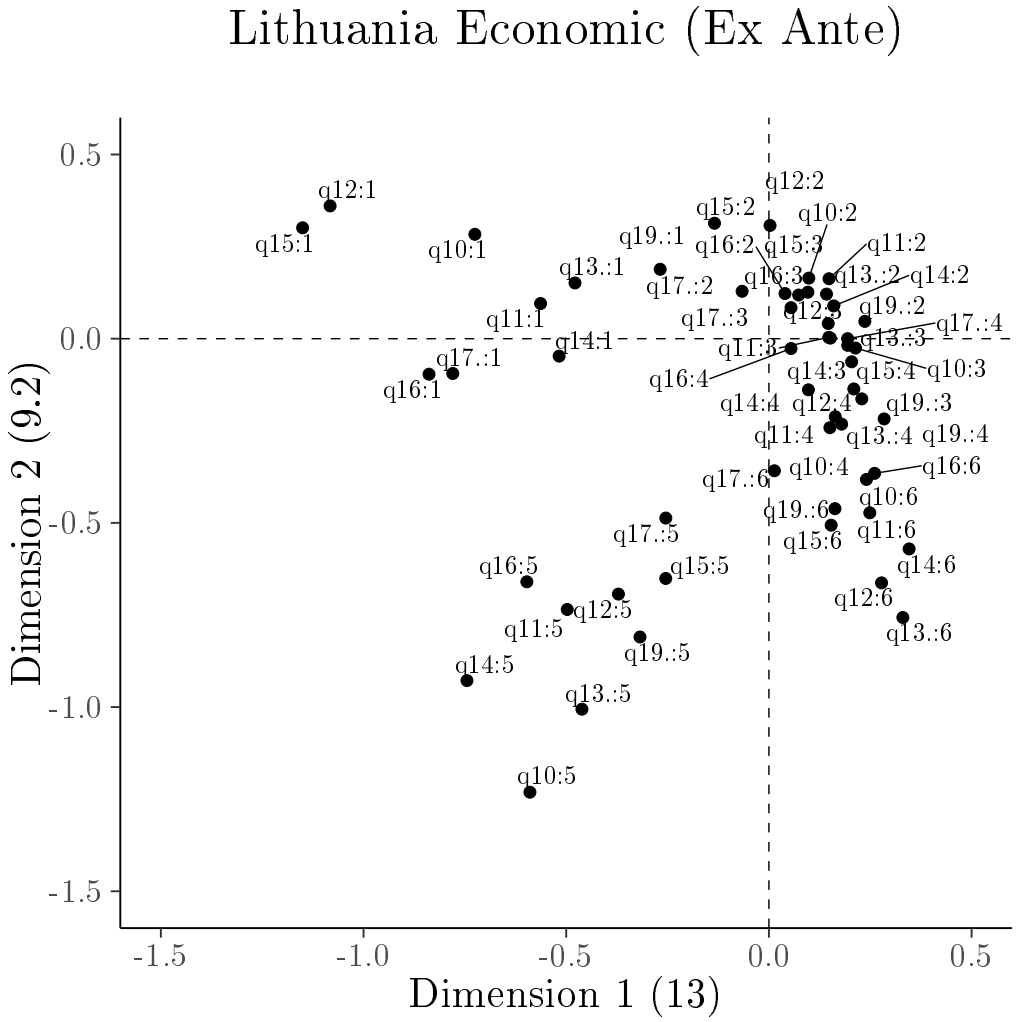


Figure 5.4: MCA for the 1st and 2nd Dimension of the original economic scale in Lithuania

To get an even better view of how such answer patterns can emerge, I run MCA on the data[3]. Figure 5.4 shows the results of this. Here, the first dimension, which should allow for a substantive interpretation (economy in this case), instead distinguishes between extreme and non-extreme response options. The horseshoe, which we would expect to find on the second dimension, is thus formed on the first dimension instead. On the second dimension, we have the dimension we were interested in (running from the top (economic left) - to the bottom (economic right)). Yet, this dimension explains less of the variation in the data than the difference between the extreme and non-extreme categories (13% vs. 9.2%). When we

[3]Note that for this MCA, as for all others in this chapter, the categories are as follows: (1) Completely Agree, (2) Agree, (3) Neither Agree nor Disagree, (4) Disagree, (5) Completely Disagree, (6) No Opinion.

would plot the points on the second dimension (which we can do by drawing imaginary lines from the points to the axes of that dimension), we can see that while they maintain their ordinality, several of them cluster close together. The only bit of good news here is that the reversed questions (13, 17, and 19) have indeed been recognized as such. Finally, there is an association between the no opinion category and the non-extreme categories on the first dimension and the right side of the economic dimension.

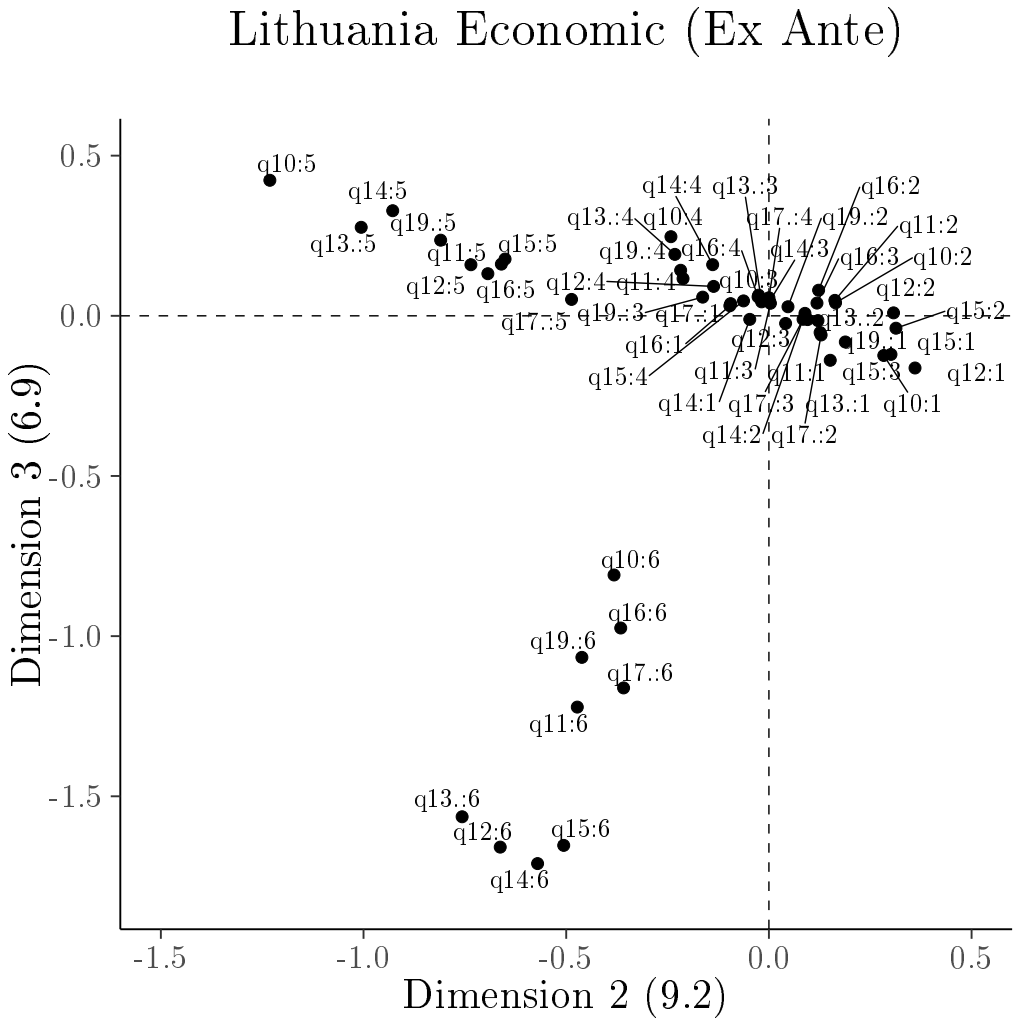


Figure 5.5: MCA for the 2nd and 3rd Dimension of the original economic scale in Lithuania

Figure 5.5 shows this in more detail and shows the second and third dimensions. While the second dimension shows the latent dimension, the third dimension separates the substantial from the non-substantial answers. These non-substantial questions have a high associated with the right side of the second dimension, as we saw above. Besides, we see that some questions (like 10 and 16) are closer to the substantial side of the dimension than the other questions. This can indicate that in these questions, the "no opinion" option was often used as a way to hide non-response. Moreover, the "completely disagree" categories are related to the positive side of the third dimension, indicating that users of these categories rarely choose a no opinion.

Thus far I have identified five reasons why the DDI for Lithuania was low. First is that

the scale is not made up of a single scale, but rather of three (or four) different ones. Second is that there are many ties in the data. The third is that the midpoints of the questions do not correspond with the actual midpoints. Fourth is that the cause of the main variation in the data is the difference between extreme and non-extreme responses. Fifth is that there is a relationship between an economic right position and providing a "no opinion" response. As such, we can hardly interpret the original economic scale for Lithuania as measuring an economic position. One of the reasons for this is because it seems that many users have simplified. Instead of using all the options of the response scale, they focused only on the extreme options - revealed in the MCA and as the ties in the catPCA. This is a form of ERS, which I discussed in the previous chapter.

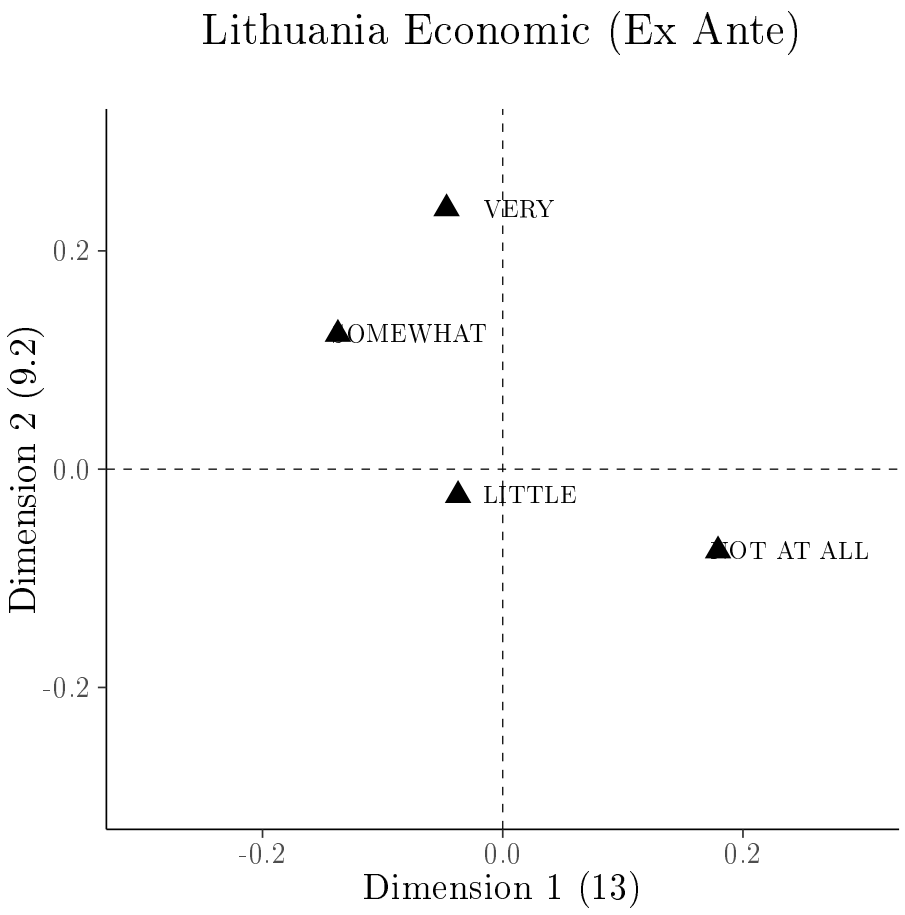


Figure 5.6: MCA for the 1st and 2nd Dimension of the original economic scale in Lithuania, with the political interest options.

ERS can occur when respondents have little interest or knowledge of the domain in question. To get a better idea of how interest shapes the way the users respond to the VAA, I use a supplementary question in the VAA. This question asked "To what extent are you interested in politics?". This question had four response options: "Very", "Somewhat", "Little" and "Not at all". Figure 5.6 shows the result of this. Note that the scale of Figure 5.6 is smaller than the scale in Figure 5.5. Here, we see the categories follow a horseshoe-shaped pattern, though the orientation of the horseshoe is opposite to the one we found earlier. The categories are thus related to both the underlying economic dimension and the

way of providing responses. Users who state they are somewhat knowledgeable have a higher tendency to use extreme response categories then users who are very interested or only a little interested. Moreover, users who are not at all interested are on the same side of the spectrum as the "neutral" and No Opinion categories, indicating that these users favoured these categories. Moreover, users are more interested as soon as their underlying economic preferences are more left-oriented, though the effect is modest.

	Completely Agree	Agree	Neutral	Disagree	Completely Disagree	No Opinion
Free market	16.2	40.6	19.7	14.5	3.1	5.9
	8.9	35.2	23.6	16.7	2.4	13.2
Public sector	20.3	39.5	17.0	16.8	2.8	3.6
	14.3	32.8	19.8	23.3	4.1	5.7
Intervention	8.6	27.3	21.1	33.9	7.9	1.2
	3.8	19.4	24.3	41.0	8.3	3.2
Redistribution	21.7	42.1	18.6	13.2	4.0	0.5
	20.0	41.0	17.8	14.8	5.3	1.3
Spending	22.1	52.4	13.7	8.9	1.6	1.3
	16.8	53.6	13.8	11.3	1.5	3.0
Firing people	5.2	21.3	21.0	41.6	10.1	0.8
	2.9	15.2	19.0	47.9	14.0	1.2
External loans	4.0	29.0	30.0	27.0	5.3	4.6
	3.0	19.1	31.4	29.3	5.5	11.6
Environment	5.0	22.9	31.3	30.7	9.5	0.6
	3.8	20.5	32.3	31.0	11.8	0.8
Sale of land	45.6	30.6	8.5	9.1	5.5	0.7
	24.8	35.0	14.9	15.6	7.7	2.0

In each cell, the first row is for high political interest and the second row for low political interest

Table 5.4: Percentages for the original Economic scale in Lithuania, divided by political interest

Still, this recounts only a part of the story to us: which categories the users would choose. To see whether they also respect ordinality, I look at them as separate groups. To ease the comparison, I merge the "Very" and "Somewhat" groups into a new "High" group and the "Little" and "Not at all" groups into a new "Low" group. First, I look at the distribution of the answers to the questions as shown in Table 5.4. Here, I see several interesting differences between the two groups. To begin with, the users with low political interest (shown in the

second row for each question), use the "neutral" and No Opinion categories more often than their those who have a high interest (except for the question on redistribution and firing people). For example, for the question on the free market, 36.8% of the users in the Low group use either the "neutral" or No Opinion categories, compared to only 25.6% for the High group, with something similar happening in the case of the question on external loans. Moreover, for both that question and the one on the environment, the percentage of users selecting the neutral category is high - in both cases for both the groups around 30%. This can either mean that users were neutral on the subject or used the category to hide their non-opinion. Of the genuine opinions of the users, there are no differing patterns. In other words, users with a high and low political interest do not think differently about the questions on this scale. If there are any large differences — such as for the question on external loans — the difference is most often explained by a similar difference in the "No Opinion" category.

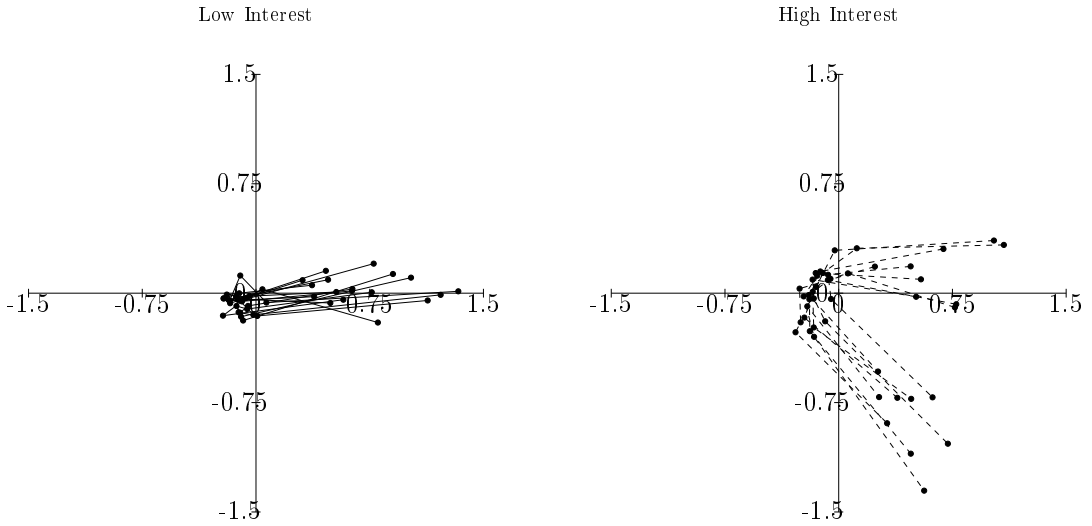


Figure 5.7: MCA Plots for the Low and High political interest groups in Lithuania, following the original economic scale

Even so, while the actual answers are similar, except for the use of the "No Opinion" and "neutral" categories, we cannot say the same for the structure of the data itself. Figure 5.7 shows an MCA for both groups on the first and second dimension. Here, we can see that where the high-interest group shows a clear horseshoe — though rotated and having the second dimension instead of the first as the main latent dimension — the graph for the low-interest group is hard to interpret. It seems that there is a little variation on the second dimension, as all points are cramped together. Instead, most variation takes place on the first dimension, which in this case represents the extremeness of the response styles.

To assess which questions are most problematic, I additionally separate the different questions, as shown in the figures on this page and the next. Starting with question 10, we see that the high group (the dotted line), follows the horseshoe pattern and has the correct order of the categories. Yet, the low group (shown by the solid line) not only has a wrong ordinality (the 5 and 1 should be on opposite sides) but also has the other three categories very close together. Thus, users with low political interest confused the two extreme categories

Q10: Free market competition makes the health care system function better

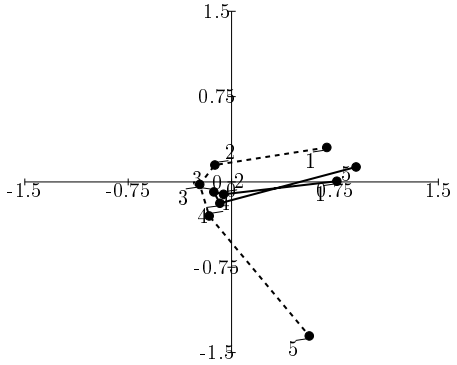


Figure 5.8: Lithania - MCA (Original Economic Scale (Question 10))

Q11: The number of public sector employees should be reduced

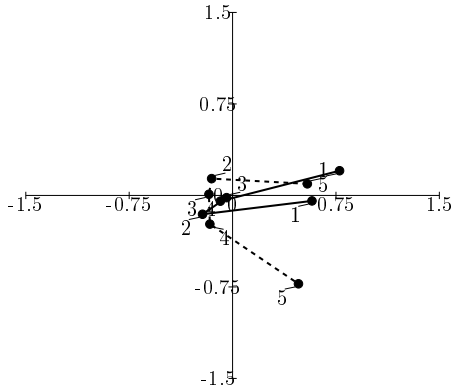


Figure 5.9: Lithania - MCA (Original Economic Scale (Question 11))

Q12: The state should intervene as little as possible in the economy

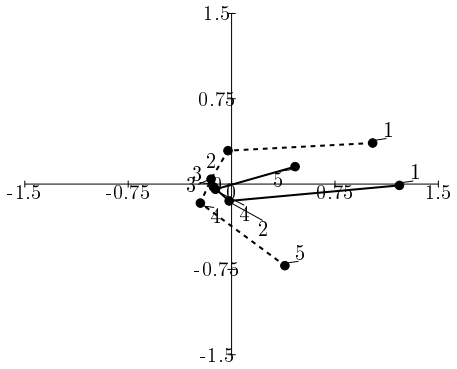


Figure 5.10: Lithania - MCA (Original Economic Scale (Question 12))

Q13: Wealth should be redistributed from the richest people to the poorest

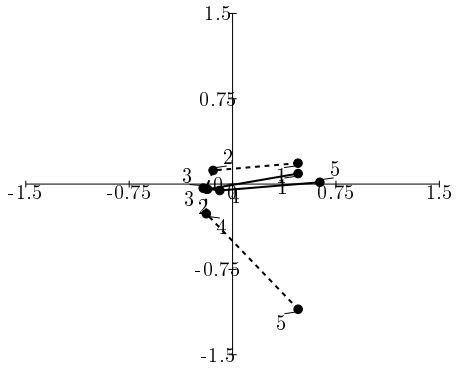


Figure 5.11: Lithania - MCA (Original Economic Scale (Question 13))

and did not distinguish between the others. A similar story goes for both questions 11 and 12. Here, the ends of the categories are again opposite to where they should be. Moreover, the categories for the low group run as $5-3-4-2-1$, with 3 ("neither agree nor disagree") and 4 ("disagree") being in the same place. As such, users with a low interest did for this question not distinguish between these two categories and treated them as similar.

For questions 13 and 14, we see that while the categories are in the right order, both the 1 and 5 category are on the positive side of the second dimension, while the other categories cluster on the negative side. This indicates that users neither distinguished between these categories, but also that they used both "completely disagree" and "completely agree" to express agreement and the other categories to express disagreement. In the case of questions 15, 16, and 19, we find that category 1 - indicating "completely agree", relates to the first dimension, indicating that in both cases this was an extreme category. We already saw this

Q14: Cutting government spending is
ı good way to solve the economic crisi:

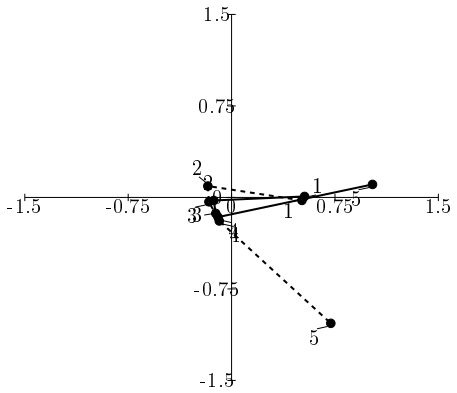


Figure 5.12: Lithania - MCA (Original Economic Scale (Question 14))

Q15: It should be easy
for companies to fire people

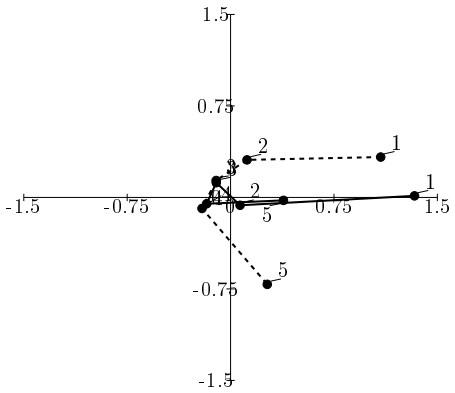


Figure 5.13: Lithania - MCA (Original Economic Scale (Question 15))

Q16: External loans from institutions
such as the IMF are a good
solution to crisis situations

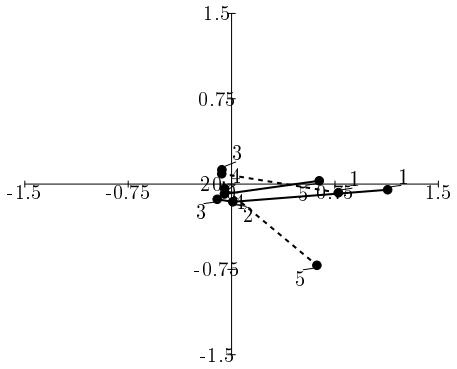


Figure 5.14: Lithania - MCA (Original Economic Scale (Question 16))

17: Protecting the environment is mo
ıportant than fostering economic grow

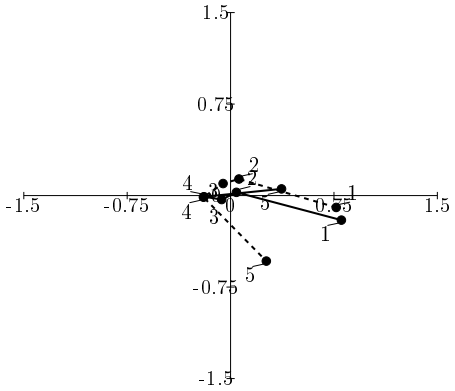


Figure 5.15: Lithania - MCA (Original Economic Scale (Question 17))

when studying the catPCA quantifications, where the category had a rather large negative score of -3.17. But while for the high group, the remaining categories form an arc, the categories for the low group behave differently. After category 1, category 2 and 4 are both placed on the other side of the second dimension, while the "neutral" category 3 is placed on the positive side. Moreover, categories 2 and 4, which should have been placed on similar places on the 1st dimension, are instead separated. Again, this shows that users with a low interest did not use the categories as intended. Question 17 finally, seems to be problematic for both the high and low group. Yet, where for the high group the only problem is that the "completely agree" category is on the wrong side of the dimension, for the low group the other problem is that again the "completely disagree" category is on the wrong side. Also, the other three categories are so close together that they are interchangeable.

As such, for both the high and low-interest groups, the main dimension of variation is

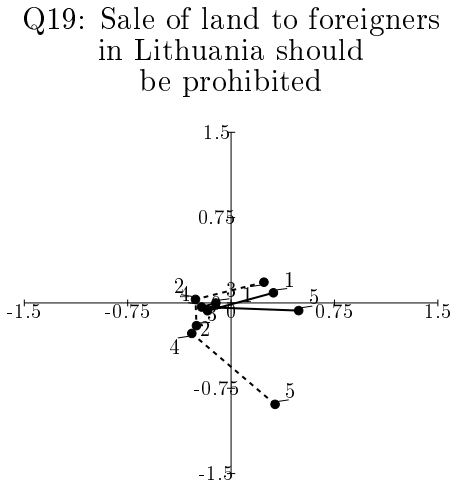


Figure 5.16: Lithania - MCA (Original Economic Scale (Question 19))

the extremeness of the answer categories and not the underlying economic dimension. Yet, these problems are way more serious for those users with a low political interest then for those having a high political interest. For the latter, the horseshoe shape is intact, and the ordinality of the data is in order. For the former, the distribution is confusing and categories are often mixed.

Table 5.5: catPCA Quantifications for the DSV Economic scale in Lithuania. Questions with an asterisk (*) were reversed in the scale.

	Topic	CA	A	N	D	CD
Q10	Free market	−1.96	−0.07	0.66	1.08	1.35
Q12	Intervention	−2.69	−0.52	0.29	0.75	0.90
Q13	Redistribution*	−1.70	0.03	0.85	1.13	1.13
Q15	Firing people	−3.08	−0.91	0.00	0.59	1.15
Q19	Sale of land*	−1.13	0.54	1.01	1.25	1.31

All this leads us to conclude that the original economic dimension for Lithuania was far from valid. So, I will now turn to the DSV economic dimension which aimed to improve on the situation. For this, DSV dropped 4 of the original 9 questions. All these questions showed a considerable number of tied categories and were in two separate groups orthogonal to each other. As a result, the DDI for the DSV version improved to 0.27 indicating the data is on the far end of being "good" in the terms of Blasius and Thiessen (2012).

This shows in the quantifications. Table 5.5 and Figure 5.17 show only a single tie in the data. Also, they show a midpoint of the underlying dimensions that is more in the middle than before. Yet, there are still differences between the questions. For questions 10 and 13, the neutral point lies at the "agree" category (2) instead of in the middle, while for question 19, the neutral point is between the "completely agree" and the "agree" categories. Given that the distances between the other categories are rather small, this indicates users saw this question as binary. Either one completely agreed that the sale of land to foreigners should be

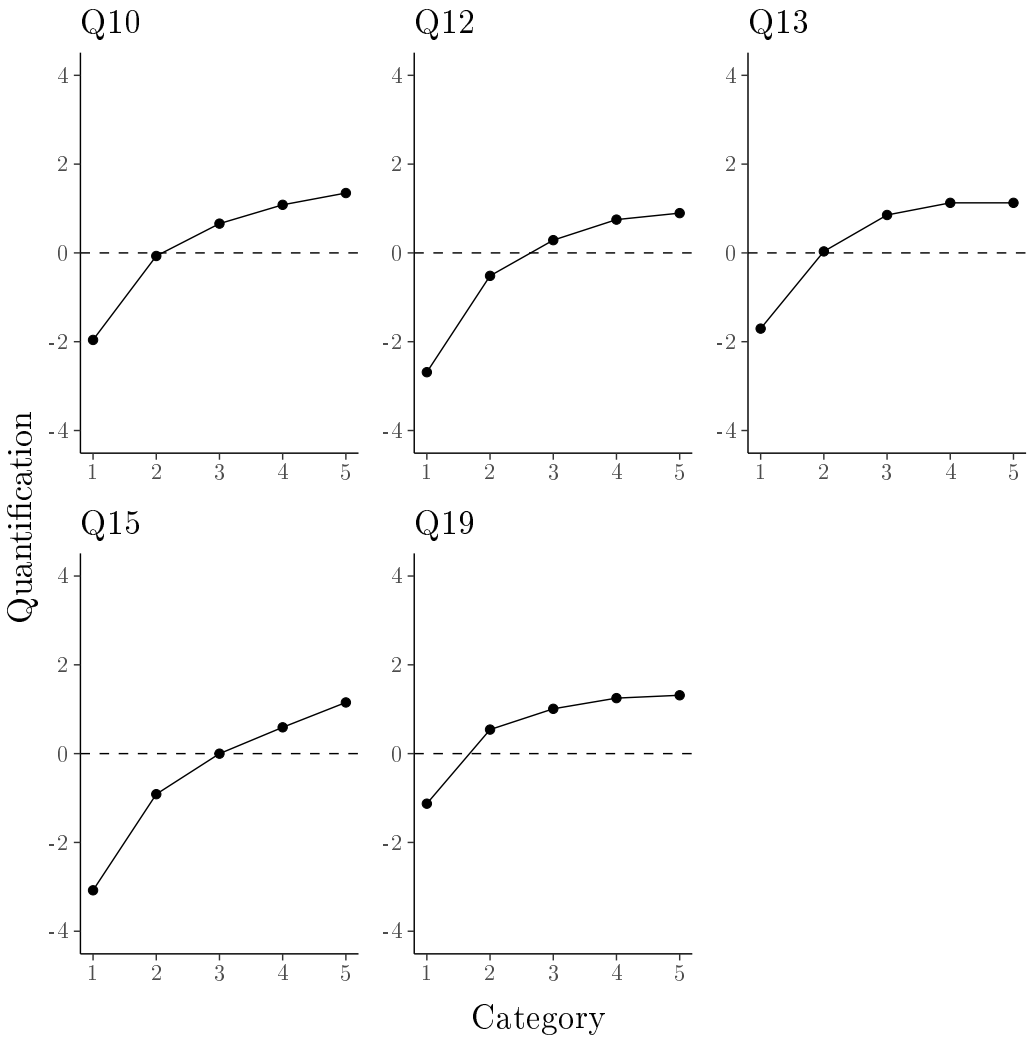


Figure 5.17: Line plot showing the relation between the quantifications and the original categories

prohibited, or one did not, with no marked distinction in whether one agreed, disagreed, or was neutral on the matter. Finally, question 15 performs best of all, with the neutral category at 0.00 and a straight diagonal line except for the extreme "completely agree" category.

Looking at the biplot for this version of the scale in Figure 5.18 I see that the original four clusters have been reduced to two. This also explains why the scalability increased from 0.17 to 0.35. Looking at the biplot for this version of the scale in Figure 5.18 I see the original four clusters have been reduced to two. Both groups of questions are perpendicular to each other, indicating they are not related to each other, which is problematic if I want them to form a single scale.

We can find further problematic evidence when we run MCA on the data. Figure 5.19 shows the MCA for the 1st and second dimension. The shape of the horseshoe is the same, leading to a similar conclusion as for the original scale: the key driver for the variation is the extremeness of the response. Yet, the effect is not equally distributed. The "completely disagree" categories (5) are associated way more with the second dimension as are the others. Besides, they seem to form a separate cluster on the far negative side of the second dimension.

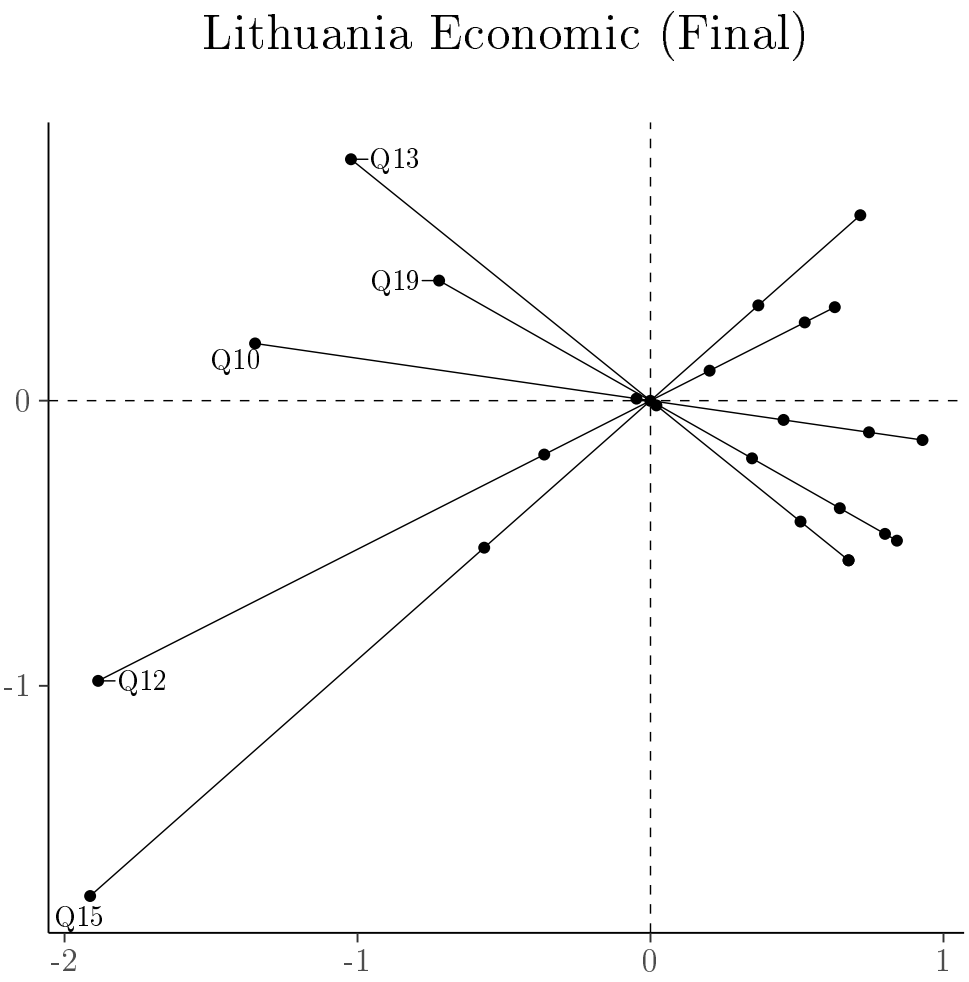


Figure 5.18: Biplot axes of the catPCA solution for the DSV Economic Dimension in Lithuania

Moreover, the horseshoe shape which we expected on the second dimension is instead located somewhere between the first and second dimension, along with an imaginary line that runs at a 45° angle through the origin. We can thus interpret the first dimension as running from the "completely agree"/"completely disagree" questions on the left, through a collection of "no opinion" and other substantial questions, to three of the "No Opinion" questions on the right. We can interpret the second as running from the "agree" category through the "completely agree" category, and then through the "neutral", "disagree" and No Opinion categories to end up at the "completely disagree" categories.

Figure 5.20 shows this by visualizing the second and third dimension. Here, we can see the incorrect ordering of the categories on the second dimension better and note that the third dimension separates the No Opinion categories from the others. But, the No Opinion categories are related to the second dimension as well, meaning that users who were on the disagreeing side of the argument, often chose the No Opinion category as well.

To sum up, for Lithuania, I have seen that while the dirty data index improved between the original and DSV scales, the actual scalability did not. Using catPCA, I discovered this is

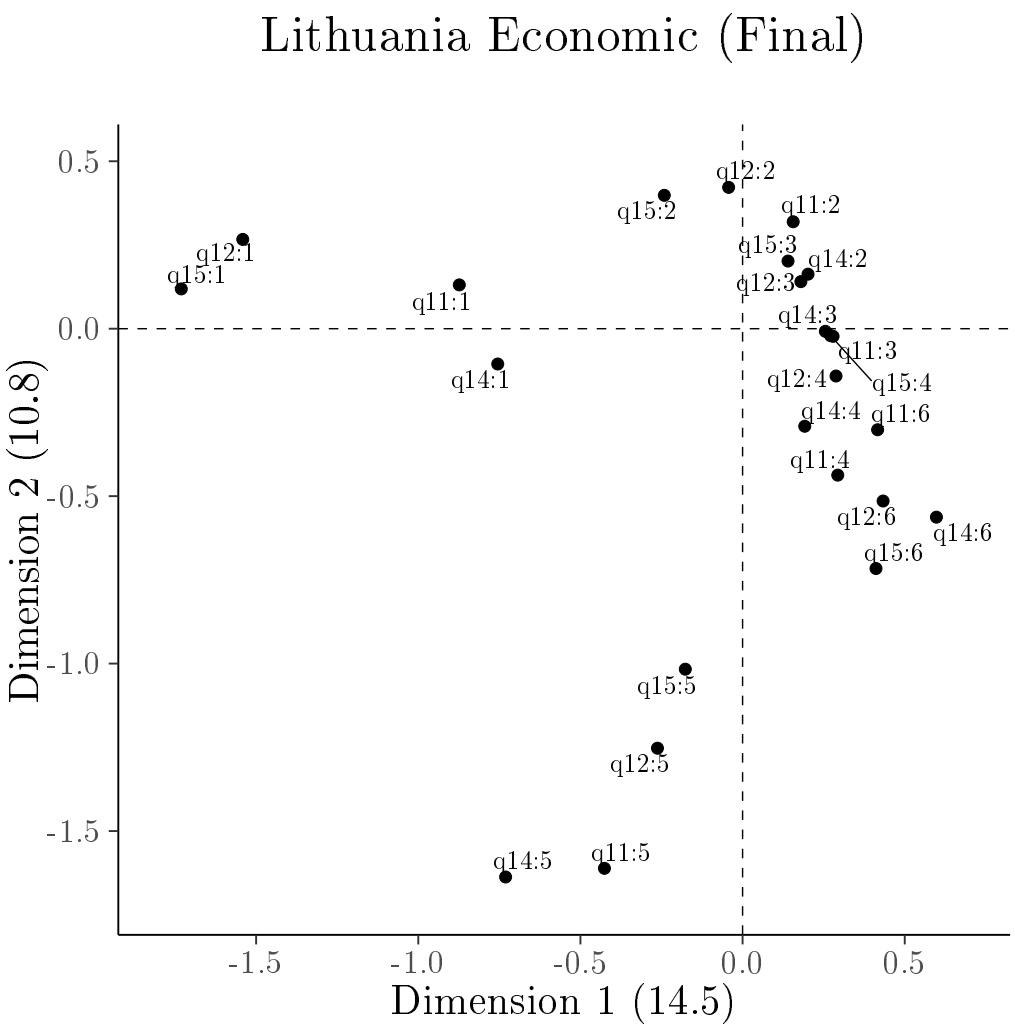


Figure 5.19: MCA for the 1st and 2nd Dimension of the DSV economic scale in Lithuania

because while the quantifications are more alike their PCA counterparts, the economic scale itself is still problematic. The catPCA biplot revealed that in fact there are still two distinct clusters in the unidimensional scale, and the MCA revealed that we can interpret neither the 1st or second dimension with ease. Even in its DSV form, the economic scale in Lithuania is not useful either to position users in a political space or to compare with against other countries. Lithuania also reveals a limitation of the DDI. As it assumes the underlying scale is unidimensional, the DDI itself might be misleading if this is not the case. To make good use of the DDI, we should thus first make the underlying scale which it assesses unidimensional using a Mokken Scaling Analysis.

This is exactly the aim of the MSA. But while the number of questions on the scale was further reduced to only three questions, neither the DDI nor the H value of the resulting scale improved when compared to the DSV scale, while the LCRC even sustained a slight drop. The resulting scale, containing only three questions, which are related to one's opinion on the free market, without any references to other macroeconomic issues. As the DDI is 0.31, I expect the PCA solution for this scale to be like the catPCA solution.

Figure 5.22 and Table 5.22 show the results for both the rotated and unrotated version.

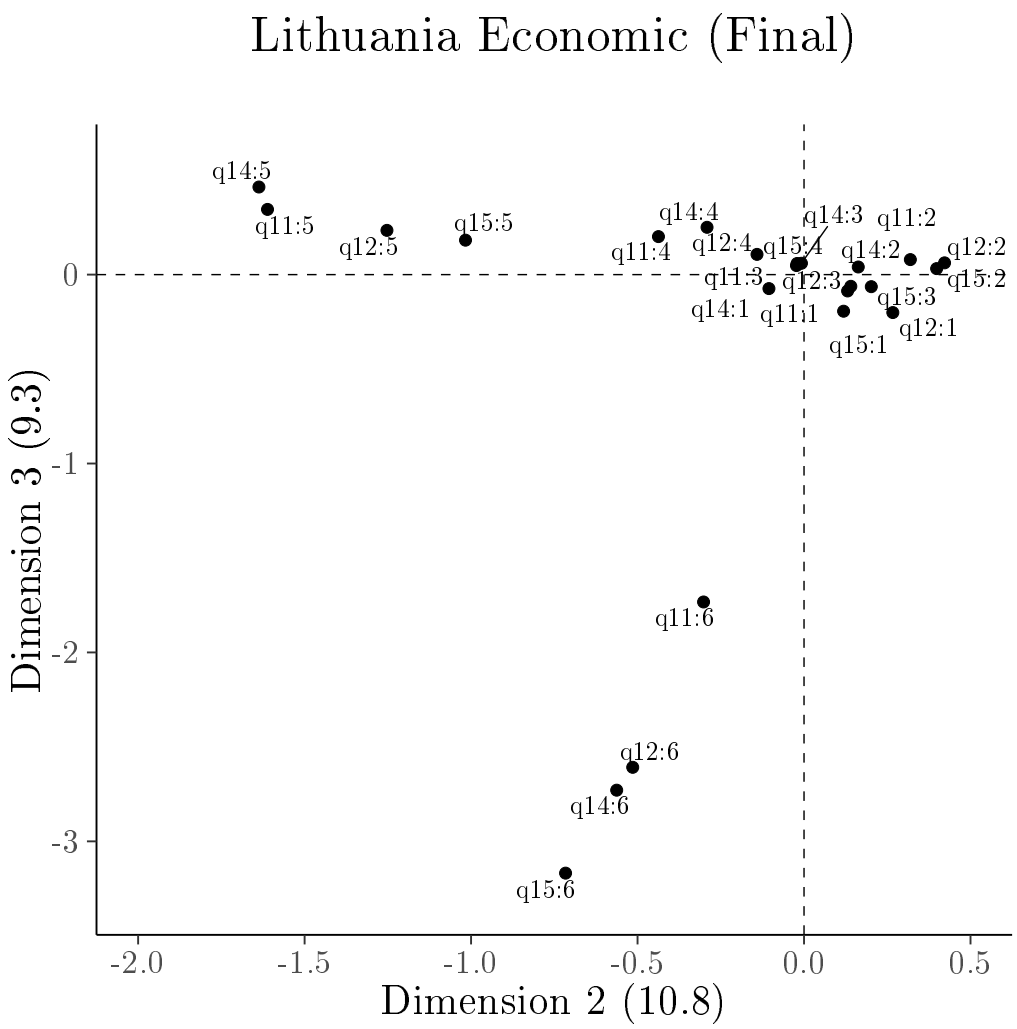


Figure 5.20: MCA for the 2nd and 3rd Dimension of the DSV economic scale in Lithuania

Here we see why the scale is still problematic, as questions 10 and 15 are perpendicular to each other. While both dimensions do explain the principal part of the variance in the data (78.69% for both dimensions), especially the rotated solution shows that whereas question 10 is strongly related to the second dimension, question 15 is related to the 1st dimension, with question 12 falling somewhere in-between.

Applying catPCA to the data does not seem to do much to the results as Figure 5.24 and Table 5.24 show. Though the eigenvalue on the first dimension has increased, the explained variance is the same and the loading plots show an identical pattern in both their unrotated and rotated form.

Looking at the quantifications in Figure 5.25 and Table 5.8, I see there are no tied categories. Moreover, in the case of question 15 the 0 line crossed close to the third category, while in the case of question 12 it is crossed between the "neutral" and "disagree" categories. The worst performing question is 10, where all quantifications until the "disagree" category are under the line. Also, both for questions 10 and 12 the distances between the first three questions are small compared to the significant jump between the "disagree" and the "completely disagree" category. This reflects the idea that users — especially for the questions

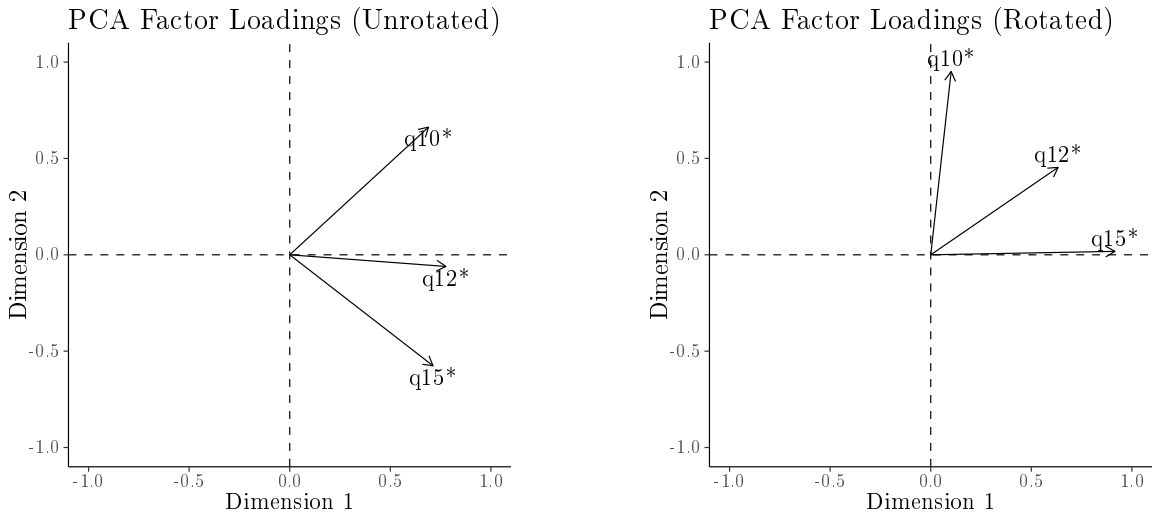


Figure 5.22: Lithuania - Unrotated and rotated PCA for the Quasi-Inductive scales

		Unrotated		Varimax Rotation	
	Topic	Component 1	Component 2	Component 1	Component 2
Q10	Free market*	0.69	0.66	0.10	0.95
Q12	Intervention*	0.78	-0.06	0.63	0.45
Q15	Firing people*	0.71	-0.58	0.92	0.02
Eigenvalue		1.59	0.77	1.25	1.11
Variance		52.89	25.80	41.67	37.03

Table 5.6: Lithuania - Unrotated and Rotated PCA (DSV Economic Scale)

10 and 12 — considered the first three categories to be close together, while they viewed "completely disagree" as an extreme option.

Running MCA on the first and second dimension (the left graph of Figure 5.26) shows a similar picture as we saw earlier - an arc-shape under a 45° angle with the 1st dimension. Projecting the categories on this 1st dimension, we find this dimension separates a collection of the categories "completely agree" — "neutral" — No Opinion — "disagree" on the left side and the centre from the "completely disagree" categories on the right side. The second dimension similarly separates the "disagree" — "neutral" — "completely disagree" — "agree" — No Opinion on the lower side of the dimension from the "completely agree" categories on the upper side of the dimension. Both dimensions seem to suggest that users treated the response categories as providing three options: "completely agree", "completely disagree" and a collection of "neutral"/No Opinion positions. The third dimension shown in the right graph of Figure 5.26 separates the No Opinion categories from the others, though especially in the case of question 10 the No Opinion category is close to the "completely disagree" category, indicating that some switching took place between them.

What then can we conclude from the case of Lithuania? To begin with, the simple observation that the economic dimension in Lithuania is problematic in any form. The original version was especially problematic, and any position on this dimension is not indicative for

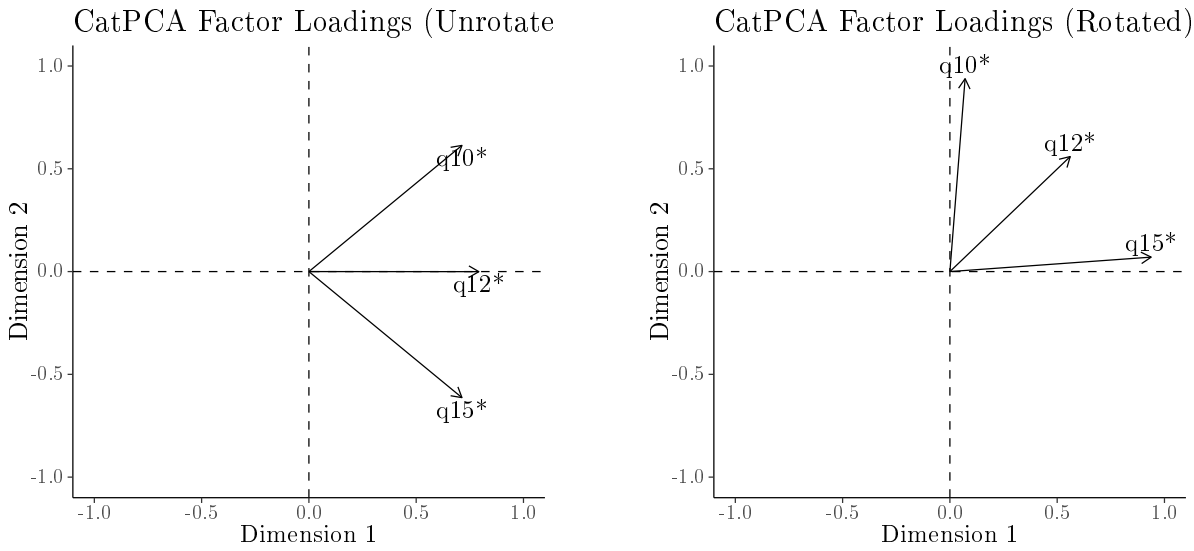


Figure 5.24: Lithuania - Unrotated and Rotated catPCA (DSV Economic Scale)

		Unrotated		Varimax Rotation	
	Topic	Component 1	Component 2	Component 1	Component 2
Q10	Free market*	0.71	0.61	0.07	0.94
Q12	Intervention*	0.79	-0.00	0.56	0.56
Q15	Firing people*	0.71	-0.61	0.94	0.07
Eigenvalue		1.65	0.75	1.20	1.20
Variance		54.83	25.09	39.98	39.95

Table 5.7: Lithuania - Unrotated and Rotated catPCA (DSV Economic Scale)

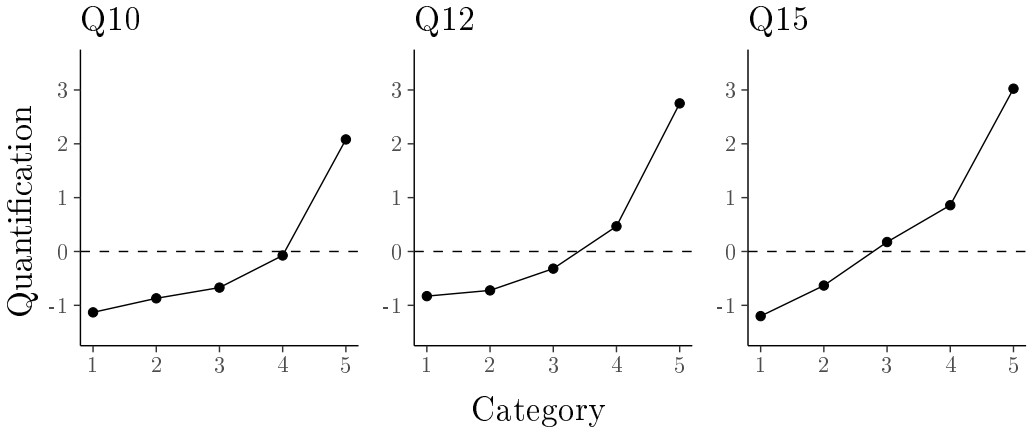


Figure 5.25: Lithuania - Transformation Plot (Quasi-Inductive Economic Scale)

	Topic	CA	A	N	D	CD
Q10	Free market	−1.13	−0.87	−0.67	−0.07	2.08
Q12	Intervention	−0.83	−0.72	−0.32	0.47	2.75
Q15	Firing people	−1.20	−0.63	0.18	0.86	3.02

Table 5.8: Lithuania - catPCA Quantifications (Quasi-Inductive Economic Scale)

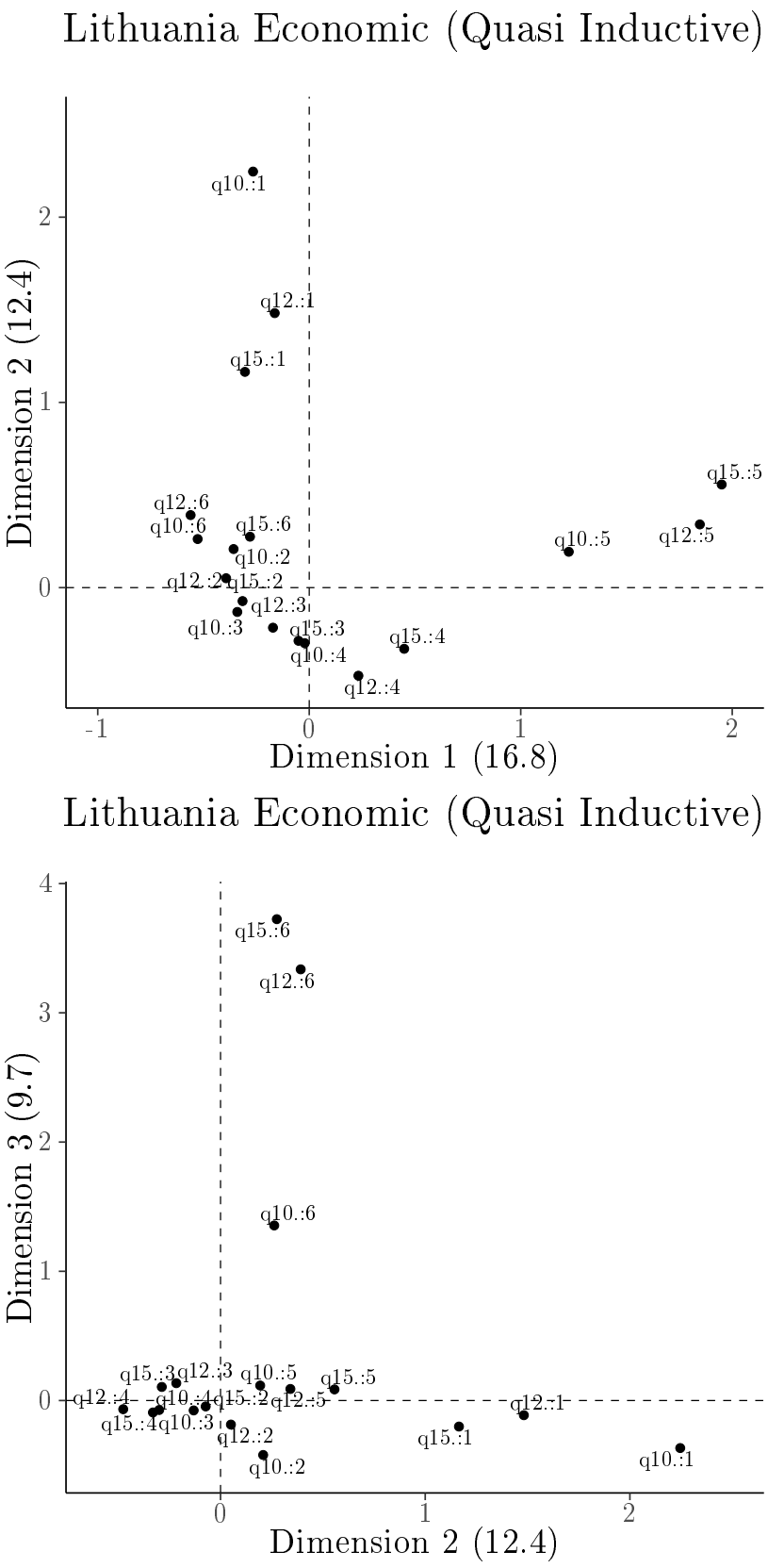


Figure 5.26: MCA for the 1st and 2nd dimension (left) and 2nd and 3rd dimension (right) of the quasi-inductive economic scale in Lithuania

the user's true position on the economy. The DSV scale fared better, but still had problems and only just approached the level of required values (especially in the case of Loevinger' H). The quasi-inductive scale meanwhile proves that while Mokken Scaling Analysis attempts to determine unidimensional scales, it can do only so much, especially if most of the other questions better line up on another dimension (in this case the EU one). More important yet is that this little overview proves that one can only do so much with bad data. It seems that the respondents in Lithuania often resorted to simplifying their responses and often did not understand how to use the Likert scale responses. Using this data in any kind of analysis is thus problematic, especially if one assumes the data to be near metric. Moreover, we have seen that where MSA can be certain about the consistency of the scale, MCA is useful for assessing the underlying distribution of the categories and pointing out problems with the data that would have gone otherwise unnoticed.

5.2 Ireland

One reason for the poor performance of Lithuania might be the country is still new to the European Union. To see if this is the case, I take a look at the EU dimension in Ireland. As Ireland has been a member of the EU since 1973, we should expect respondents to have ample experience with the topic.

	Topic	CA	A	N	D	CD
Q1	Euro*	-0.69	-0.13	0.78	1.44	3.98
Q2	Treaty change*	-0.78	-0.59	-0.18	0.33	3.18
Q3	Right to work*	-1.11	0.19	0.88	1.47	2.99
Q4	Common foreign policy	-0.64	-0.64	-0.50	-0.11	2.44
Q5	Redistribution	-2.29	-0.05	0.81	1.10	1.10
Q6	EU Membership*	-0.74	-0.03	1.07	2.04	3.98
Q8	Borrow money	-0.37	-0.37	-0.22	0.24	4.28
Q9	Russia	-0.39	-0.39	-0.25	0.19	3.58

Table 5.9: catPCA Quantifications for the Original EU scale in Ireland. Questions with an asterisk (*) were reversed in the scale.

The original scale for this dimension had a DDI of 0.39, which is explained by the quantifications in Table 5.9. To start with, we have four cases of ties (questions 4, 8, and 9 for ties between the "completely agree" and "agree" categories and question 5 for a tie between the "disagree" and "completely disagree" category). Moreover, only for questions 1,5 and 6 does the neutral category seem to represent the middle, while in the other case this is either the "agree" or "disagree" category. Furthermore, the "completely disagree" Category is often far removed from the other category, with the most extreme case between questions 2,8 and 9. In the case of question 4, a tie between the "completely agree" and "agree" categories and small distances between them and the other categories seem to suggest that users considered "completely disagree" to represent an extreme answer.

In its DSV version, the DDI dropped to 0.26 and lost questions 5,8, and 9. Yet, from Table 5.10 I find there are still problems. While there are no ties any longer, the "completely

	Topic	CA	A	N	D	CD
Q1	Euro*	-0.79	-0.05	0.90	1.60	3.58
Q2	Treaty change*	-0.85	-0.62	-0.14	0.40	3.09
Q3	Right to work*	-0.92	-0.05	0.92	1.74	3.26
Q4	Common foreign policy	-0.66	-0.65	-0.50	-0.09	2.43
Q6	EU Membership*	-0.82	0.05	1.19	2.08	3.50

Table 5.10: catPCA Quantifications for the DSV EU scale in Ireland. Questions with an asterisk (*) were reversed in the scale.

disagree" categories are still extreme and in most cases, the centre is not given by the "neutral" category. For question 6, all the categories from "completely disagree" to "agree" indicated disagreement with the question, with the opposite being the case for question 4. Questions 1 and 3 seem to have the neutral point somewhere between the "agree" and "neutral" categories, while in the case of question 2 it is between the "neutral" and "disagree" categories.

	Topic	CA	A	N	D	CD
Q1	Euro*	-3.80	-1.54	-0.87	0.13	0.71
Q2	Treaty change*	-2.96	-0.47	0.02	0.64	1.06
Q4	Common foreign policy	-2.29	-0.11	0.43	0.80	0.98
Q6	EU Membership*	-3.82	-2.08	-1.17	0.05	0.74
Q7	Treaties	-1.23	0.42	1.11	1.34	1.34
Q18	Water*	-1.92	-0.20	0.44	0.80	1.07
Q19	Tax	-2.06	-0.37	0.25	0.74	1.04

Table 5.11: catPCA Quantifications for the Quasi-Inductive EU scale in Ireland. Questions with an asterisk (*) were reversed in the scale.

In its quasi-inductive version, the EU scale emerged as one of three similarly strong scales and gained two unrelated questions. Table 5.11 shows the quantifications and reveals only a single tie at question 7. Once again, the "neutral" category is rarely found in the middle, while the distances between the separate categories differ considerably at some points. Still, the DDI score of 0.28 indicates we can still speak of "good" data.

Turning now to the MCA plots for the scales, Figure 5.27 shows the results for the original scale. The figure on the left, showing the first and second dimensions, reveal a similar pattern as we saw earlier in the case of Lithuania and have the same problematic interpretation. Basically, the first dimension a cluster of "completely disagree" categories from the others, while the second dimension runs from No Opinion through "agree" — "neutral" — "disagree" — "completely disagree" — "completely agree". The third dimension meanwhile, shown on the right side of the figure seems to the No Opinion categories from the others. But while this is very strong for question 3, the other questions, especially question 8, are still quite close to the other categories, indicating that a large amount of switching took place. Furthermore, all the No Opinion questions are positively related to the negative side of dimension 2 - the more negative the dimension the greater the distance. Both indicate that the No Opinion option for especially question 3 ("The right of EU citizens to work in Ireland should be restricted") was rare.

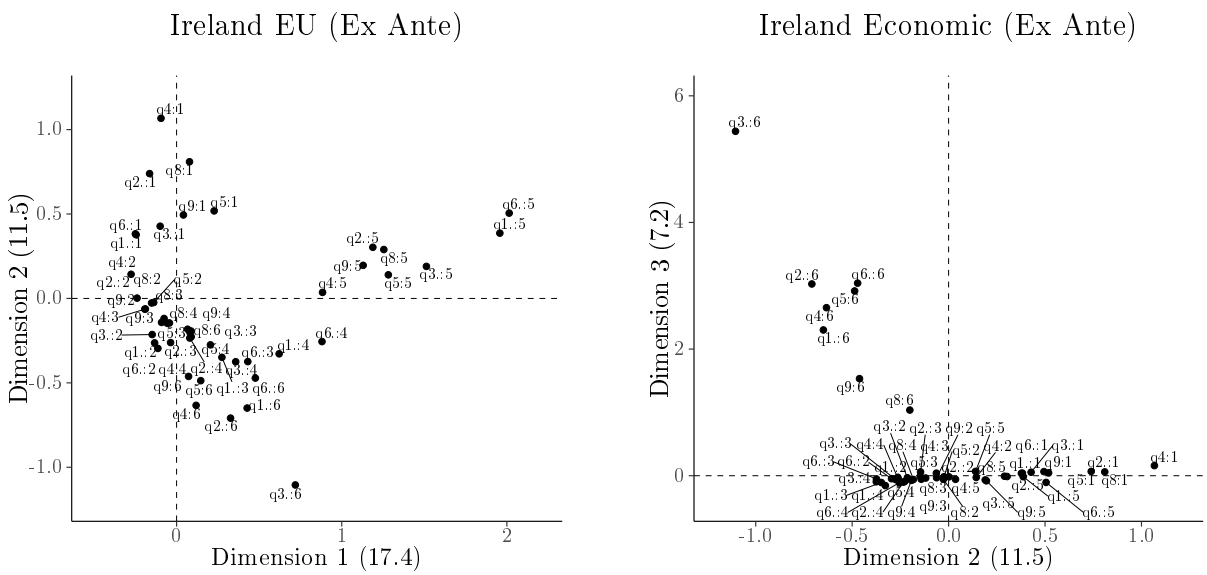


Figure 5.27: MCA Plots for the Original scale in Ireland, with the 1st and 2nd dimension on the left, and the 2nd and 3rd dimension on the right

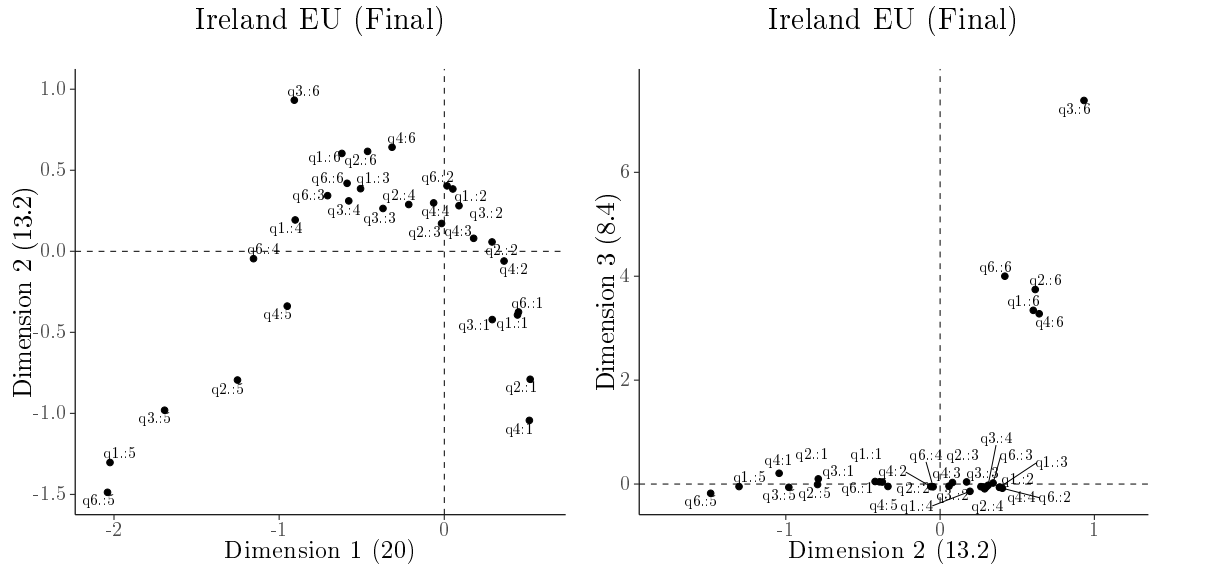


Figure 5.28: MCA Plots for the DSV scale in Ireland, with the 1st and 2nd dimension on the left, and the 2nd and 3rd dimension on the right

The DSV version of the scale, of which Figure 5.28 shows the MCA, fares much better. While the neutral categories are not in the centre, we can distinguish a horseshoe shape. So, the first dimension holds the order of the categories intact and runs from "completely disagree" to "completely agree", with the No Opinion categories somewhere in the middle. The second dimension than the extreme categories from the less extreme and the No Opinion categories. The third dimension furthermore the No Opinion categories from the others, though these categories do show a relation with the positive side of dimension 2, meaning they are associated with the less extreme categories.

The quasi-inductive scale, the MCA of which is shown in Figure 5.29. Here, the horse-

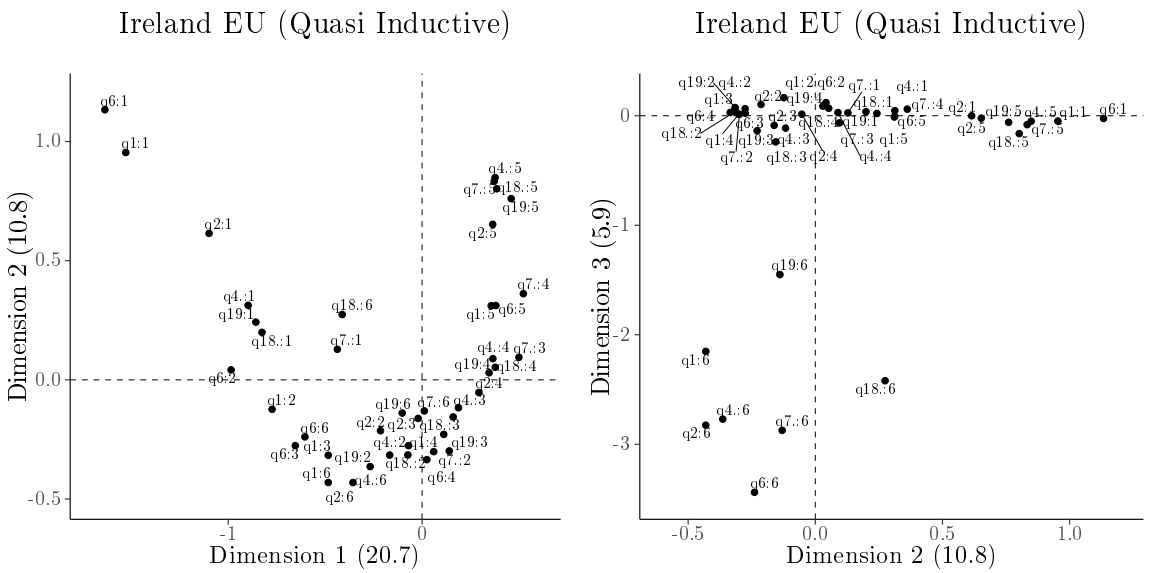


Figure 5.29: MCA Plots for the Quasi-Inductive scale in Ireland, with the 1st and 2nd dimension on the left, and the 2nd and 3rd dimension on the right

shoe shape is even more clear and the ordinality of all the questions is correct, except for the final two categories of question 7, which I already saw indicated when studying their quantifications. The third dimension also the No Opinion categories from the other, with these categories have a slight relation with the non-extreme side of Dimension 2, except for question 18, which seems associated with the extreme side.

The EU dimension in the Irish case presented a more optimistic picture than the one painted earlier for Lithuania. While the EU dimension was problematic initially, both the DSV and quasi-inductive version of the scale improved upon the original. Moreover, while there were nonetheless some ties (which increased the DDI somewhat) the underlying structure of the data seems to be in order. One point of interest about the quasi-inductive scale is how two unrelated questions are included into the scale, with both having H values of 0.36. On closer inspection, yet, the question on the water tax is related to the EU dimension because of the relation of the tax to the EU water framework, while the question on the local property tax is related to the EU dimension as the incomes in the tax were significant for the partly EU sponsored recovery programme after the financial crisis of $2007-2009$. This shows that MSA cannot only be used to ensure unidimensional scales, but also to discover which questions belong together to gain a better understanding of the structure of the political space during elections.

5.3 Hungary

Hungary presents an intriguing case. On its EU dimension, the original scale scored a DDI of 0.20, while its DSV version only scored 0.26. After the DSV it was the only dimension that remained with a DDI of 0.24. Its H values were high, with the original scale scoring 0.42, the DSV scale 0.52, and the quasi-inductive scale 0.48. Especially interesting is that the

0.10 increase in the H value between the original and DSV scale was the result of switching question 5 with question 11. Table 5.12 shows this more clearly. Here we see that question 5 in the original scale negatively covaries with the other questions in the scale. Put different, one's opinion on this question is the opposite of their position on the main EU dimension. Eliminating this question thus improves the original scale. The other question of interest — question 8 — is noteworthy because in both cases it barely makes the 0.30 cut-off mark. But its value of 0.31 in the original scale and 0.34 in the DSV scale are altogether different from the other values.

#	Question	Original H_i	DSV H_i
Q1	Hungary should never adopt the Euro	0.53	0.58
Q2	A single member state should be able to block a treaty change, even if all the other member states agree to it	0.43	0.48
Q3	The right of EU citizens to work in Hungary should be restricted	0.46	0.51
Q4	There should be a common EU foreign policy even if this limits the capacity of Hungary to act independently	0.50	0.57
Q5	The EU should redistribute resources from richer to poorer EU regions	−0.10	–
Q6	Overall, EU membership has been a bad thing for Hungary	0.55	0.62
Q7	EU treaties should be decided by the Hungarian parliament rather than by citizens in a referendum.	–	–
Q8	To address financial crises, the EU should be able to borrow money just like states can	0.31	0.34
Q9	European integration has gone too far	0.52	0.58
Q10	The Hungarian Land Law should be suspended	–	–
Q11	The policy of 'eastern opening' harms Hungary's relations with the EU	–	0.46
	Scalability Coefficient (H)	0.42	0.52

Table 5.12: Loevinger's H values for the original and DSV scales in Hungary

This is further shown in Figure 5.30. Here, the questions 8 and 5 are separated from the others and are perpendicular to each other. In other words, the position of a user on whether the EU should be able to borrow money like states can is unrelated to whether the EU should redistribute resources. Moreover, while question 8 is still related to the other questions, question 5 is unrelated to the other questions as well. The rest of the questions are all tied together on the first dimension, indicating they belong together.

Turning to the quantifications in Table 5.13, I find no evidence of ties and a neutral category that seems to be predominantly in the middle. The only other point of note is the "completely disagree" categories are rather far from the centre, though the values are not

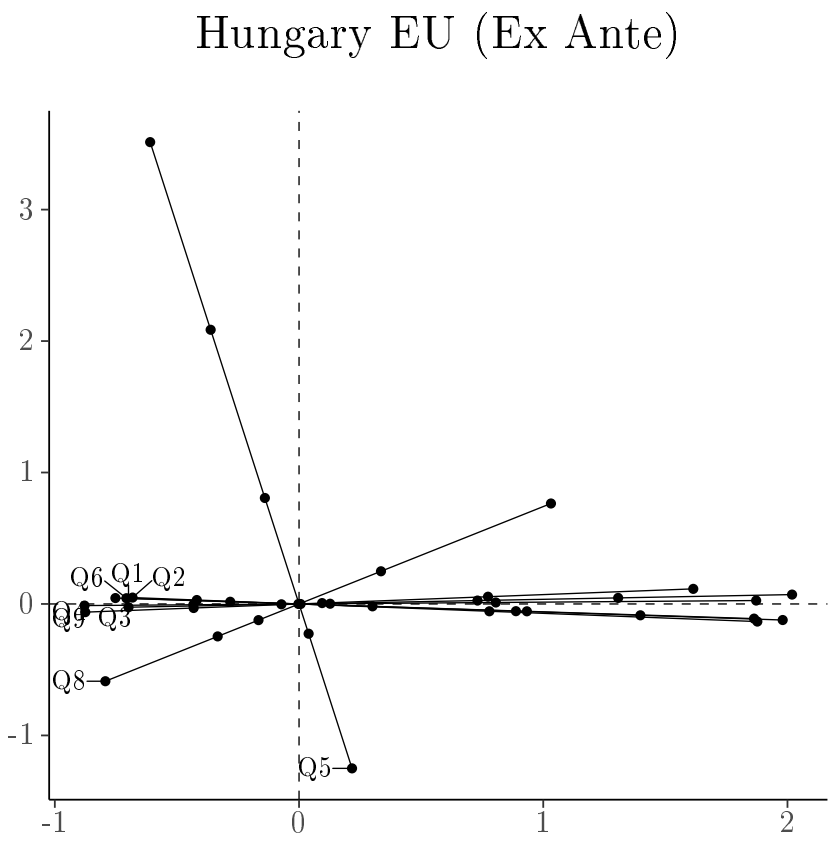


Figure 5.30: Hungary - catPCA Biplot (Original EU Scale)

	Topic	CA	A	N	D	CD
Q1*	Euro Adoption	-0.85	−0.09	0.89	1.59	2.45
Q2*	Block Treaty Change	−0.95	−0.59	−0.00	1.09	2.62
Q3*	Right to work	−1.01	−0.38	0.40	1.25	2.51
Q4	Common Foreign Policy	−1.09	−0.54	0.12	0.96	2.01
Q5	Redistribution	−1.34	−0.24	0.87	2.24	3.77
Q6*	EU Membership	−0.81	0.01	1.02	1.60	2.26
Q8	EU borrowing	−1.57	−0.66	−0.33	0.67	2.05
Q9*	EU Integration	−1.05	−0.52	0.15	0.96	2.23

Table 5.13: Hungary - catPCA Quantifications (Original EU Scale)

as extreme as those I saw in the case of Lithuania. Based on this evidence, it would seem admissible to not only remove question 5 from the set of questions but to remove question 8 as well, even though its H value is sufficient.

We can see the reason for this can in the left plot of Figure 5.31, which shows the biplot for the DSV scale. Here, we find that the new question 11 correlates with question 8 - the reason why it was likely included. Yet, in combination with question 8, it belongs to a cluster separate from the other questions, undermining the unidimensionality of the scale. Turning to the quasi-inductive scale on the right side of Figure 5.31, I note that here most of the questions cluster together to produce a stable bundle of questions.

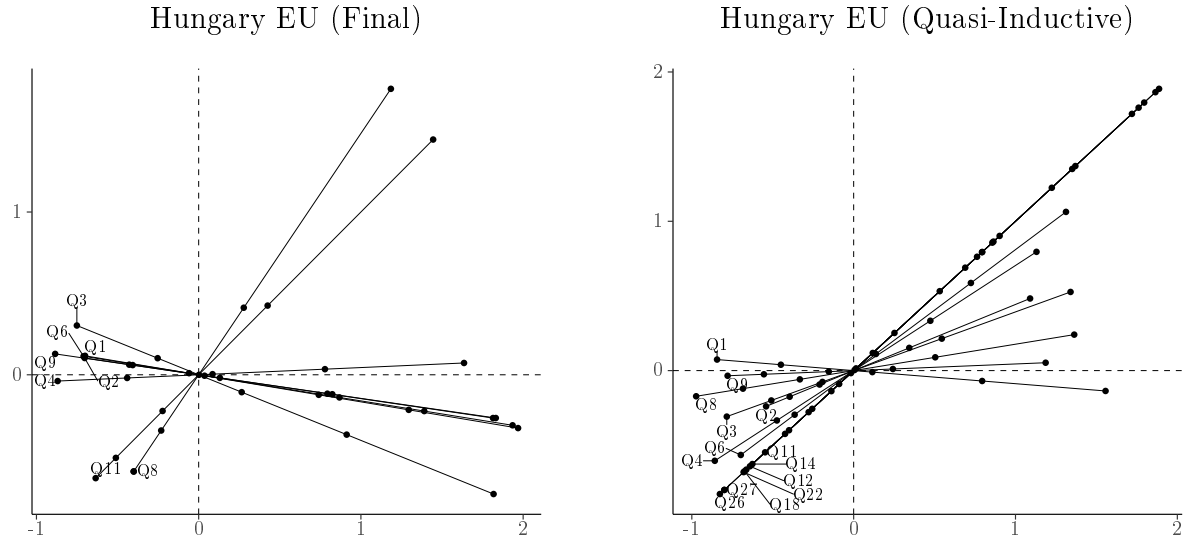


Figure 5.31: Bilots for the DSV and Quasi-Inductive EU scale in Hungary

Turning to the MCA for the first and second dimension in Figure 5.32, we find it are the questions 5 and 8 that do not live up to the ordinality of the questions. When plotted on the first dimension, the categories for question 5 run as $3-4-2-1-5-6$. The categories for question 8 do run in order, except for the "completely disagree" category, which does seem unrelated to the other categories of that type. For both the DSV and quasi-inductive scales, the ordinality of each of the questions is in order, and we can see a clear horseshoe. Moreover, in both cases, the first dimension explains a significant amount of the variation in the data (32.8% for the DSV version and 38.2% for the quasi-inductive version, while the second dimension separates the extreme from the non-extreme categories.

To understand if the "neutral" and No Opinion categories are separated from each other or whether they have been used interchangeably, we have until now made use of the third dimension, which shows this distinction. While this is also the case for all three scales in Hungary, we can also assess this distinction using a subset multiple correspondence analysis (SMCA). Figure 5.33 shows these, with only the No Opinion and "neutral" categories shown. Starting with the original and DSV scales, I find that especially in the case of the problematic question 8 the category is close to its "neutral" category, indicating that users might have used this latter category to hide their possible No Opinion. In the case of the quasi-inductive scale, I find most of the questions well separated from each other, with the possible exception of — again — question 8 and question 22.

Hungary thus provides evidence of a well-defined EU scale. Moreover, I saw how a biplot can identify the problematic questions, even after they have been placed into the scale based on their H value. This again shows that using automated algorithms like the ones employed by MSA is not without risks and that we should include no question for the reason that its values are sufficient. Instead, especially when the value for a question substantially deviates from the others we need clear reasons as to why to include the question. Here, this would have led to this removal of not only question 5 but also question 8 from the original scale.

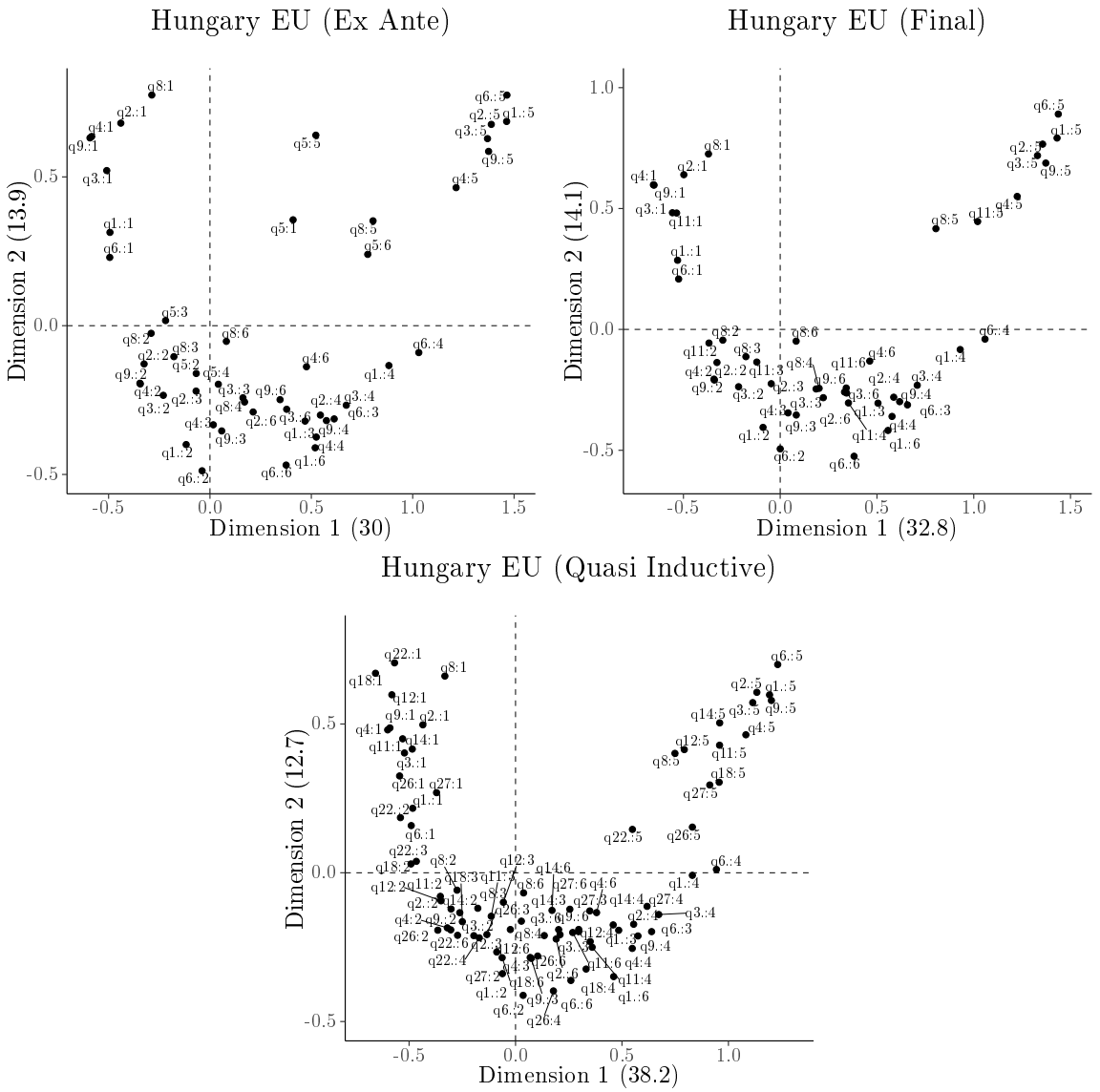


Figure 5.32: MCA Plots for the Original, DSV and Quasi-Inductive EU scale in Hungary, 1st and 2nd dimension

This, in turn, would also have led us to not include question 11.

5.4 Estonia

I now turn to Estonia, which displayed the most unsatisfactory performance on all the indicators. To see where the problem lies, I take a look at its most problematic Cultural dimension. This dimension scored a DDI of 0.52 on the original scale, 0.60 on the DSV scale, and 0.39 on the quasi-inductive scale. Here, I will identify for each of these scales where the problems lie and how we should address them and whether we should bother about the dimension as a whole.

Let us start once more by considering the data as if they were metric. To picture what this leads to, Figure 5.14 shows the scree-plot that would emerge after I would run a PCA

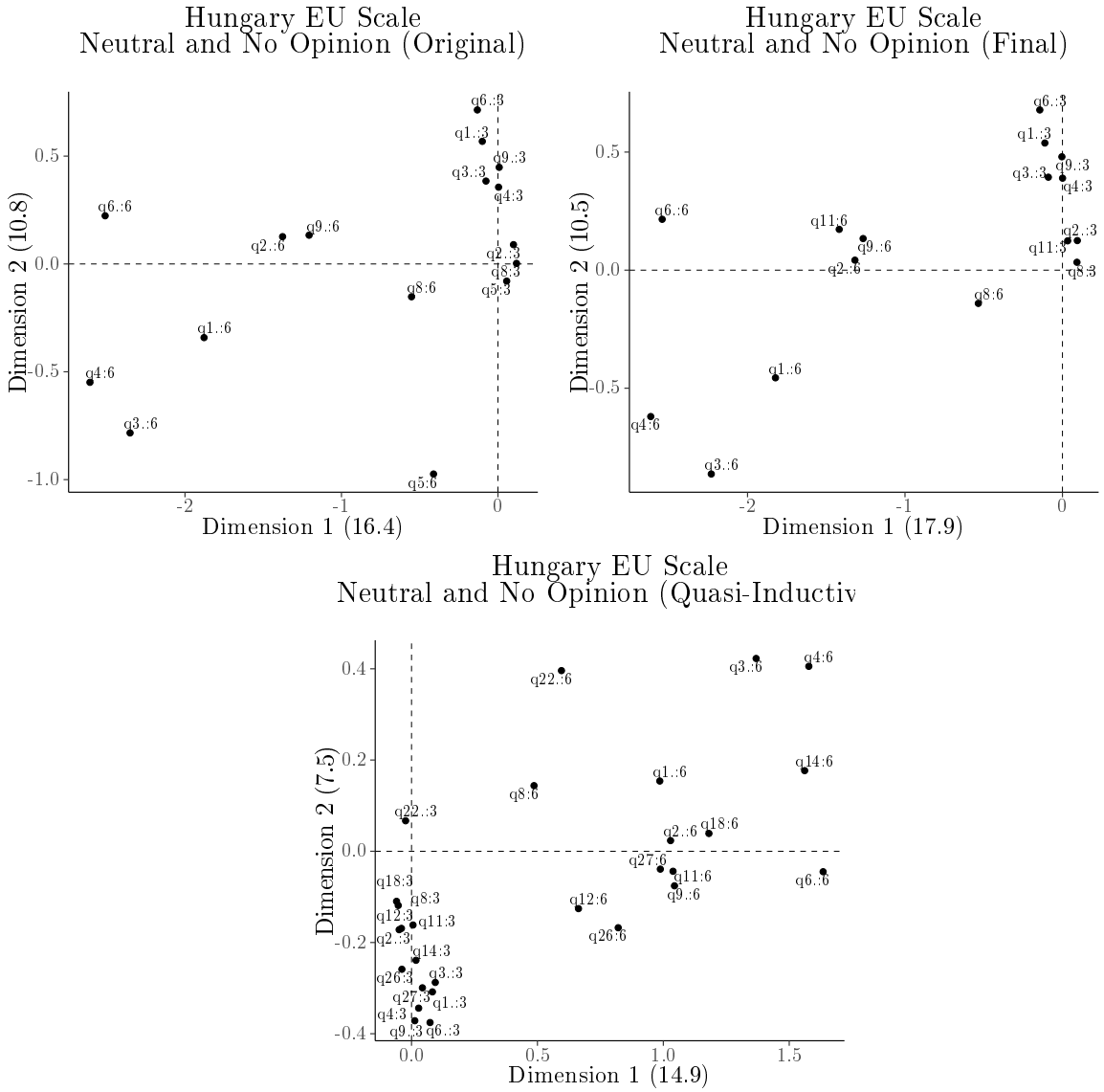


Figure 5.33: SMCA for the Original, DSV, and Quasi-Inductive EU scale in Hungary

on the scale. As there are 8 questions in the scale, there are also 8 possible dimensions. A dotted line shows where the eigenvalue equals 1. The custom to only include eigenvalues larger than 1 is better known as the Kaiser criterion. Here, this would lead to us recognizing three factors underlying the data. Yet, if we look at the shape of the scree plot, we see that we find the "elbow" in the second dimension. The values in Table 5.14 substantiate this. The difference between the first and second dimension is larger than the difference between the second and third dimensions (and all the dimensions after that). Whether to include the second dimension is up for debate, though for visualization purposes I will include it from here on. Note though that from a PCA point of view the underlying dimension is already problematic as I assume the scale has a single dimension only.

I now move on to look at the PCA factor loadings, which Figure 5.36 shows in their rotated and unrotated version, with the values also given in Table 5.15. Here, we notice two things. First, that there are two clusters of questions, each loading on its own, and second,

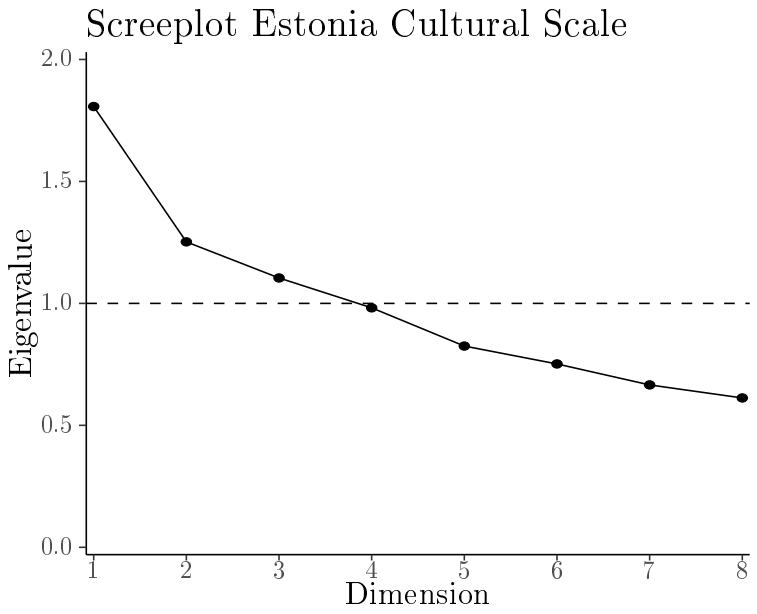


Figure 5.34: Estonia - Screeplot (Original Cultural Scale)

Dimension	Eigenvalue (λ)	% of Variance
1	1.81	22.59%
2	1.25	15.65%
3	1.10	13.80%
4	0.98	12.27%
5	0.83	10.32%
6	0.75	9.40%
7	0.67	8.32%
8	0.61	7.65%

Table 5.14: Estonia - Eigenvalues and Variance (Original Cultural Scale)

that the loadings are not very large. This is especially the case for question 23, which does not load well on both of the dimensions. As the rotation seems to do a good job of making the structure of the data clearer, I will concentrate on these values from here on. To begin with, questions 25,24,26 and 23 load on the second dimension. These questions all relate to topics that deal with progressive issues, such as marriage, abortion and cannabis possession. The fourth question, 23, does load on this dimension, albeit very little. The first dimension comprises the other questions. These questions are all closer together, indicating a larger amount of cohesiveness on this dimension. This dimension seems to deal with immigration, citizenship, privacy and public order and can be said to cover conservative issues.

Conducting a similar analysis using catPCA - as shown in Figure 5.38 and Table 5.16 reveals that according to catPCA, the main structure of the data is altogether different in its unrotated form. Most interesting is that the explained variance for the second dimension has increased from 15.65% to 19.68% in the unrotated solution. This indicates the PCA is missing out on a large amount of the variation that the catPCA did capture. For the rotated solution both of the dimensions pick up more variance than in their unrotated counterparts.

The key difference between the PCA and catPCA solutions is that in its unrotated version,

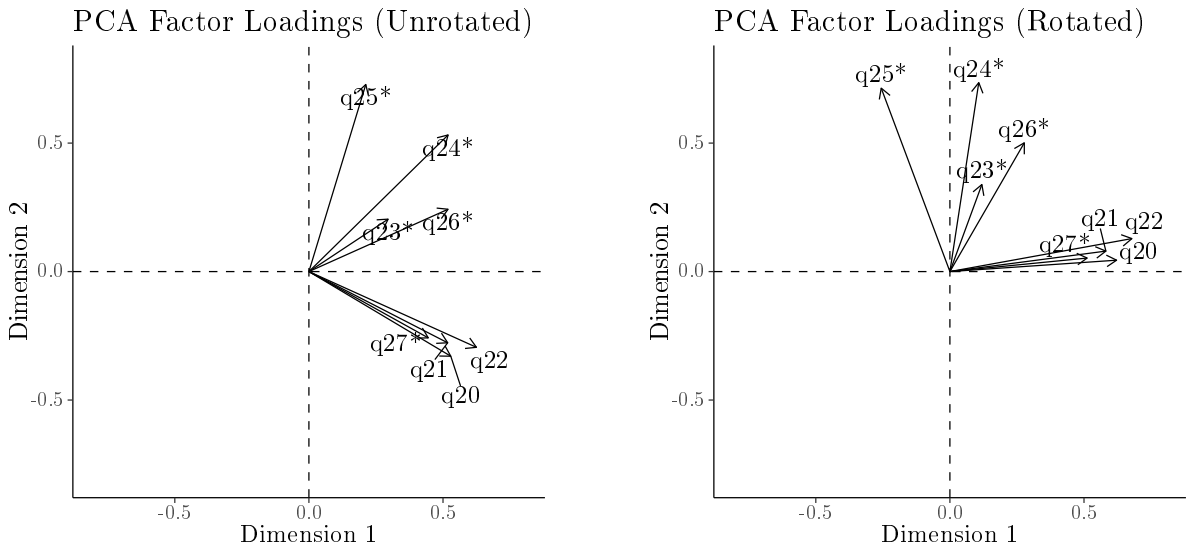


Figure 5.36: Estonia - Unrotated and Rotated PCA (Original Cultural Scale)

	Topic	Unrotated		Varimax Rotation	
		Component 1	Component 2	Component 1	Component 2
Q20	Immigrants	0.53	−0.33	0.62	0.04
Q21	Privacy	0.52	−0.28	0.58	0.08
Q22	Public Order	0.62	−0.29	0.68	0.13
Q23*	Community	0.30	0.20	0.12	0.34
Q24*	Same Sex	0.52	0.53	0.11	0.74
Q25*	Abortion	0.21	0.73	−0.26	0.71
Q26*	Cannabis	0.52	0.24	0.28	0.50
Q27*	Citizenship	0.45	−0.26	0.51	0.05
Eigenvalue		1.81	1.25	1.62	1.44
Variance		22.59%	15.65%	20.19%	18.05%

Table 5.15: Estonia - Unrotated and Rotated PCA (Original Cultural Scale)

the questions of the second dimension correlate negatively with the questions on the first dimension, while this was not the case in the PCA solution. In the rotated version, not only question 25 but also question 23 now correlate negatively with the first dimension. Moreover, the loadings are higher for all the questions, especially for question 23, which did not load very strong on any dimension in the PCA. This again indicates PCA is ignoring certain information about this question that catPCA is picking up.

Figure 5.39 shows the unrotated solution as a biplot. This reveals that the clusters of questions and several categories are very close together as the points on the axes of the categories are distributed unevenly over the lines. Moreover, note that as in the above loading plots for the catPCA, I can identify three, instead of only two clusters of questions. Most important is that question 23, located in the PCA between questions 24 and 26, is now associated with question 25, establishing a subcluster of sorts. Given that for both questions only the endpoints are visible, this means that the middle categories are very close together near the centre, pointing at some severe problems.

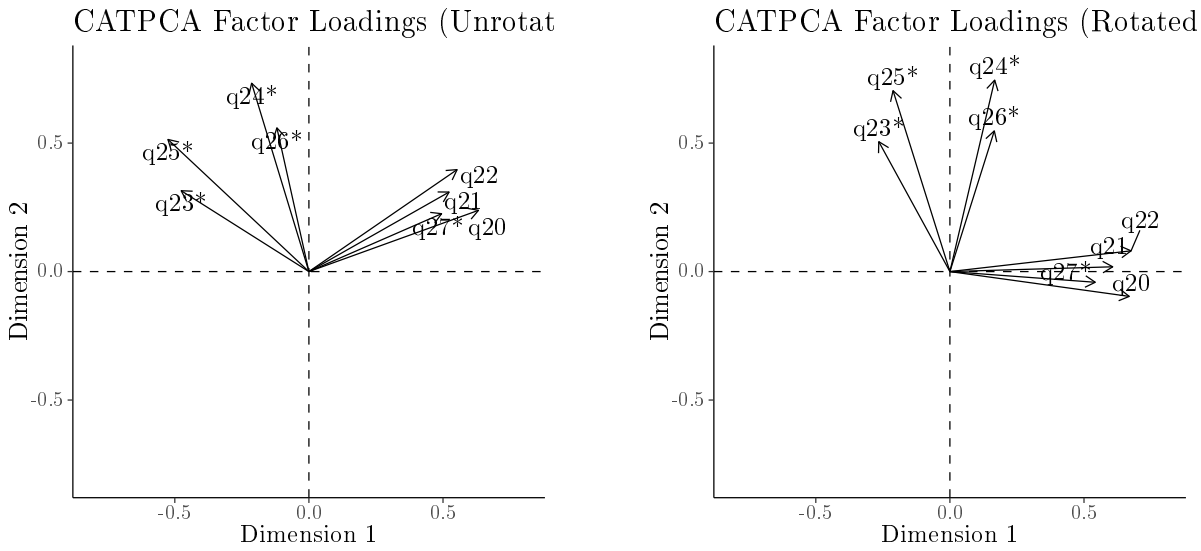


Figure 5.38: Estonia - Unrotated and Rotated catPCA (Original Cultural Scale)

	Topic	Unrotated		Varimax Rotation	
		Component 1	Component 2	Component 1	Component 2
Q20	Immigrants	0.63	0.24	0.67	−0.10
Q21	Privacy	0.52	0.31	0.61	0.02
Q22	Public Order	0.55	0.40	0.68	0.08
Q23*	Community	−0.48	0.31	−0.27	0.51
Q24*	Same Sex	−0.21	0.73	0.17	0.75
Q25*	Abortion	−0.53	0.51	−0.21	0.70
Q26*	Cannabis	−0.12	0.56	0.17	0.55
Q27*	Citizenship	0.49	0.23	0.54	−0.04
Eigenvalue		1.79	1.57	1.74	1.62
Variance		22.33%	19.68%	21.71%	20.30%

Table 5.16: Estonia - Unrotated and Rotated catPCA (Original Cultural Scale)

This is revealed in Table 5.17 and Figure 5.40, where I present the quantifications of each of the questions and plot them against their original categories. Apart from question 26, each of the questions has multiple ties (or is very close to this, for example, question 22). Worst-case scenarios are questions 23 and 25 (and to a lesser degree 24), which have four tied categories, all on the negative side. This means that all the categories up till that point meant the same to the user and that they did not distinguish between them. We find the opposite effect for the other questions. Given the fact that the quantifications seem to distinguish between two levels only, it seems like the users simplified the five possible response options to a binary version.

Before I turn to the MCA plots, I will focus on a technique I have not discussed so far. This technique is related to the IRT origins of Mokken scaling and refers to the Item Response Functions (IRF) and the Item Step Response Functions (ISRF). The idea behind these is simple. If the question belongs to a true Mokken scale, then both the IRF and the ISRF will monotonically increase. This means that for a higher score on the underlying latent variable,

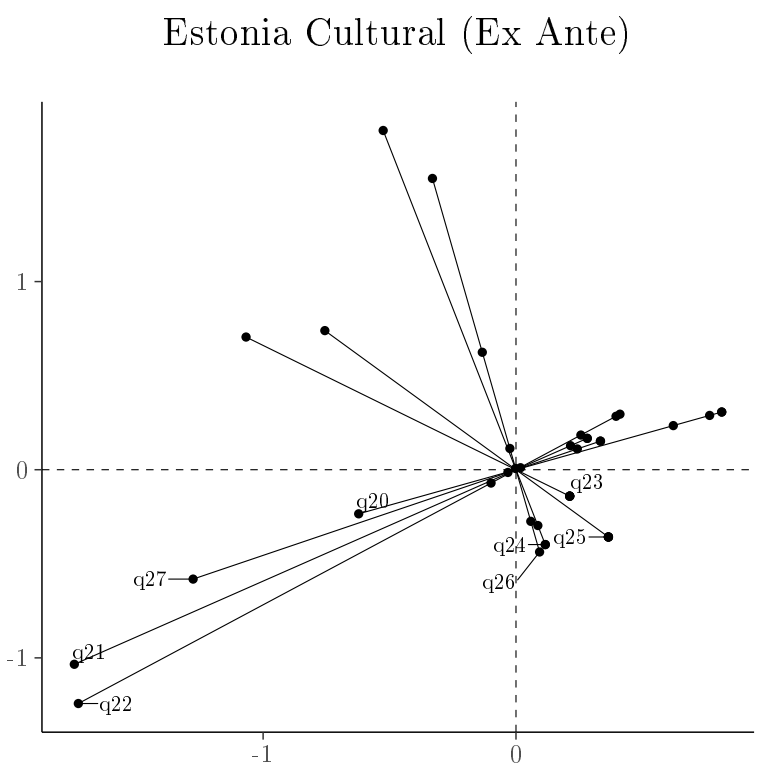


Figure 5.39: Biplot for the Original Cultural Scale in Estonia

one will have a higher chance of obtaining a high score on that variable. High and low here depends on the orientation of the variable. Here, I have a scale on which a higher score indicates a progressive standpoint, while a low score indicates a conservative standpoint. As the question is reversed, the higher one's score on the question "The recreational use of cannabis should be legal", the more progressive one should be. To check whether this assumption of monotonicity holds, I can plot the IRF. This IRT should be roughly S-shaped, meaning one would have a low response when the score on the scale is low, while one would have a high response when the score on the scale is high.

The IRF for question 25 is shown on the right-hand side in Figure 5.41. Here, the axis shows the rest score, which is the score of the user on all the other questions in the scale, while the y-axis shows the score of the user on that question, ranging from $0-4$. The reason it ranges from $0-4$ is that there are four "steps" on a five-point scale. Therefore, a value of 1 indicates the step between the "completely disagree" and the "disagree" option, 2 indicates the step between the "disagree" and the "neutral" option, and so further. From this graph, it emerges that question 25 is problematic. It seems that whatever the score of the user on the scale, the response of the user to the question is stable, only going up when users have a very high score on the scale.

As the question is polytomous, we can also look at the categories themselves. We do this with the ISRF, shown on the left-hand side in Figure 5.41. Once more, we note only four lines here, each one referring to a "step" between the five categories. Ideally, we would want to see four S-shaped lines, each starting higher than the other, being parallel to each other. This

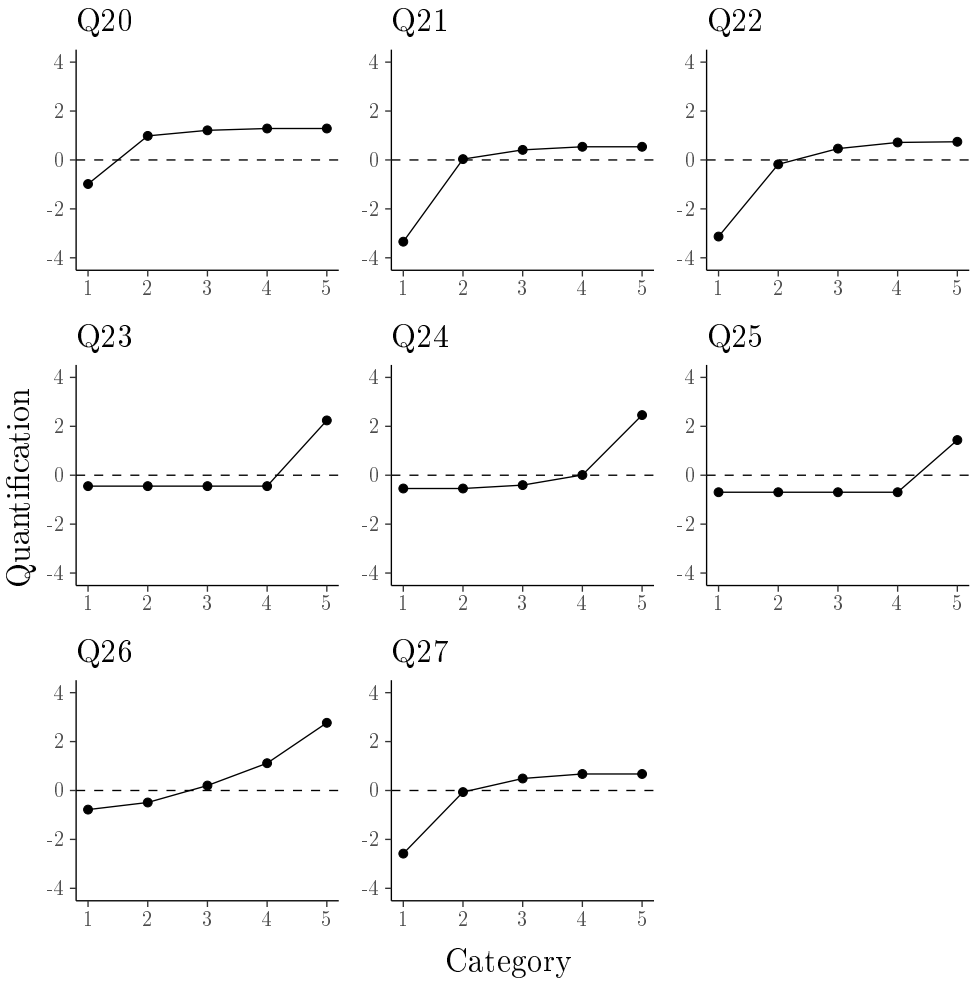


Figure 5.40: Estonia - Transformation Plot (Original Cultural Scale)

	Topic	CA	A	N	D	CD
Q20	Immigration	−0.98	0.98	1.21	1.29	1.29
Q21	Privacy restrictions	−3.34	0.03	0.41	0.54	0.54
Q22	Restrict demonstrations	−3.13	−0.18	0.46	0.72	0.74
Q23*	Community Service	−0.45	−0.45	−0.45	−0.45	2.24
Q24*	Same Sex Couples	−0.54	−0.54	−0.40	0.01	2.46
Q25*	Abortion	−0.70	−0.70	−0.70	−0.70	1.44
Q26*	Cannabis Use	−0.78	−0.49	0.20	1.12	2.77
Q27*	Citizenship	−2.58	−0.06	0.49	0.68	0.68

Table 5.17: Estonia - catPCA Quantifications (Original Cultural Scale)

would indicate that for the lower restscores, the probability of selecting the highest category is low, while it is higher for each lower category. Instead, we see that four the highest three steps are horizontal and that there is no relation between them and the position of the user on the scale. The first step meanwhile is even more problematic, as it would mean the chance of moving from the "completely disagree" to the "disagree" category would first decrease (instead of increase), while later on, it would increase. This behaviour, together with the behaviour of IRF, partly explains the low H value of question 25 of 0.08 and the high crit value of 143.

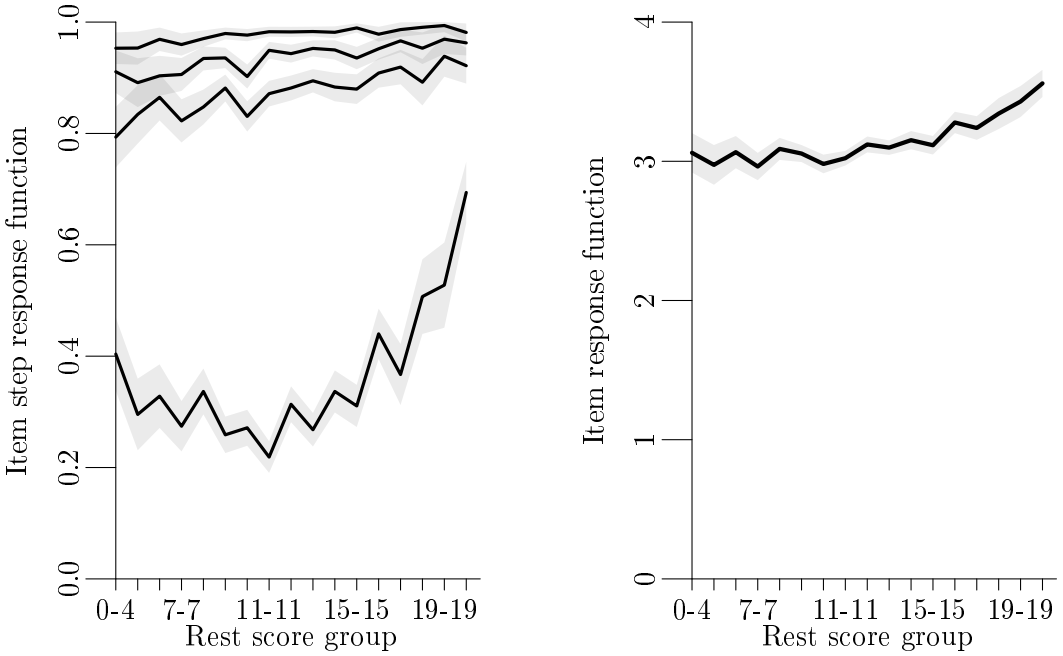


Figure 5.41: Item Step Response Function (Left) and Item Response Function (Right) of Question 25

The fact that the whole scale is problematic is further when we look at the MCA for the first two dimensions as shown in Figure 5.42 where, if we are very willing, we can see a horseshoe shape on the second dimension. Looking at the first dimension, we see that runs from the extreme responses on the left to the non-extreme responses on the right. The second dimension has the categories from high to low. Yet, how it does so depends a lot on the individual questions. For example, projected on the second dimension, we see that for question 20, the categories 5,4, and 3 are close to each other. This means that users switched between these questions and that there was in their eyes little difference between them. Moreover, note that each of these categories is quite far from the $(0, 0)$ centre of the plot, while similar categories for the questions 23 and 25 were quite close.

We can see further problems in Figure 5.43 where I have plotted the second and third dimension. We see that the third dimension the No Opinion categories from the "neutral" categories, the latter of which all load negatively on this axis. This means that users either opted for a "neutral" or a No Opinion response, and rarely switched between them. While this may seem beneficial, another interpretation of the third dimension is between three groups: neutral, completely agree/agree/disagree/completely disagree, and no opinion (except for category 5 for question 20). This would mean that the users see "neutral" as an alternative to No Opinion and not as belonging to the other categories and hence is different from them.

In whatever way we look at it, the original cultural scale for Estonia is problematic. We can find many bundles of questions, the quantifications show many ties, and we cannot interpret the first dimension as the underlying dimension. Thus, any position of the user on this dimension relates us very little of what that person actually thinks about cultural issues.

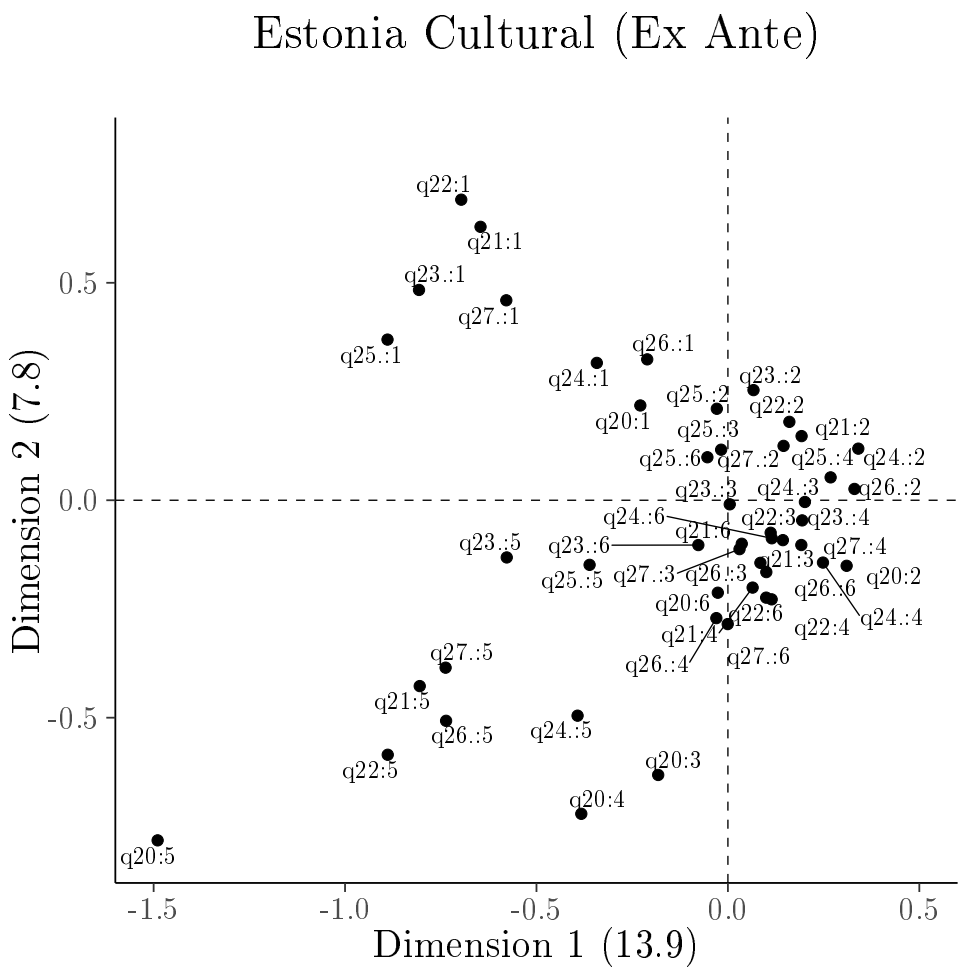


Figure 5.42: MCA for the first and 2nd dimension of the original Cultural scale in Estonia

	Topic	CA	A	N	D	CD
Q24*	Same Sex Couples	−2.48	0.09	0.40	0.51	0.51
Q25*	Abortion	−1.44	0.70	0.70	0.70	0.70
Q26*	Cannabis Use	−0.89	−0.86	−0.74	−0.46	1.50

Table 5.18: Estonia - catPCA Quantifications (DSV Cultural Scale)

It was thus no surprise that for the DSV scale a considerable number of questions were deleted. Still, the total H value was still low at 0.23, while also none of the individual questions reached the benchmark of 0.3, while 2 of the questions also had crit values > 80. Moreover, looking at the quantifications in Table 5.18, I find that question 24 has a tie at the "disagree" and "completely disagree" categories and a centre that lies somewhere between the "agree" and "completely agree" categories. Question 26, on the other hand, has the centre between the "disagree" and "completely disagree" categories, and while having no ties, comes very close to it between the CA and A categories. Question 25 has three ties, which means that the A, N, D and CD categories conveyed an identical meaning to the user. Moreover, for each of the three questions, the distances between the categories differed, especially for the CA category in question 24. In all, these problems explain the low DDI and show the

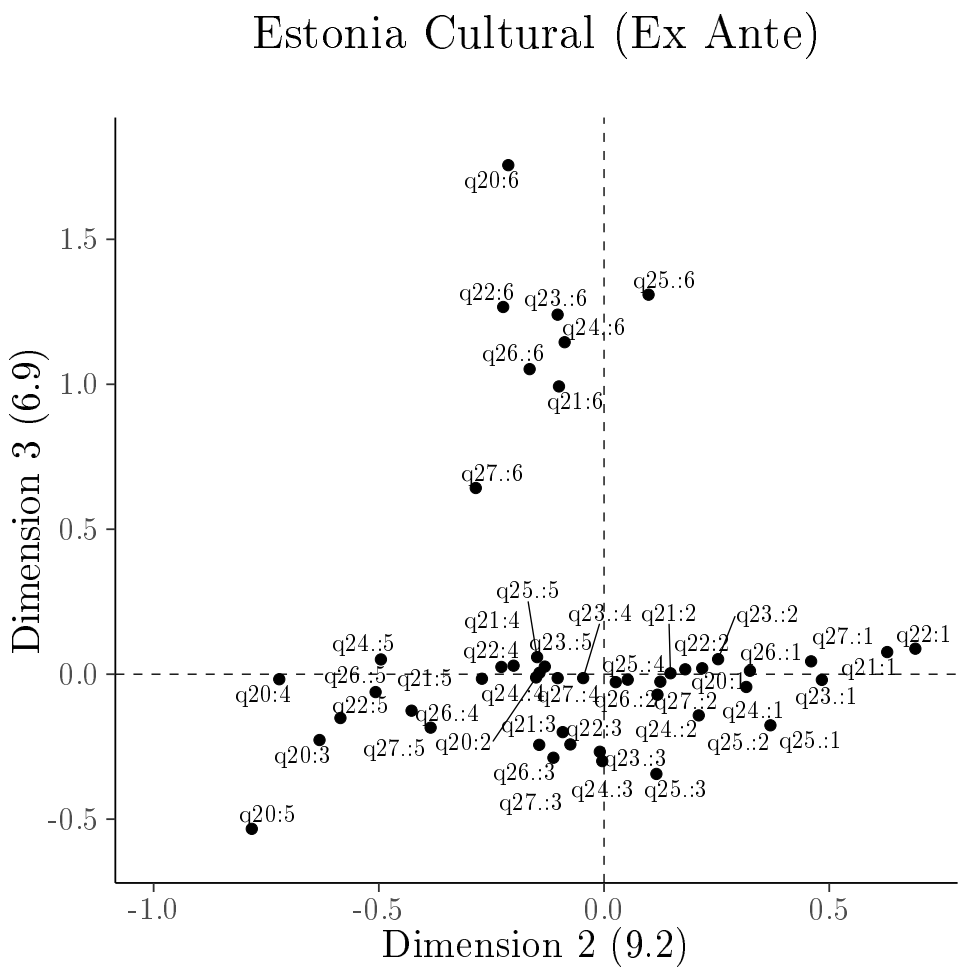


Figure 5.43: MCA for the 2nd and 3rd dimension of the original Cultural scale in Estonia

data is far from metric.

The problems continue when I look at the MCA. Starting with the first and second dimension - shown left in Figure 5.44, I can best describe the resulting shape as a under a 45° angle with the first dimension. Projecting the categories on this first dimension shows some serious violations of the order of the categories. For example, the categories from question 25 run as $1-5-4-3-2$, those from question 24 as $1-5-2-3-4$ and from question 26 as $1-2-5-3-4$ (ignoring the No Opinion categories). This means the ordinality of the questions was non-existent for each of the questions. Moreover, we cannot interpret this dimension in any significant way. The same goes for the second dimension, which seems to serve to distinguish category 5 from question 25 from the others. When we project the categories on this axis, the ordinality is maintained for the questions 25 and 24, but not for question 26. However, the distances between these categories differ. The third dimension, meanwhile, is interpretable and can be seen as the No Opinion responses from the other responses.

Apart from the original and DSV scales, even the quasi-inductive scale contained problems. While the MSA resulted in a scale that had a H value of 0.36, the questions included seem to refer only to a small part of the original cultural scale and seem to deal with na-

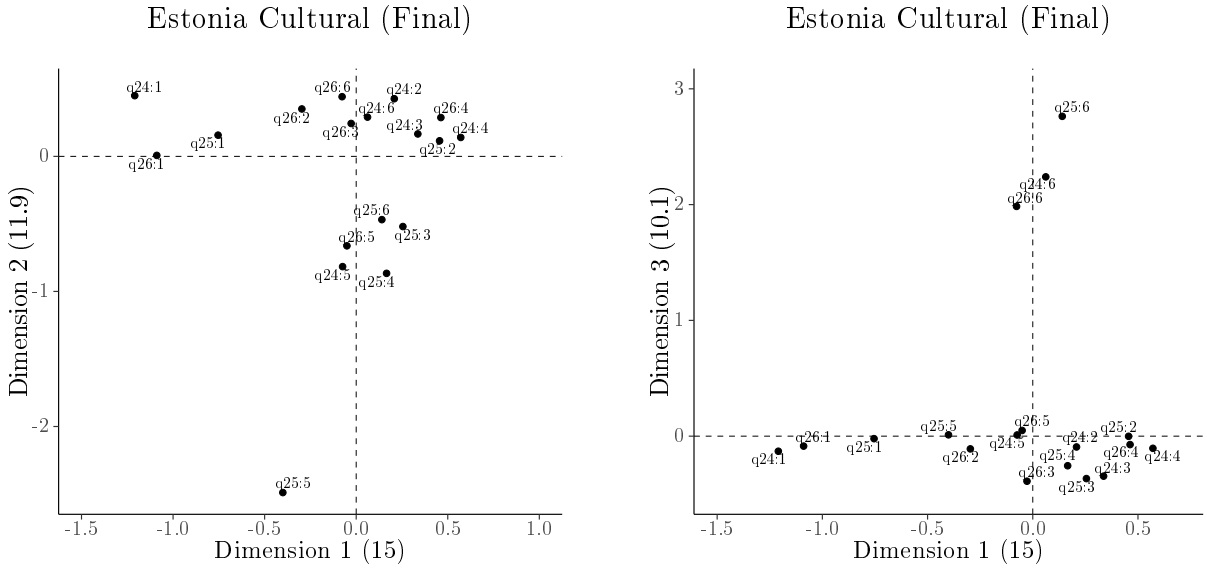


Figure 5.44: MCA for the 1st and 2nd (left) and 1st and 3rd (right) dimension of the DSV Cultural scale in Estonia

	Topic	CA	A	N	D	CD
Q9	Russia Sanctions	−0.44	−0.20	0.22	1.05	4.31
Q20	Immigration	−0.62	0.34	1.76	3.36	9.25
Q27*	Citizenship	−0.58	−0.51	−0.27	0.18	3.62

Table 5.19: Estonia - catPCA Quantifications (Quasi-Inductive Cultural Scale)

tionality and immigration issues. Looking at the quantifications in Table 5.19, I find no ties (explaining the lower DDI of 0.39 - which is still high), but some categories come very close to being ties (the CA and A categories of question 27 for example), and the distances between the categories differ. Moreover, question 20 is still unbalanced in that we can find the middle somewhere between the questions CA and A.

Turning again to the MCA I find a horseshoe-shape on the first and second dimension and a correct ordinality for all the questions. The only problem here is the ordinality of question 9 is incorrect as the ordering (from left to right on the first dimension) should be the same but is in fact mirrored. This means the users did not note this during the filling out of the questionnaire. The third dimension meanwhile seems to separate the No Opinions from the other categories. But especially in the case of question 27, the No Opinion category is very close to its "neutral" category indicating that users switched between the two. Moreover, question 20 has a category that seems to be on the opposite side of the dimension than its No Opinion category, making this interpretation of the third dimension problematic.

In all three its versions, the cultural scale of Estonia is problematic. When taken as a scale, a score on the scale tells little about the user's true position on cultural issues. Moreover, there are many pieces of evidence for users having simplified the response options into a simpler binary version, having used the "neutral" option as a way to hide non-response, having missed the reversal of certain questions, and not observed the correct ordinality. This

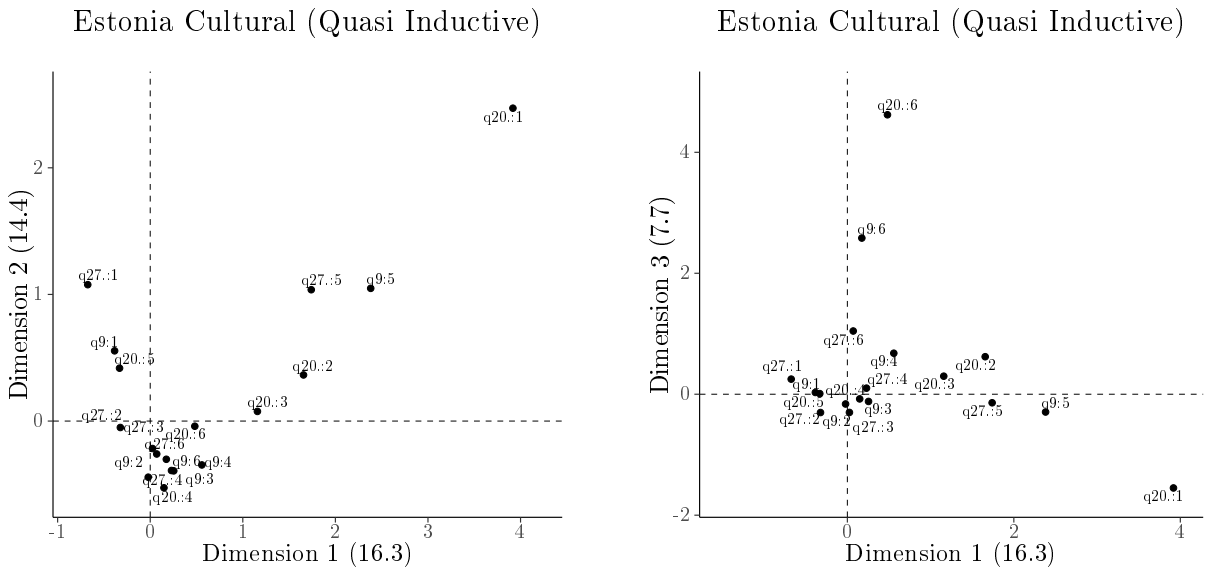


Figure 5.45: MCA for the 1st and 2nd (left) and 1st and 3rd (right) dimension of the Quasi-Inductive Cultural scale in Estonia

shows that even while a scale - in its quasi-inductive form, might be a correct Mokken scale, this does not mean it is useable. Mokken scaling only tells us whether the questions actually relate to each other, not how the users used the categories within it. As such, we should not use the cultural scale for Estonia in any shape or form as the problems within it are so substantial that a different scaling does not improve it.

5.5 United Kingdom

For scales that are successful, I turn to the EU scale in the United Kingdom. This scale had both the high H values and low DDI scores in all three different versions. This would mean we would expect the data for each of the three scales to be near metric, to show correct ordering on the latent dimension, and to contain identifiable dimensions.

Besides, we expect there to be little difference between the PCA and catPCA solutions. We can confirm this by comparing Figures 5.47 and 5.49. Both the rotated and unrotated versions of the plot show the same: a single bundle containing all questions except question 8, which loads on the second dimension. This question, on EU sanctions against Russia, is from the others and is perpendicular to it. This means whatever one thinks of the sanctions is unrelated to one's other opinions on the EU.

Looking at tables 5.20 and 5.21, I find only a little increase in the eigenvalues between the PCA and catPCA solutions, again indicating data of high quality. About the loadings for the rotated solutions, we find that each question loads strongly on one component and only marginally on the other. Taken together, I can conclude the questions in the original scale would form a single scale if question 8 were to be deleted.

The outlier status of question 8 is confirmed when I look at the biplot in Figure 5.50. Here, we can again identify question 8 as not forming part of the main bundle of questions. Moreover, we can see the question is responsible for lowering the DDI as the categories do

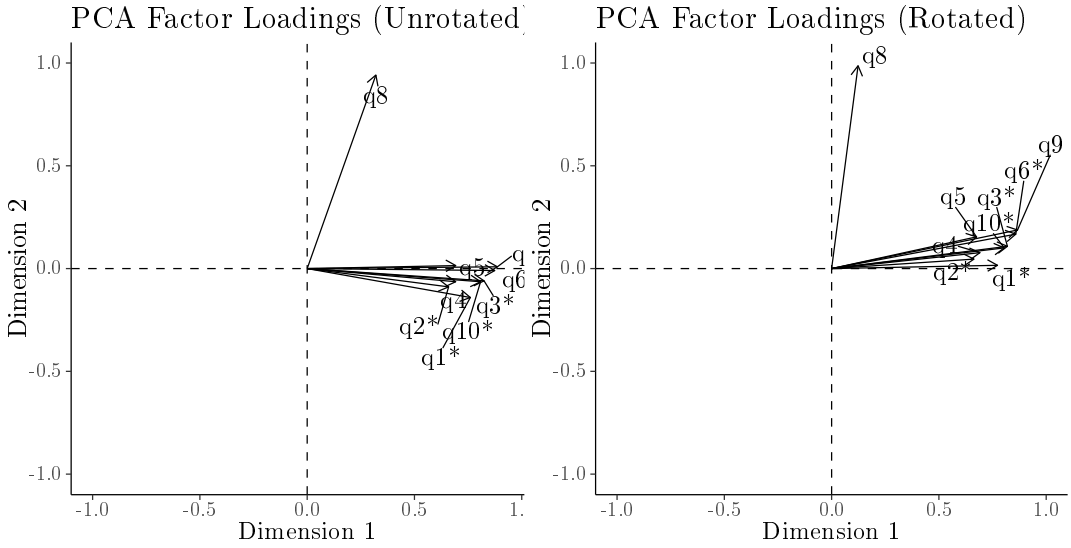


Figure 5.47: United Kingdom - Unrotated and Rotated PCA (Original EU Scale)

	Topic	Unrotated		Varimax Rotation	
		Component 1	Component 2	Component 1	Component 2
Q1*	Euro	0.76	−0.14	0.77	0.02
Q2*	Treaty change	0.66	−0.09	0.66	0.05
Q3*	Right to work	0.82	−0.06	0.82	0.11
Q4	Foreign Policy	0.69	−0.06	0.69	0.08
Q5	Redistribution	0.69	0.01	0.68	0.15
Q6*	EU Membership	0.87	−0.01	0.86	0.17
Q8	Referendum	0.32	0.94	0.12	0.99
Q9	Remain	0.89	0.01	0.87	0.19
Q10*	Membership	0.81	−0.06	0.80	0.10
Eigenvalue		4.96	0.93	4.79	1.09
Variance		55.08%	10.29%	53.25%	12.12%

Table 5.20: United Kingdom - Unrotated and Rotated PCA (Original EU Scale)

not seem to be evenly distributed. Especially its "completely disagree" category seems to carry a considerable amount of weight, indicating that users who completely disagreed with the question "The EU should impose economic sanctions on Russia, even if this jeopardizes gas supplies to EU countries", were seriously opposed to any sanctions at all.

That the other questions did not contain any problems is shown in Figure 5.51 and Table 5.22. Apart from question 8, for which the outlier status of the last category is clear, the other questions present no evidence of ties and all increase in a more or less even fashion. Especially well-performing questions are questions 6 and 9, with the latter one being the best performing question of all. The sole point of note here is that for many questions the "neutral" category was not located in the middle but was either located more at the negative or positive part of the scale.

Turning now to the MCA, Figure 5.52 provides nice evidence that all the methods I have used so far are in essence complementary and provide alternative ways of viewing the

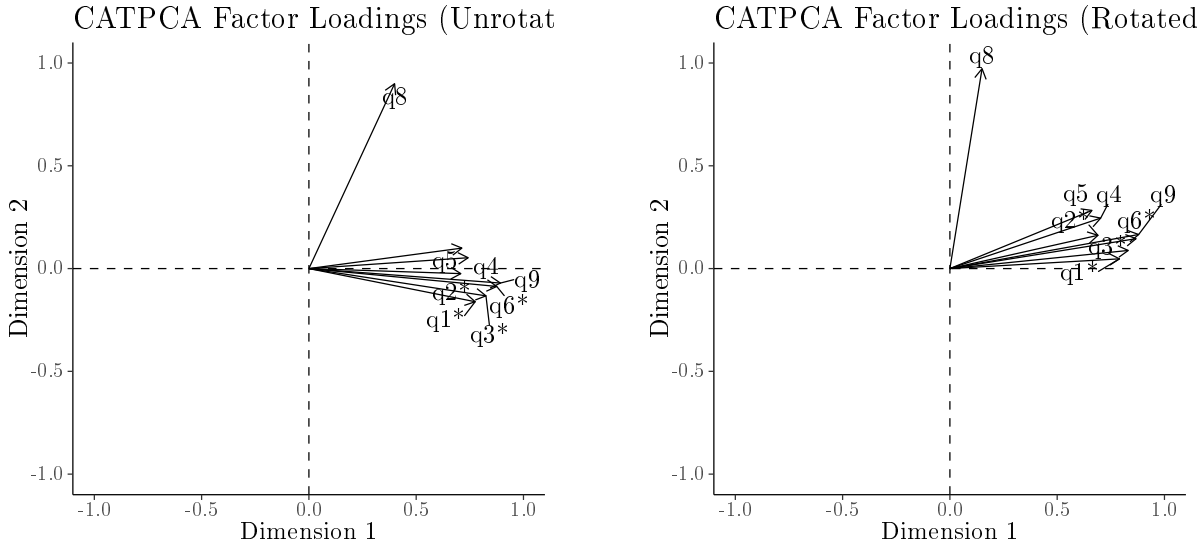


Figure 5.49: United Kingdom - Unrotated and Rotated catPCA (Original EU Scale)

	Topic	Unrotated		Varimax Rotation	
		Component 1	Component 2	Component 1	Component 2
Q1*	Euro	0.77	−0.16	0.79	0.05
Q2*	Treaty change	0.71	−0.02	0.69	0.16
Q3*	Right to work	0.83	−0.13	0.83	0.09
Q4	Foreign Policy	0.74	0.05	0.70	0.25
Q5	Redistribution	0.71	0.10	0.66	0.28
Q6*	EU Membership	0.88	−0.09	0.87	0.15
Q8	Referendum	0.40	0.90	0.15	0.97
Q9	Remain	0.89	−0.07	0.88	0.17
Q10*	Membership	0.84	−0.09	0.83	0.13
Eigenvalue		5.27	0.89	4.97	1.19
Variance		58.52%	9.85%	55.18%	13.19%

Table 5.21: United Kingdom - Unrotated and Rotated catPCA (Original EU Scale)

same data. To begin with, I can identify a clear horseshoe shape, with the first dimension indicating one's position on the underlying EU scale, and the second dimension indicating the extremeness of the category. The sole exception to this rule is again question 8. Seen from the first dimension, they run as $2-3-1-4-5$, pointing at serious problems with the ordinality. Apart from this, all the other questions perform as expected.

Looking at the third dimension in Figure 5.53, we identify a pattern we have not seen thus far. The usual job of this dimension to the No Opinion categories from the others has now been taken over by the fourth dimension. The third dimension itself is hard to interpret and seems not to have any meaning. The fourth dimension meanwhile seems to seperate the No Opinion categories from the others well, though especially in the case of question 8 the category is closest to the others, indicating that some between the No Opinion and "neutral" categories has taken place.

Given the problematic nature of question 8 and its low scalability of $H_i = 0.22$, it was

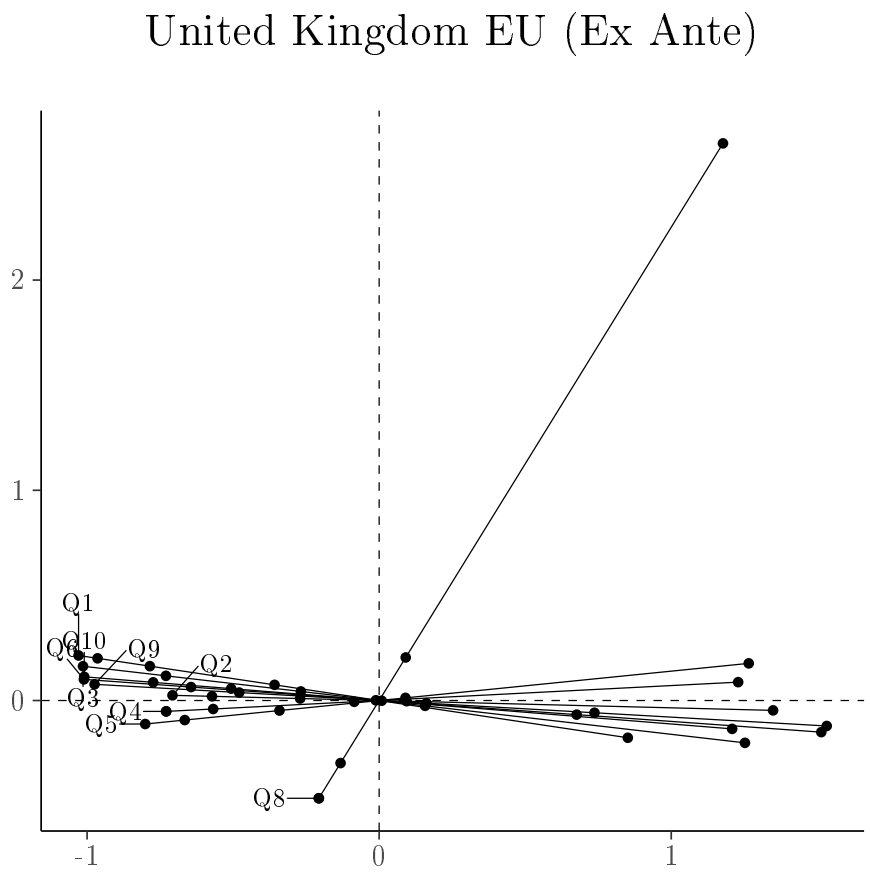


Figure 5.50: Biplot for the Original EU Scale in the United Kingdom

logical that the question was scrapped from the DSV scale. Looking at the biplot in Figure 5.54, this also means we can now better distinguish between the different questions which in the previous biplot were too close together to distinguish. As such, we find that the questions on the Common Foreign Policy and whether the EU should distribute resources (questions 4 and 5) are found in a bundle that is different from the rest. If we would thus want to further improve the scale in this way, eliminating these questions could help to further improve the scale. Another reason to delete these questions is that when looking at the quantifications in Table 5.23 I find the only tie in question 4 between the "disagree" and "completely disagree" categories. Moreover, the question seems to be unbalanced with the middle point being around the "agree" category. For the other questions, we only find a few problems and can view the data to be close to metric.

Looking at the MCA in Figure 5.55 reveals a clear horseshoe shape under a small angle with the first dimension. This has as a result that, when projected on the first dimension, some categories have an incorrect ordinality, for example as in the case of question $2, 4$ and 5, where the first and second categories are reversed, though they are very close together. The third and fourth dimension (not shown here), meanwhile, reveal an identical situation as for the original scale, where the third dimension is hard to interpret and the fourth dimension separates the No Opinion categories from the others.

Turning now to the quasi-inductive version of the scale we find some good evidence that

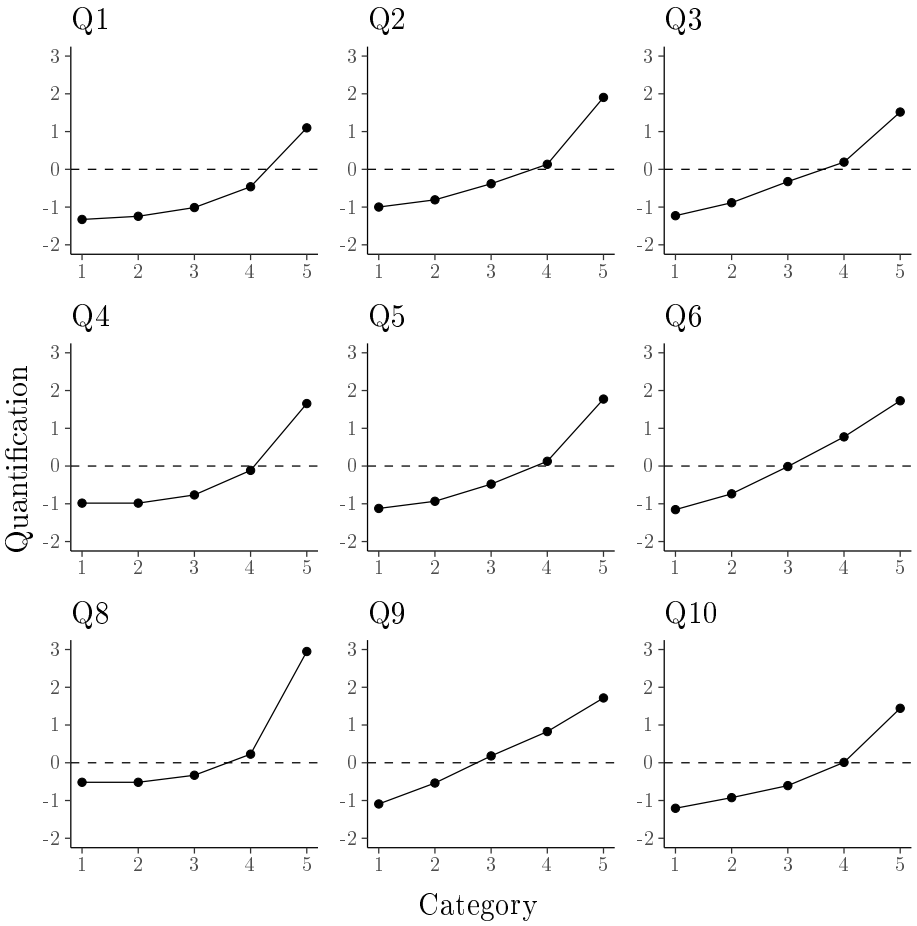


Figure 5.51: United Kingdom - Transformation Plot (Original EU Scale)

	Topic	CA	A	N	D	CD
Q1*	Euro	−1.33	−1.24	−1.01	−0.46	1.10
Q2*	Treaty change	−1.00	−0.81	−0.38	0.13	1.90
Q3*	Right to work	−1.23	−0.88	−0.32	0.19	1.52
Q4	Common Foreign Policy	−0.98	−0.98	−0.77	−0.12	1.66
Q5	Redistribution	−1.12	−0.93	−0.48	0.13	1.77
Q6*	EU Membership	−1.15	−0.74	−0.01	0.77	1.73
Q8	Russia sanctions	−0.52	−0.52	−0.33	0.23	2.95
Q9	Remain in EU	−1.09	−0.54	0.18	0.83	1.72
Q10*	Membership referendum	−1.21	−0.92	−0.60	0.01	1.44

Table 5.22: United Kingdom - catPCA Quantifications (Original EU Scale)

a more extended scale is not necessarily better and that a Mokken scale does not necessarily need to be unidimensional (Smits, Timmerman, and Meijer 2012). Put differently, scalability does not imply unidimensionality. This is clearly shown in the biplot of the scale in Figure 5.56. Here, we can distinguish between two different bundles of questions. The first contains both the original EU-questions and the cultural questions; the second contains all the economic questions. As such, the resulting quasi-inductive scale thus returns a highly scalable scale, but in essence, represents two dimensions instead of a single one. If one would want a pure unidimensional scale, it would thus be best to consider them as two separate scales.

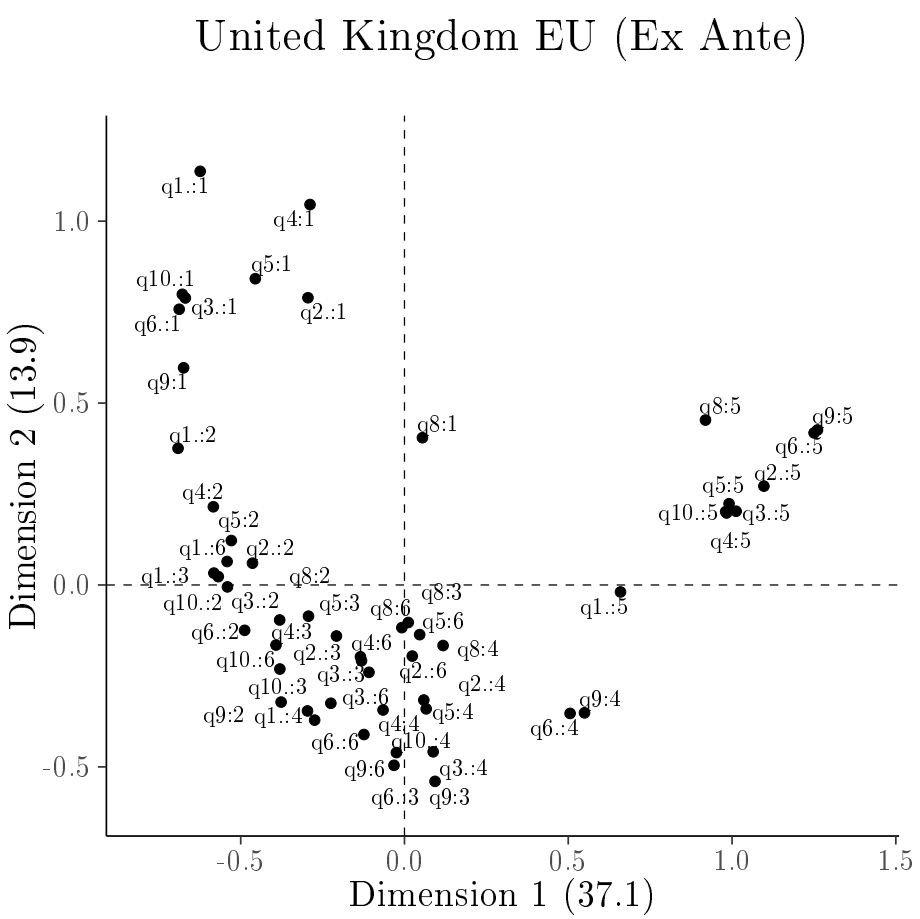


Figure 5.52: MCA for the 1st and 2nd dimension for the Original EU scale in the United Kingdom

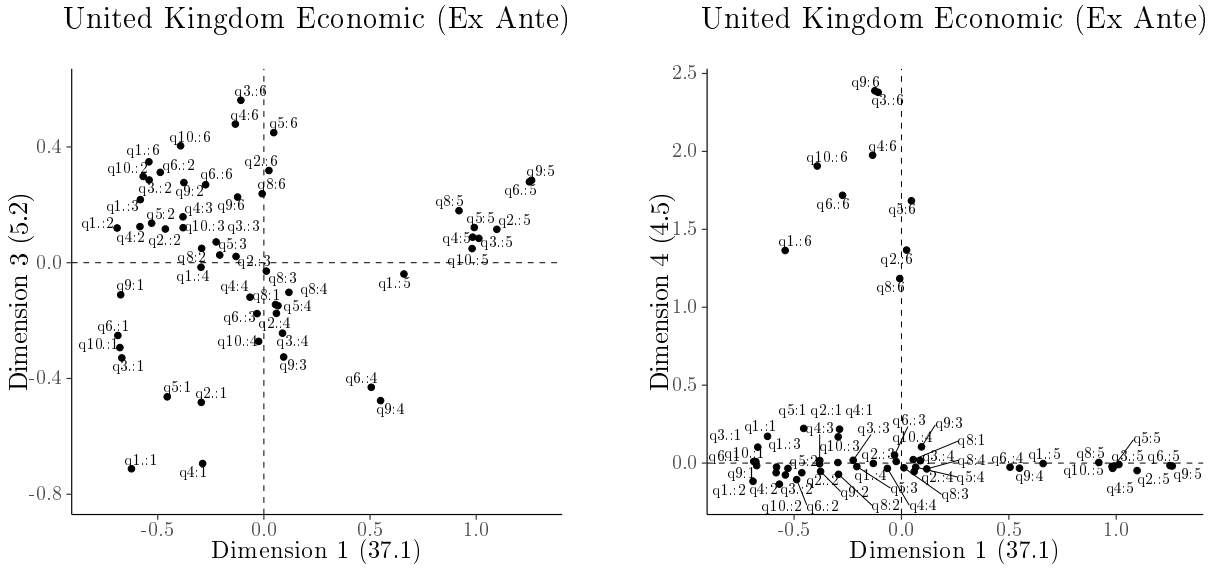


Figure 5.53: MCA for the 1st and 3rd dimension (left) and 1st and 4th dimension for the Original EU scale in the United Kingdom

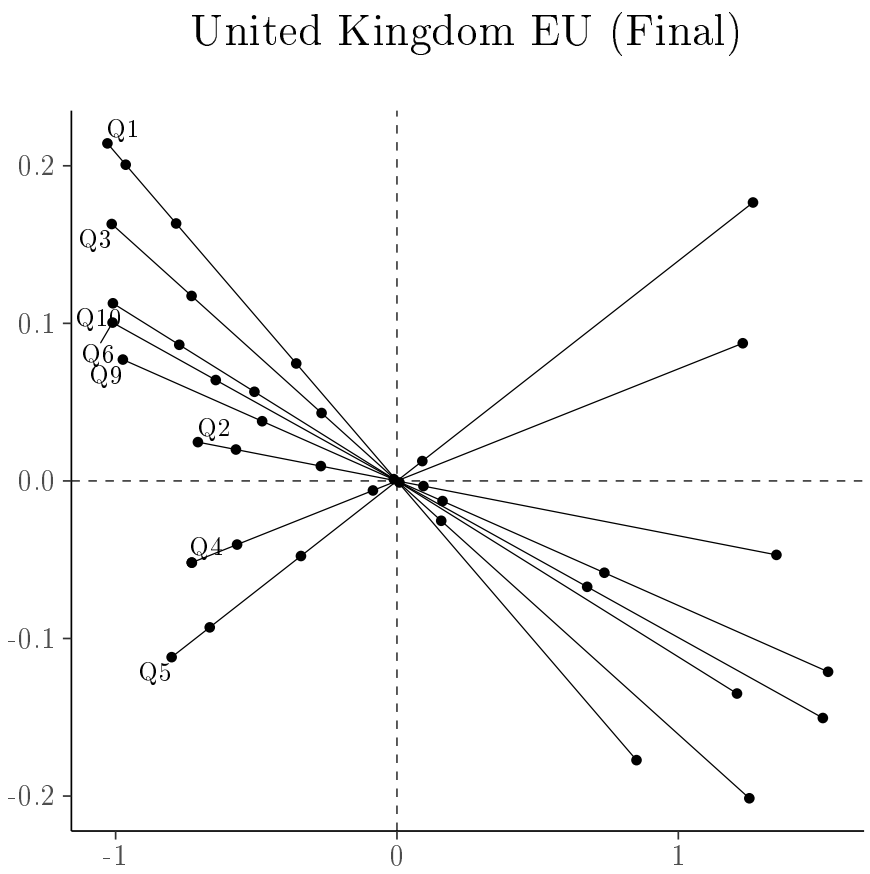


Figure 5.54: United Kingdom - catPCA Biplot (DSV EU Scale)

	Topic	CA	A	N	D	CD
Q1*	Euro	−1.09	0.43	1.05	1.27	1.27
Q2*	Treaty change	−1.86	−0.21	0.37	0.84	1.04
Q3*	Right to work	−1.48	−0.27	0.35	0.91	1.25
Q4	Common Foreign Policy	−1.61	0.02	0.79	1.05	1.05
Q5	Redistribution	−1.12	−0.93	−0.48	0.13	1.77
Q6*	EU Membership	−1.67	−0.83	−0.04	0.75	1.18
Q9	Remain in EU	−1.65	−0.89	−0.24	0.52	1.13
Q10*	Membership referendum	−1.42	−0.07	0.61	0.95	1.21

Table 5.23: United Kingdom - catPCA Quantifications (DSV EU Scale)

Looking at the quantifications in Table 5.24, I find only a single tie in question 4, which I already identified as problematic earlier. Another partly problematic question is 21 on whether "Immigrants must adapt to the values and culture of the United Kingdom". Here, while not tied, the "neutral", "disagree" and "completely disagree" responses are close together, while even the "agree" option is placed well at the positive side of the dimension. The other question for which the same goes is question 1, "The United Kingdom should never adopt the Euro" where even agreeing with the question is seen as adopting a negative stance. Turning to the MCA, I find a clear horseshoe shape with the ordinality completely intact. Moreover, while again the third dimension is hard to interpret, the fourth dimension separates the No

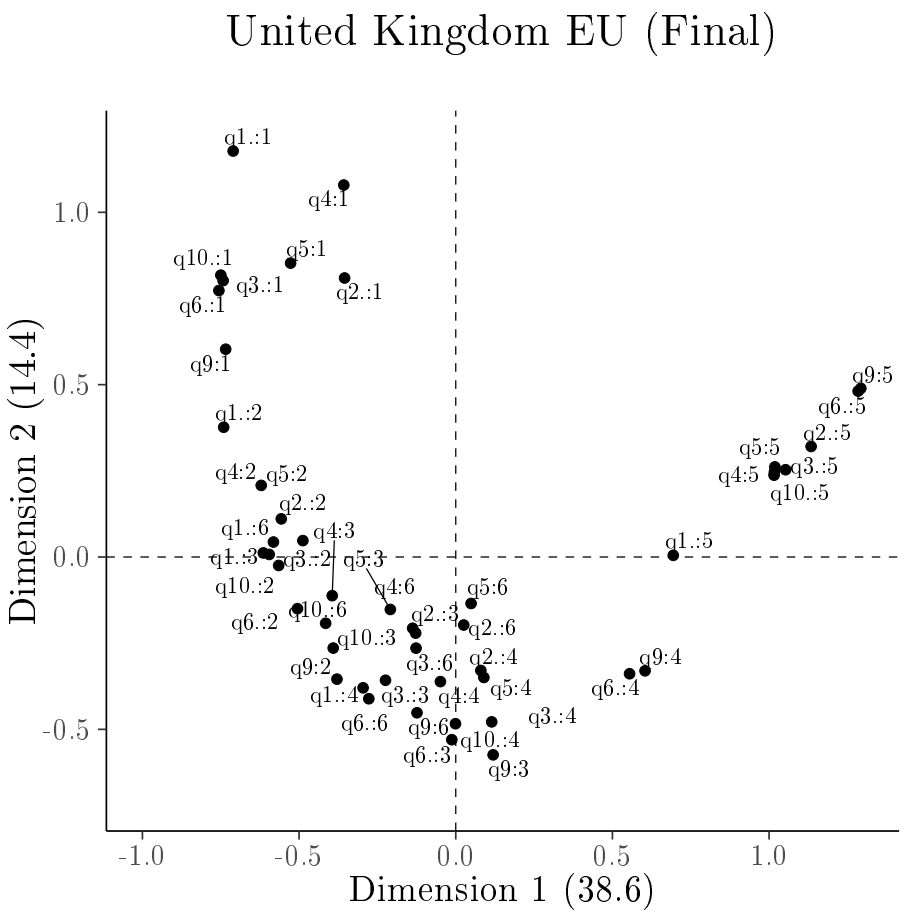


Figure 5.55: MCA for the 1st and 2nd dimension for the DSV EU scale in the United Kingdom

Opinion responses from the other responses (not shown here).

Before I move on, let us first see if I can further improve the quasi-inductive scale. As I already mentioned, MSA ensures scalability, not unidimensionality. According to the biplot in Figure 5.56 we can find not one, but two dimensions in the data - one containing the economy questions and one containing the EU and cultural questions. To see what would happen when I would separate these two, I run a new MSA on the questions for each of the two clusters. Table 5.25 shows the results. Here, I see that instead of a single scale with a H value of 0.46, I gain two scales - one with a H value of 0.57 and the other with a H value of 0.44. Moreover, especially for the EU-Cultural scale, the reliability coefficients are high. This provides us with even more evidence that it is best to view the political space of the United Kingdom as having two separate dimensions as opposed to having only one. Running a PCA on the EU-Cultural dimension seems to confirm the unidimensionality of this dimension as the first dimension has an eigenvalue of 7.24 (with 55.66% explained variance), while the second dimension only has an eigenvalue of 0.90. Loadings for the individual questions on the first dimension range between 0.59 and 0.85, again supporting the unidimensionality of this new scale. For the Economic dimension the results are less, with an eigenvalue for the first of 3.08 and 51.35% explained variance (with the second dimension having an eigenvalue of 0.70 and 11.74% explained variance), and loadings between 0.64 and 0.79.

The United Kingdom presents a compelling case of the limits of Mokken Scaling Analysis

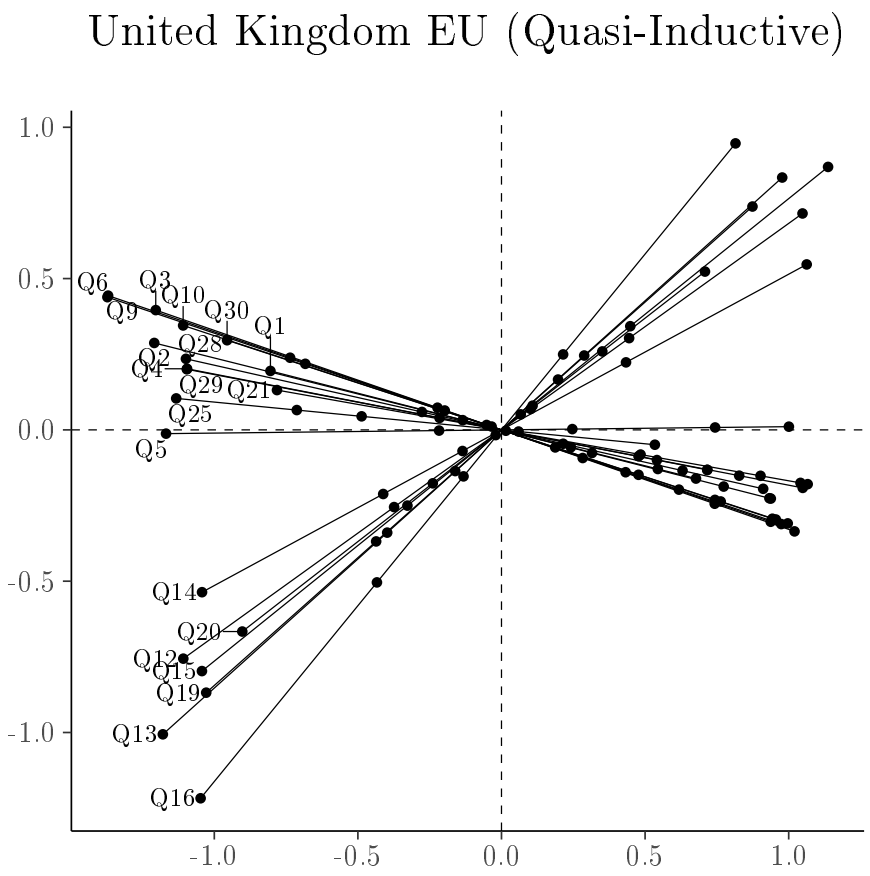


Figure 5.56: Biplot for the Quasi-Inductive EU Scale in the United Kingdom

and the value of graphing, in this case through the use of a biplot. While MSA found a single long EU scale with high scalability, inspection of the biplot showed that the scale is not unidimensional. Instead, while the EU did supersede the separate EU and cultural scales, we could best omit the economic scale as the questions it contains are different from the others. This strengthens the claim by Smits, Timmerman, and Meijer (2012) that scalability and unidimensionality are not identical.

5.6 What Makes a Good Scale?

Both this and the previous chapter assessed the validity of the questionnaires that VAAs use by looking at the responses users gave to them. The previous chapter used the DDI to assess whether users used the full range of responses the Likert scale offers them, which showed there was a considerable difference between countries and scales. Still, the DDI assumes the scale it analyses is unidimensional and can be interpreted. Using MCA, we saw this was not always the case. Even in cases where the scalability values were above the minimum of $H = 0.3$, the main dimensions were often hard to interpret using MCA, or the catPCA showed evidence that there were actually many dimensions present in the scale. In some cases, like the cultural scale in Estonia, even in its DSV form, we could not deem the scale valid. Instead of measuring a supposed underlying economic dimension, the scale separates users with different

	Topic	CA	A	N	D	CD
Q1	Euro	−1.09	0.43	1.05	1.27	1.27
Q2	Treaty change	−1.86	−0.21	0.37	0.84	1.04
Q3	Right to work	−1.48	−0.27	0.35	0.91	1.25
Q4*	Common Foreign Policy	−1.61	0.02	0.79	1.05	1.05
Q5*	Redistribution	−1.63	−0.30	0.34	1.04	1.40
Q6	EU Membership	−1.67	−0.83	−0.04	0.75	1.18
Q9*	Remain in EU	−1.65	−0.89	−0.24	0.52	1.13
Q10	Membership referendum	−1.42	−0.07	0.61	0.95	1.21
Q12	Public Sector	−1.75	−0.59	0.16	0.70	1.65
Q13	State Intervention	−2.18	−0.74	−0.04	0.53	1.81
Q14*	Redistribution	−1.63	−0.64	−0.21	0.68	1.66
Q15	Government Spending	−1.79	−0.56	0.11	0.77	1.95
Q16	Fire People	−2.28	−0.94	−0.29	0.47	1.77
Q19	Income Tax	−1.84	−0.78	−0.29	0.35	1.56
Q20	Fracking	−1.77	−0.47	0.21	0.69	1.39
Q21	Values	−1.04	0.65	1.20	1.39	1.42
Q25*	Same Sex Marriage	−1.91	−1.20	−0.82	0.10	0.90
Q28	Islam	−1.47	−0.37	0.29	0.84	1.22
Q29*	Asylum Seekers	−1.45	−0.29	0.63	1.10	1.39
Q30	Immigration Quota	−1.22	0.24	0.97	1.22	1.27

Table 5.24: United Kingdom - catPCA Quantifications (Quasi-Inductive EU Scale)

response styles. Ireland performed better in this respect. Especially for the DSV scale, the main dimension was interpretable, and there was only a little evidence of users trying to hide their non-response by opting for a neutral question. Hungary, meanwhile, showed a promising EU scale, both in its original and DSV version. Even so, there was evidence of a significant amount of switching between the neutral and no opinion categories, indicating a considerable number of users tried to hide their non-response by choosing the neutral option.

The validity of the questionnaire in EUVox is thus a mixed bag. While in some cases the scales perform as expected, they often do not, and employing them might lead to misguided conclusions. This should serve as a warning for those that use these data to draw conclusions on the differences between countries, as the scales might not measure the same concept. It is thus advisable to them to establish first whether they can interpret the scale by running MCA. More problematic is the use of the scales for low-dimensional matching. As the matching algorithm positively matches those users with a high value on the scale with parties with a similarly high value, it expects that in both cases these high values have a substantial meaning. Still, as we saw in Estonia, this needs not always be the case. As such, the information this scale provides for the eventual match is thus not valid, as it does not provide a true idea of the economic position of the user.

The logical follow-up question is this: what *does* make a good scale? Or, put in more practical terms for designers of VAAs: what do I need to make a valid and reliable scale? Based on what I have seen thus far, I can distinguish the following:

Inclusiveness To begin with, there should be enough information within the available ques-

	EU-Cultural		Economic	
Question	H-value	crit-value	H-value	crit-value
q1	0.57	0	–	–
q2	0.47	30	–	–
q3	0.64	0	–	–
q4	0.49*	53	–	–
q5	0.52*	0	–	–
q6	0.65	0	–	–
q9	0.67*	0	–	–
q10	0.58	50	–	–
q12	–	–	0.49	0
q13	–	–	0.44	0
q15	–	–	0.47	0
q16	–	–	0.42	0
q19	–	–	0.44	0
q20	–	–	0.39	0
q21	0.57	0	–	–
q25	0.48*	27	–	–
q28	0.56	1	–	–
q29	0.57*	0	–	–
q30	0.62	0	–	–
Total H	0.57		0.44	
α	0.93		0.81	
α (Ord)	0.95		0.84	
ω (Ord)	0.95		0.82	
LCRC†	0.94 (13)		0.81 (6)	
N	61435		61435	

$*$ Questions with an asterisk ($*$) were reversed.
† I used the Bayesian Information Criterion to determine the number of latent classes.

Table 5.25: New Quasi-Inductive Scales for the United Kingdom (England)

tions to form a scale. The reason we call quasi-inductive methods "quasi" is that they are dependent on which information is available in the questions. If the questions fail to measure some fundamental ideas on a certain scale, no amount of analysis is going to discover this scale.

Reliability The scale should have a high level of reliability. We can use various methods to assess the reliability, with Cronbach's α giving the lower bound. Given the fact that most VAA data is polytomous as it uses Likert scales, the LCRC is a good indicator for the reliability. Given that we need reliable placements on the underlying scales, I should aim for a LCRC > 0.9.

Scalability The questions on the scale should be well scalable, with both the scale and the individual questions having a H score of > 0.3. Moreover, we should assess questions

with H scores that deviate from the others to establish the reason why the scores differ so much. Also, the crit scores for each of the questions should be < 80.

Unidimensionality The scale should be unidimensional. Note that MSA does not guarantee unidimensional scales - it only guarantees that the scales have a good scalability (Smits, Timmerman, and Meijer 2012). Thus, we should assess the dimensionality separately. We can do this by using various IRT methods (cf. Smits, Timmerman, and Meijer 2012, p.536). We can also visualize this using a biplot resulting from a principal component analysis. Here, we can identify clusters of questions on which we can then run separate MSA procedures.

Validity Ensure the scale "means" something and that we can interpret it. Scaling methods only associate questions with each other and do not look at their meanings. So, we should use no scale when we cannot assign it with a clear meaning. This also goes for the individual questions, each of which we should be able to assign a reason why it belongs to the scale.

Quality The DDI of the scale should be at least < 0.3 and preferably < 0.15. This is to ensure the distances between the categories of the questions are as equal as possible and can thus be assumed to be metric. This step is to ensure the users treated all questions in the same way - i.e. answered using the same number of response options. Questions that have one or more ties or categories whose quantification differs substantially from the others should be reconsidered for inclusion. Moreover, in the MCA of the scale, the first dimension should be interpretable as the dimension under interest, while the second dimension should help form a horseshoe shape. Here, the 1st dimension should correct order the questions, with the negative questions reversed when necessary. The second dimension should separate the extreme responses from the moderate responses. Besides, in further dimensions, the No Opinion categories (if present) should be well separated from the other categories. In a plot, the No Opinion categories should be highly associated with that dimension, while the other categories are unassociated with it.

Given the nature of the data, it seems all but impossible that all VAA scales will be able to meet these criteria. Most problematic is that most VAAs have only a limited number of questions, thus restraining the amount of data there is to work with. Moreover, given that VAAs are online platforms without any supervision, there is a considerable risk that users will not pay the desired amount of attention to each of the questions. While cleaning the data will aid to get rid of the most problematic questions, this will not overcome the problem of how users actually respond to the questions and whether they comprehend them as intended. Besides, each of these procedures requires a certain amount of data to work with. Thus, as with Dynamic Scale Validation, we need a certain number of users to run the analyses described here.

This chapter has taken a critical look at the scales that most VAAs use and the way in which the users respond to them. Besides, in the previous chapters, I already introduced the

idea that when looking at VAA generated data it is important to keep both the users and the individual questions in mind. In the next chapter, I take this idea one step further. There, I will look at how different questions that are only a little different, can still lead to very different responses. This should serve as a further sign to VAA designers that questions, and the way they are formulated, belongs to the core of what a VAA is.

6 | The Effects of Statements

In this chapter, I will take a closer look at what makes up those scales we have discussed so far: the questions. In the previous chapter, we already saw that users struggled with some questions more often than with other questions. Yet, we only related that to the content of the question. In other words, we assumed that questions were difficult for the users because of what they were *about*. Yet, a question is more than what it is *about*. This is because there are many ways to ask the same question. Most VAA designers assume that the actual wording of the question does not change much. Still, question-wording matters on many levels. To begin with, we know that parties own certain issues. Yet, it might be they only own these issues when we word them as such. For example, a party which is against the European Union might own the issue if it reads that "our country should *leave* the European Union", but might not own it if it reads that "our country should *remain in* the European Union". Because including questions which certain parties own influences the match between the user and party (Lefevere and Walgrave 2014), choosing which wording to use might not be a neutral affair after all. Also, question-wording also influences users in a direct way. From the psychological literature, it is well-known that as soon as the wording of the questions becomes more complicated, the number of misunderstandings about what the question means increases. This is especially so for those users that have less experience with questionnaires or have a lower knowledge of the field. While this is more related to the difference in response that the user gives, here I will focus on the latter effect: the match between the user and the party.

This leads to my second main research question: *how does the wording of the questions influence the match between the user and the party*? There are many ways in which a question in a VAA can represent a similar topic, and the potential differences between different wordings are plentiful. Here, following Holleman, Kamoen, Krouwel, et al. (2016), I focus on one of them: whether the wording of the questions is positive or negative. Thus, a question reading "soft drugs should be allowed" is positive, while a question reading "soft drugs should not be allowed" is negative. I will thus make the research question more specific:

Q Does the match between the user and the party differ when the wording of the question is either positive or negative?

In the following section, I will consider what wording effects are and which possible factors might influence their strength. Besides, I also state some hypotheses. Following that, I will continue with the description of the experiment I will use to test these hypotheses.

6.1 Influences on Question Wording Effects

That the way in which users answer the questions depends on their wording is well-established (e.g. Schuman and Presser 1981; Hippler and Schwarz 1986; Tversky and Kahneman 1981). For example, Schuldt, Konrath, and Schwarz (2011) found that referring to the phenomenon as "global warming" rather than "climate change" made Republican respondents 16% more likely to endorse the phenomenon. This and other examples show that a simple change in wording can draw the users' focus to different aspects of the phenomenon, leading to different associations. Whether a question is positive or negative is also called the *valence* of the question. A question as "Soft drugs should be forbidden" is negative, while "Soft drugs should be legalized" is positive. Both questions aim to measure the opinion of the user on soft drugs, but do so in different ways. Note that whether a question is negative or positive has no relation to *what* we say in the question, but *how* we say it. Thus, a question like "The death penalty should be introduced" is positive, regardless of what one's opinion on the subject might be. Most of the time, designers of VAAs hold that it is not the wording of the question, but the content of the question that decides on the response of the user. Still, Holleman, Kamoen, Krouwel, et al. (2016) showed that this is not always the case: users do respond in a different way to a similar question if we include it in either a negative or positive form.

H1 There is a difference in the match between the user and the party when we word a question in a positive or negative way.

Negative questions come in two forms: explicit or implicit. To which of the two they belong depends on how we operationalize the negative in the question. For example, the question "Soft drugs should be forbidden" is an implicit negative as the negative is implicit in the specific word "forbidden" that we use. Explicit negatives place an extra word like "not" or "no" in front of the positive version of the verb to ensure the negativeness of the question (Clark 1976). An explicit negative version of the question on soft drugs would be "Soft drugs should **not** be legalized".

Whether we make a negative question implicit or explicit is dependent on the designer. Still, explicit negatives are often found to be confusing for users (Schriesheim, Eisenbach, and Hill 1991). This is because explicit negations are less complex than their implicit counterparts (Horn 1989, p.521-524). While counter-intuitive, the reasoning is that implicit negations only make sense in a context where the use of the implicit negative itself makes sense. For example, one does not talk of *keeping someone from* suicide, unless one supposes someone had the intention of doing so in the first place (Horn 1989, p.523). Yet, for an explicit negation, one first has to reconstruct the positive assumption, and then make it negative. As an example, Kaup, Lüdtke, and Zwaan (2006) asked respondents to imagine a non-open door. They found it took them twice as long to imagine this then the positive version of an open door. In VAAs, we use both implicit and explicit negatives, often to ensure that we frame a question in the way in which current politics talks about it. Yet, the process how such a selection goes is rather unclear for most of the VAAs available (Camp, Lefevere, and Walgrave 2014). Thus, the reason why we include a question as "Soft drugs should **not** be legalised", "Soft drugs should be forbidden", or "Soft drugs should be legalized" is unclear.

This is problematic insofar as a certain type of wording might be able to benefit certain parties and be detrimental to others. To gain a better understanding of how the type of negative question influences the eventual match between the user and the party, I hypothesize:

H2 There will be a difference in the effect of question-wording for implicit and explicit negatives.

6.1.1 Question, Party, and User Effects

Question effects refer to those effects that a question has on a user, apart from their actual opinion on the topic the question tries to discover. The assumption here is that the framing of the question influences the user. This framing comes in two kinds: *issue framing* and *valence framing*. While issue framing looks at the angle from which we present an issue, *valence framing* looks at whether we express the issue in a positive or negative way. Of these two, valence framing is the more subtle variant. For example, while both the statements *Country X has to remain in the European Union* and *Country X has to leave the European Union* aim to capture the respondents' opinion towards their countries' membership of the European Union, each does so in a different way. The main difference is the negation adds extra information to the sentence that was not present in the positive form. Because there is no marker to show a positive statement is positive, while such a marker is present in a negative sentence, positive statements are less complex than their negative counterparts (Just and Carpenter 1971, p.248-249). Negative statements, thus need closer attention of the respondent. Because of this, respondents may not realize while answering the statement that "not forbidding" is the same as "allowing" (Schuman and Presser 1981). Stated otherwise, the use of negation does not imply that one has the opposite in mind of what the positive version of the statements would be (Giora et al. 2007). Thus, negative formulations do not correspond with their positive counterpart (Horn 1989)[1]. About the direction of this difference, users are more likely to disagree with negative statements than to agree with similar positive statements[2](Kamoen, Holleman, Mak, et al. 2017; Kamoen 2012).

Negations themselves come in two kinds: explicit and implicit. Explicit negations are negatives like "never" or "no", while implicit negations are words that contain a negation within themselves, such as "abolish". Perhaps counter-intuitively, Horn (1989, p.521-524) points out that because explicit negations are less complex than their implicit counterparts, they are harder to answer. This is because implicit negations only make sense in a context where the use of the implicit negative makes sense. For example, one does not talk of *keeping someone from* (negative) suicide, unless one supposes someone had the intention of doing so (Horn 1989, p.523). Yet, for an explicit negation, one first has to reconstruct the positive assumption and then make it negative. As an example, Kaup, Lüdtke, and Zwaan (2006)

[1] As such, we might excuse designers of VAAs might not to include any negatives at all. Yet, one reason why they do include them is to avoid acquiescence bias (Pasek and Krosnick 2010, pp.38-39) as in general, respondents seem to be biased towards a positive response instead of a negative response. Besides, they can use negatives to awaken "sleeping" respondents and make them aware of the fact that statement content varies (Swain, Weathers, and Niedrich 2008, p.116).

[2] The direction of the difference between negative-positive has been the subject of debate. In a meta-analysis, Holleman (1999) finds that a response of "no, not forbid" is obtained more often than "yes, allow", a conclusion later on supported by Kamoen (2012). Still, Reuband (2003) found that such an effect disappears if the response options clearly and explicitly state the difference ("allow or not allow" or "allow or forbid").

asked respondents to imagine a "non-open" door and found it took them twice as long to imagine this then the positive version of an "open" door.

Explanations for the cause of the difference between positive and negative statements emerge from two schools of thought: those based on the *response process* and those based on *attitude strength*. The former looks at where in the response process the difference between answering a negative or positive statement occurs. We already saw this process in a previous chapter. Simplified, this process consists of two steps (Holleman 2000). First, there is an *acquisition stage* of comprehension and retrieval. Then, there is an *answering stage* of translation and judgement. So, a difference during the first stage points to users understanding the positive and negative versions in a different way. A difference during the second stage points to users understanding the versions in the same way, but translating and mapping them in a different way onto the response options (Chessa and Holleman 2007, p.204). Most current literature points to the second stage as the main source of difference (Holleman 1999; Kamoen 2012). Chessa and Holleman (2007) find that while the time to read positive and negative questions was similar, the time to answer those questions was different. While this would go against the above-mentioned idea that negative statements are more complex and that the difference should, thus, be found during the first stage, Holleman and Kamoen (2017) argue that most studies have a too coarse measure grained and that using more precise measurements and linking these to eye-tracking might lead to different results.

The latter explanation, meanwhile, depends on the strength of the attitude a user has towards a certain statement. The strength of this attitude can be measured using what Converse (1964) calls the "attitude continuum", which runs from clear and concrete *strong* attitudes to unclear and vague *weak* attitudes. Krosnick and Petty (1995) provide four features that characterize strong attitudes: they are stable over time, resistant to change, and influential on cognition and action. Weak attitudes are characterized as fluctuating and inconsequential (Dalege et al. 2017). Whether an attitude is strong or weak depends on the importance one attaches to them. As a result, users with weak attitudes are more susceptible to question effects than those with strong attitudes (Petty and Krosnick 1995; Converse 1974; Payné 1951). Moreover, while this effect has been criticized (e.g. Krosnick and Schuman 1988; Lavine et al. 1998; Bassili and Krosnick 2000), Holleman and Kamoen (2017) argue that much depends upon exactly how and in what way attitude strength is being conceptualized. Here, we construct attitude strength on the basis of the political knowledge of the user, as this would seem to be the most logical in the context of VAAs.

While we expect an effect of question-wording on the degree of match between the voter and the party, we expect this effect will not be the same everywhere. We expect variation at three levels: (1) on which question we calculate the match (2) with which party we calculate the match, and (3) what the characteristics of the user are for which we calculate the match.

Of the questions, we expect differences because not all questions are the same. Some questions may be embedded in ideological cleavages, while others are cross-cutting. Also, questions that were salient during the elections and with which parties have become identified are likely to show larger differences between the two types of question-wording than other questions.

In the same way, we expect that within these questions, the effect of question-wording differs for each party. Research on the match between voters and parties shows that the

amount of salience a party gives to an issue is a key mediator for the match between the voter and the party on that issue (Giger and Klüver 2016; Klüver and Spoon 2016). So, parties who own certain issues might show larger effects as a result of the question-wording. Given that we assume that changing the valence of the questions means that we have different questions instead of the same ones, we are in fact talking about two different questions. As such, it can be a party owns the positive version of the issue but not the negative version. Given the question on soft drugs, it might be a party has a strong opinion on the positive version of the question "Soft drugs should be legalized", but not so much on its negative version "Soft drugs should be forbidden". Given that including questions which parties "own" effects the number of matches they receive (Lefevere and Walgrave 2014; Ramonaitė 2010), we can expect differences between parties as well.

Of users, we expect that question-wording does not affect each of them to the same degree. To begin with, users with a higher political sophistication should have fewer problems understanding what the questions are *about*. Also, users with more experience of answering questions should be more likely to understand the structure of the questions. Users with a low political sophistication and less experience with questionnaires will pay less attention to the questions and the wording of the questions is thus more likely to affect them (Bassili and Krosnick 2000).

H3a The effect of the question wording on the match between voters and parties will depend on the topic of the question.

H3b The effect of the question wording on the match between voters and parties will depend on the party with which we calculate the match.

H3c The effect of the question wording on the match between voters and parties will depend on the political sophistication of the user.

6.2 A Look at the Data

To measure the effect of the valence of the question wording, we would, if possible, launch two similar VAAs, one with only positive and one with only negative questions. Still, given that we can expect the wording of the questions can influence the users and the outcome of the VAA, this would be unethical. To account for this, I designed a VAA with two different versions (hereafter Version A and Version B). Both versions had 25 questions in total, with 13 questions being exactly similar for both (the wording of these questions was positive). Of the remaining 12 questions, the wording differed for each version. When the wording was positive in Version A, it was negative in Version B, and when the wording was negative in Version A, it was positive in Version B. This means that where Version A had 7 negative questions, Version B had 7 positive questions and that where Version B had 5 negative questions, Version A had 5 positive questions. I chose the questions based on similar questions occurring in other VAAs, and on other topics I deemed relevant. Whether I formulated the questions in a negative or positive wording depended on whether a negative wording would make sense, both in a practical and grammatical way. Of the negatives, 4 of them were explicit negatives,

while 8 were implicit negatives (see also Table 6.1). I decided which negatives were implicit and which explicit at random. For a full overview of the questionnaire see Appendix B.

Set	Number of statements	Version A	Version B
I	3 statements	Negative explicit	Positive
II	1 statement	Positive	Negative explicit
III	3 statements	Negative implicit	Positive
IV	5 statements	Positive	Negative implicit
V	13 statements	Positive	Positive

Table 6.1: Distribution of Positive and Negative Questions over the two Groups

I ran the Stem-Consult VAA (`www.stem-consult.nl`), launched for the Dutch parliamentary elections of 2017, was active between the 1st and 17th of March 2017, and in cooperation with the PreferenceMatcher network[3] (Gemenis, Bruinsma, et al. 2017). Stem-Consult reached potential users through word-of-mouth and Facebook advertisements. Upon entering the website, the VAA assigned users at random to either Version A or B of the VAA. Besides the main questionnaire, the VAA asked users optional questions on age, political interest, education and sex, amongst others. The VAA included 14 or the 28 parties participating in the elections. Parties included were all parties that were already present in parliament, as well as parties that showed a consistent chance of obtaining at least a single seat in the polls. The included parties are 50Plus, a pensioners' interests party, the CDA, a Christian-democratic party, the CU, a social Christian party, D66, a social-liberal party, the FvD, a new conservative Eurosceptic party, GL, a Green party, the PvdA, a social-democratic party, DENK, a PvdA splinter party focusing on ethnic minority rights, the PvdD, an animal rights party, the PVV, a radical right populist party, VNL, a radical right splinter party from the PVV, the SGP, an orthodox Calvinist Christian party, the SP, a radical left party, and the VVD, the main centre-right liberal party[4]. All parties expect VNL gained seats in the parliament after the elections. The previous government was a coalition of the VVD and the PvdA, and the government resulting from the election would be a coalition between the VVD, CDA, CU and D66. A team of coders positioned the parties using the iterative Delphi-process (Gemenis 2015). The coders positioned all parties on the questions as appearing in version A of the VAA. The positions for Version B were then generated by reversing the positions were necessary. See for an overview of the party positions Appendix B.

The two experimental conditions were assigned at random and as a result, the two groups are not different in Sex ($\chi^2(1) = 0.07$, $p = 0.79$), Age ($t(2670) = -0.61$, $p = 0.54$), and Education ($\chi^2(1) = 0.39$, $p = 0.53$). There is a significant difference in Political Interest ($\chi^2(1) = 4.75$, $p < 0.05$), though the actual differences 2.40 for Version A and 2.47 for Version B on a 5-point scale, are small.

In total, 6283 unique users completed the VAA before I took it offline. Like the EU-Vox data-set, I cleaned the Stem-Consult data in accordance with the proposed cleaning by

[3] `http://www.preferencematcher.org/`

[4] CDA=Christen-Democratisch Appèl, CU=ChristenUnie, D66=Democraten 66, FvD = Forum voor Democratie, GL = GroenLinks, PvdA = Partij van de Arbeid, PvdD=Partij voor de Dieren, PVV=Partij voor de Vrijheid, SGP = Staatkundig Gereformeerde Partij, SP = Socialistische Partij, VNL = VoorNederland, VVD = Volkspartij voor Vrijheid en Democratie.

Mendez and Manavopoulos (2018). As with EUVox, I used para-data measuring the time taken to complete each question. I removed users when: the time taken to complete the total of 25 issue statements was less than 75 seconds; they answered at least one issue in less than 2 seconds; they answered 12 or more consecutive statements in the same way. Besides, I removed returning users — identified by similar entries from the same computer — as well as all users taking the VAA after March 15, which was the date of the election and users taking the VAA between $10-12$ March, when the VAA was taken offline for a security update. Finally, I removed the first 50 entries as these most likely were made during initial testing. The cleaned data-set contained 4178 users, with 2051 assigned to Version A and 2127 assigned to Version B. After selecting only those users that provided information on age, political interest, sex, and education, 2674 users remained, with 1328 assigned to Version A and 1346 assigned to Version B. I removed most users because they had at least a single statement with a response time less than 2 seconds (1517 users), or had missing data (1504). While we could call the missing data criterion too stringent, I made the choice to use the same data for all analyses to ensure comparability. Other criteria showed less loss of users, with a total response time of less than 75 seconds losing only 329 users, and only 2 users having more than 12 similar responses. Users who filled out the VAA between the 10th and 12th of March and after the 15th numbered 43 and 20 respectively, while I dropped 125 users because they were returning users.

To reach users, Stem-Consult was advertised using targeted Facebook advertisements that ensured that a representative sample of the population could be reached. To see whether this was indeed the case, we compare the characteristics of the users as found in the raw data-set with those of the general population.

Table 6.2 shows the results of the comparison. Of sex, we find that Stem-Consult attracted more women than men, which is in accordance with the general population, though the differences are less pronounced. The average age is also in accordance with the general population, which is unexpected, as most VAAs seem to reach predominantly younger users (Pol et al. 2014). Of education, we find more expected results as a larger part of the VAA users was highly educated. For the provinces, we find a reasonable overlap, with only Overijssel being overrepresented. This was most likely caused because of a partnership with a university in that province. In all, we find Stem-Consult is a reasonably accurate sample of the general population of the Netherlands except for education.

6.2.1 A First Idea

To get a first idea of whether question-wording might have any effect at all, we take a look at the differences between the responses between the different versions. Figure 6.1 shows the differences between the responses users gave to the questions that were altered in the two versions of the VAA. The differences in Figure 6.1 are not always as expected. While in Q18 the difference in the negative change for the CA and A categories are compensated by the D and CD categories (which is to be expected when the direction of the question is changed), in Q11 it seems that the loss in the A category went to a gain in the CA category. Put differently, the reversal of the question did not lead users who otherwise would have agreed to disagree but moved them from agreeing to completely agreeing. Similar behaviour occurs

		Stem-Consult	Population
Sex	Men	45.8%	49.6%
	Women	53.9%	50.4%
Age	Men	38.2	40.7
	Women	43.0	42.5
Education	Primary	1.7%	8.6%
	VMBO	9.4%	17.6%
	HAVO/VWO/MBO	37.2%	31.6%
	HBO/WO Bachelor	37.5%	14.9%
	Post-Graduate	14.2%	8.6%
Province	Drenthe	3.5%	2.9%
	Flevoland	2.6%	2.4%
	Friesland	3.9%	3.8%
	Gelderland	13.1%	12.0%
	Groningen	4.9%	3.4%
	Limburg	7.4%	6.5%
	Noord-Brabant	12.5%	14.7%
	Noord-Holland	11.8%	16.5%
	Overijssel	13.8%	6.7%
	Utrecht	7.0%	7.5%
	Zeeland	2.2%	2.2%
	Zuid-Holland	17.1%	21.4%

Population data are from Centraal Bureau voor de Statistiek (CBS)[5]

Table 6.2: Comparison of the characteristics of Stem-Consult users with the general population.

in Q2 and Q25. Besides, the size of the difference is not the same for each question. Some questions, like Q5, show only minimal differences, whilst others, such as Q15, show much clearer differences. What causes these differences is not clear. While it might be because Q15 contains an explicit negative and Q5 and implicit negative, this is not substantiated by other questions (like Q18, which has similar behaviour as Q15 but does have an implicit negative)[6].

6.2.2 A Look at the Validity

Besides, let us take a brief look at what happens when we apply MCA to our data-set (see Figure 6.2). The results are worrying. In two of the Figures (6.2a and 6.2b) representing the economic scale, we fail to see the expected horseshoe shape. Instead, the 1st dimension separates the intensity of the categories and does not represent the underlying scale. Also, the first dimension fails to keep a consistent ordering. This indicates the first dimension might not represent the dimension we expect and that the dimension needs to be re-calibrated to be useful to interpret in MCA. The other two figures — representing the cultural scale —

[6] In a similar analysis - not shown here - we see that for the questions that remained the same, differences were observed between the two different versions.

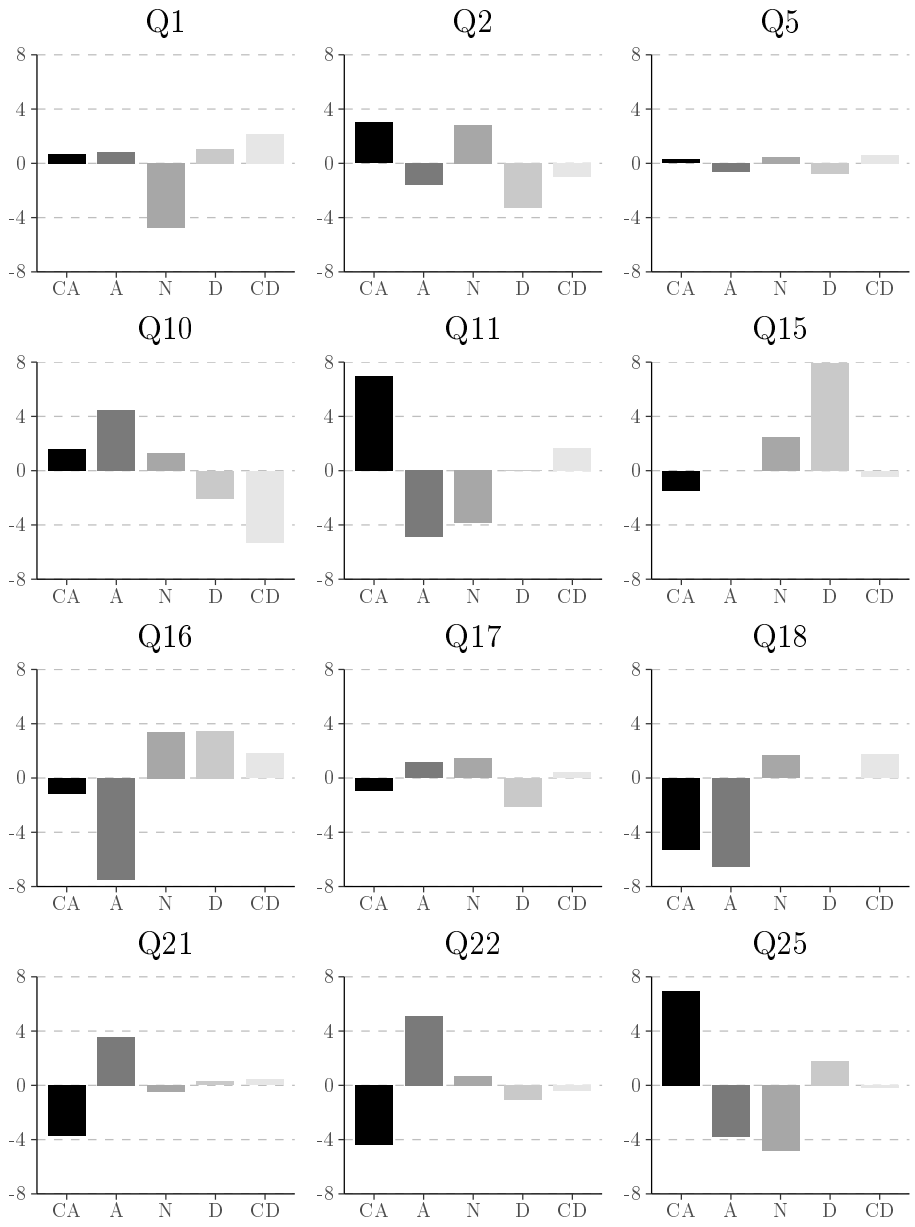
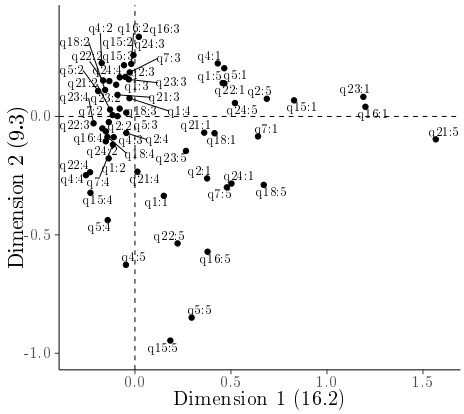


Figure 6.1: Differences between the responses for both of the versions of Stem-Consult, displayed by question

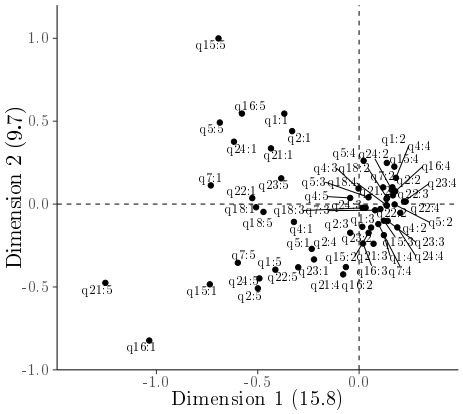
fare better. Except for some unusual behaviour (such as $q13:5$), in both cases, we can find a horseshoe, with the ordering of the responses in the right order. Moreover, the dimension on which the questions are ordered has high inertia (30.4% and 27.1%). This means that the underlying scale handles a sizeable amount of the variation in the data. As such, the cultural scale has better chances of emerging from a Dynamic Scale Validation than the economic scale, to which I will turn now.

6.2.3 Another New Look at the Validity

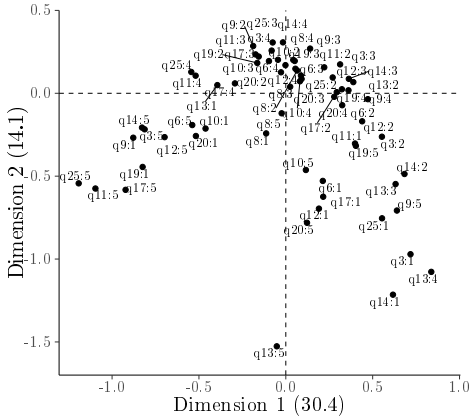
The Appendices show the results for both analyses. Clear is that both ex-ante scales are deficient for both versions. The ex-ante economic dimension never reaches the minimum H_i



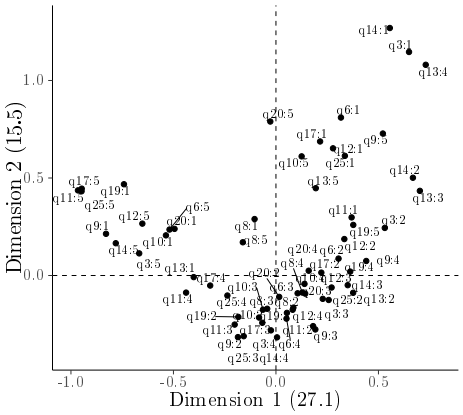
(a) Version A - Economic Scale



(b) Version B - Economic Scale



(c) Version A - Cultural Scale



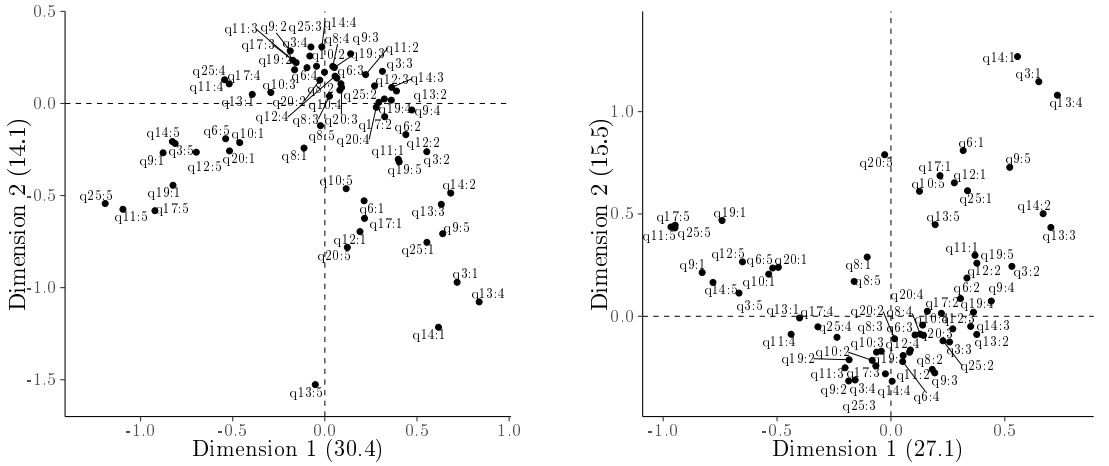
(d) Version B - Cultural Scale

Figure 6.2: Original Economic and Cultural Scales for both versions of Stem-Consult, 1st and 2nd dimension

of 0.30, while the cultural scale only does so on 5 occasions. Moreover, the H values for the scales are all below 0.30. In the same way, the crit values are high to very high on average, with values reaching 644, and only a few values beneath the cut-off point of 80. Besides, the reliability estimates are low as well, with none of them reaching satisfactory values. For the quasi-inductive scales, the cultural scale for Version A has an H of 0.47, high H values for the questions, and reliability coefficients that range between 0.86 and 0.92. The same goes for the cultural scale in Version B, having an H score of 0.42 and reliability coefficients between 0.84 and 0.90.

Now let us return to the MCA plots. From the DSV it emerged that rather than a two-dimensional solution, a one-dimensional solution would be better fitting for the data. As expected, this solution is built around the original cultural scale. Figure 6.3 shows the results for MCA on these scales.

For the cultural scale in Version B, the expected horseshoe shape is visible, though tilted. The ordinality of the order questions is maintained, and the questions 13, 9, and 19 were inverted. This is also the case for the cultural scale in Version A, which shows a similar



(a) Version A - Quasi-Inductive

(b) Version B - Quasi-Inductive

Figure 6.3: MCA for the Quasi-Inductive Cultural Scales for both versions of Stem-Consult, 1st and 2nd dimension

horseshoe (the difference in orientation can be ignored), with the only outlier being category $q13:5$, which, when projected on the main dimension, is located between $q13:3$ and $q13:4$. Thus, in this question, the ordinality could not be maintained, though the difference is small. In all, for both versions, the first dimension can thus be interpreted as the main underlying cultural dimension, running from progressive (left) to conservative (right).

6.3 About the Measures

We now return to the main argument. To measure political sophistication, we use the number of Guttmann errors each user made. These errors refer to the mistakes a user makes while answering a Mokken-type scale[7]. To illustrate the underlying idea, imagine a scale that measures conservative versus liberal positions consisting of three questions that increase in difficulty. The difficulty here means the popularity or "extremeness" of the question. To use a well-known example, the first question can ask users whether they would agree to immigrants living in the country, the second whether they would agree to immigrants living in their city, and the third whether they would agree to immigrants living in their neighbourhood. We can say that a user who would disagree with all three questions is conservative, while a user who would agree with all the arguments is liberal. Moreover, they are consistent with their answering patterns and show no Guttman errors. But, a user who would agree with the second question, but disagree with the first, shows abnormal behaviour. These cases, in which agreement on a more difficult question precedes disagreement on an easier question, we count as a Guttman error. The higher the number of Guttman errors, the more

[7] As the original 25 question scale is not a true Mokken scale, I first carried out a DSV to determine the content of the scales. Yet, instead of maintaining the original scales, the algorithm was unrestrained. So, the resulting scale was a single social scale for both versions of the VAA. I calculated the Gutmann errors based on these scales using the **GetR** package (Beller and Kliem 2013). See for a full overview of the scale validation Appendix B.

inconsistent the user is. We can thus use Guttman errors as a simple person-fit statistic to detect abnormal behaviour while answering a questionnaire (Meijer 1994). Thus, they have been used in VAA research as a proxy measure for political sophistication (Djouvas, Mendez, and Tsapatsoulis 2016; Wheatley 2016). Here, the number of Gutmann errors ranges from 0 to 166, with $\bar{x} = 20.90; sd = 16.99$ for Version A and $\bar{x} = 22.66; sd = 18.42$ for Version B, and are different ($t(2660) = -2.56$, $p < 0.05$), though the actual difference in the mean is small. So, the number of Gutmann errors is skewed towards the left, and is low, meaning that the average level of political sophistication of the users is high. For an overview of the distribution, see Appendix B.

To calculate the match between user and party, we have the choice of different algorithms. Here, we choose the *Hybrid* algorithm as proposed by Mendez (2012, 2014b), as this was the one that implemented in the Stem-Consult used[8]. This algorithm attempts to be the middle ground between an algorithm that would use a Euclidean logic, and an algorithm that would use a Scalar logic. The resulting matches can theoretically run from -100 to 100. Here, -100 means that in all cases where the user completely disagreed the party completely agreed and in all cases where the user completely agreed the party completely disagreed; $+100$ means that in all cases where the user completely agreed the party completely agreed and in all cases where the user completely disagreed, the party also completely disagreed; 0 means that in all cases, either the party or the user was neutral while the other completely agreed or disagreed (cf. Mendez 2012, 2014b).

6.4 Mode of Analysis

To estimate the effect of the positive-negative condition on the three levels of users, parties and questions, we can adopt two different strategies: either we estimate a complex multilevel model in which we include parties, questions and users, or we estimate a simple linear model, but run it many times for each combination of question and party. This latter approach is what Andrew Gelman calls the *secret weapon* (Gelman and Hill 2007) and Tufte (1990) calls "small multiples". The common idea of both is to cut up a complex dataset in many smaller and simpler ones, and then analyse them visually. Not only does this prevent relevant information to be buried in the coefficients of a complex model, but graphs have also been known to be effective conveyors of information (Gelman, Pasarica, and Dodhia 2002; Kastellec and Leoni 2007). The result of this is a collection of small graphs ordered by party and question, with the users supplying the data within.

Within the small visuals, we have to depict how changing the valence of the question influences the match between parties and users. To do so, we assign the questions in each version of the VAA to a positive-negative condition that is 0 when the question is positive and 1 when the question is negative. We then run a linear model with the positive-negative condition as the main independent variable and the match between the user and the party as the main dependent variable. Besides, the variables age, sex, education, interest in politics, and the number of Guttman errors the user made, as well as their interactions with the positive-negative condition, were also included.

[8] Note that I did extra calculations with the three other possible algorithms as described by Mendez (2012, 2014b), but they did not show any significant differences.

Still, while this linear model does allow us to observe the effect of the experimental condition between the parties (as we calculate the match between user and party for each party separately), it does not distinguish between the separate questions. To do so, we use *marginal effects*. Marginal effects show us the effect of a single variable, in this case, the positive-negative condition, while we hold the other variables constant. Put differently, the marginal effect tells us the contribution of the positive-negative condition to the match between the user and the party. As such, a marginal effect of 0 means that the positive-negative condition does not contribute to the match i.e. there is no difference between the positive and negative condition of the question. A positive effect means that if the positive-negative condition changes from 0 (positive) to 1 (negative), the match between the user and party goes up, while a negative effect means that if the positive-negative condition changes from 0 (positive) to 1 (negative), the match between user and party goes down. Marginal effects come in three broad types: marginal effects at representative values (MER) give the marginal effect at values of positive-negative condition that are theoretically interesting, marginal effects at means (MEM) gives the marginal effects at the means of the covariates, and average marginal effects (AME) provide the average of all the marginal effects calculated for all observed values of the positive-negative condition (Leeper 2018a). Here, we use the AME as it not only gives information of the full range of values of X as MER does, it also provides us with a single value that can be interpreted, as MEM would. I calculated the marginals using the **margins** package in R (Leeper 2018b).

To establish what the influence of political sophistication is on the positive-negative condition, we have to establish what the conditional effect of political sophistication on the positive-negative condition is. This is modelled using the interaction terms as shown in the model in Appendix B. Using the **interplot** package in R (Solt and Hu 2018), we can calculate for each value of the conditional effect what the coefficient for the positive-negative condition would be. This allows us to plot these conditional effects, together with their 95% confidence intervals. The result will be a similar plot of graphs as with the marginal effects, grouped per question and per party. In these plots, conditional effects are present when the line is either increasing or decreasing (in the case of a numeric condition), or when the values of each group are different (in the case of a categorical condition). This indicates that for different values of the conditional variable, the positive-negative condition would show differently - i.e. the conditional variable influence the positive-negative condition.

6.5 Results

I will discuss the results of our analysis in three steps. To begin with, I will consider the effect of question-wording on the match, discussing each question separately. Second, I will consider the influence of political sophistication on the effect of question-wording, following a similar strategy. To conclude, I will consider the effects of the extra variables of age, political interest, education and sex.

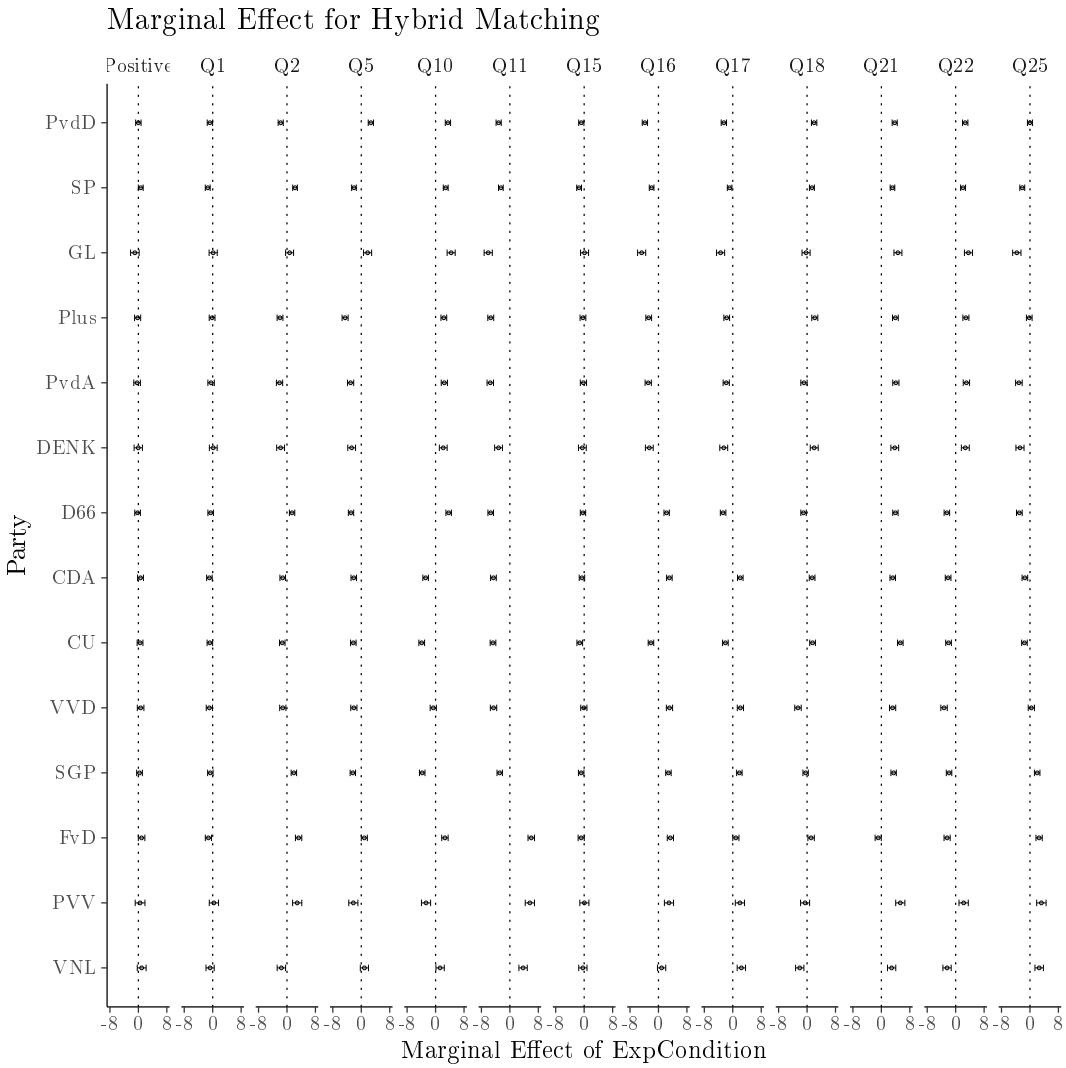


Figure 6.4: Marginal Effect of the positive-negative condition on the match between the user and the party using he Hybrid Algorithm.

6.5.1 Effects of Question-Wording

Figure 6.4 is the collection of graphs that show the average marginal effect of the positive-negative condition for each question separately, including a 95% confidence interval (these can also be found in Appendix B). For each question, I calculated the match by adding that question to the standard set of 13 questions. In most cases, the effects are significantly different from 0 but are also small. The largest negative effect is -5.51, while the largest positive effect is $+5.50$. These effects are especially small considering that the matches are on a $200 - point$ scale, ranging from -100 — full disagreement, to $+100$ — full agreement.

For Question 1 the largest effect is for the SP (-1.37), meaning that adding that question in the negative version decreases the average match between the voter and the SP. For the other parties, whether the question is in its positive or negative form does not influence the average match between the user and the party. One of the reasons this question shows so little difference between the positive and negative version is that the topic — subsidies for arts and culture may be cut — was not particularly salient in the election.

For Question 2 — The public broadcaster in its current form has to be maintained — the highest positive marginal effect is for the FvD (+3.22) and PVV (+3.34) and the lowest for the PvdD (−2.02) and the PvdA (−2.07). As such, FvD and PVV would profit if the question would appear in the negative - the public broadcaster has to be abolished. This would fit in their programme, as both parties were critical against the current system of public broadcasting. The PVV mentions in its manifesto that it wishes to spend "no public money (...) [on] broadcasting" (Partij voor de Vrijheid 2016) while FvD talks about wishes to "clean-up" (Forum voor Democratie 2016, p.16) the public broadcaster. The PvdD and PvdA, who would gain when the question is positive, are more positive about the public broadcaster, with the PvdA wishing to maintain the broadcaster and invest an extra 100 million in it (Partij van de Arbeid 2016).

For question 5, most parties would profit if the question is asked in its positive version — mortgage relief has to be maintained — while PvdD (+2.52) and GL (+1.71) gain most when its negative version is asked. When the question reads the mortgage relief has to be maintained 50*Plus* (−4.52) and D66 (−3.36) would profit. Once more, this is consistent with their opinions as both PvdD and GL support a (gradual) abolishment (GroenLinks 2016; Partij voor de Dieren 2016), while 50*Plus* notes "not to interfere" with mortgage deduction (50Plus 2016, p.5). The effect is not consistent, however, as the PvdA (−3.22) and SP (−2.22) would profit from a positive version while supporting abolishment (Partij van de Arbeid 2016; Socialistische Partij 2016).

For question 10 the largest effects are for CU (−3.45) and GL (+4.01). CU would profit when the question calls for soft drugs to be allowed, while GL profits when the question calls for them to be forbidden. This is the reverse of what is in their respective manifestos (ChristenUnie 2016; GroenLinks 2016). The same is true for the other parties on the positive (D66, PvdD) and negative (SGP, CDA) side.

Question 11 shows the largest and most clear effects. All parties would profit from the positive version - The Netherlands has to remain in the European Union — while only the PVV, VNL and FvD would profit from its negative — The Netherlands has to leave the European Union. This not only fits in with the fact that these are the only 3 parties supporting leaving the European Union, but the strength of the effect on both sides is also consistent with the saliency of the issue for these parties as not only are the effects for the PVV (+5.35), VNL (+3.52) and FvD (+5.22) large, so are those of parties supporting the EU such as GL (−5.51), the PvdA (−4.75) and D66 (−4.89).

For question 15, only limited effects can be found. Most of the effects are small and only 5 parties, SP, CU, SGP and PvdD are significantly different from 0, though even the largest effect, for the SP at −1.87 is small. This is interesting insofar as the question, whether for environmental taxes may (or may not) be raised, should be a salient topic for the green parties GL and PvdD.

For question 16 on whether the government can leave it to companies (or is allowed to force them) to take energy saving measures, the effects are as expected but reversed. Parties gaining when the question reads the government should "be able to force the companies" are VVD (+3.23), PVV (+3.51) and FvD (+3.38) are parties to the economic right of the political spectrum, while the parties that gain when the question reads the government should "leave it to" the companies, GL (−4.91) and PvdD (−4.10), are both parties to the economic

left with strong environmental profiles. Similarly, while the GL manifesto calls for strong government intervention with regard to energy-saving measures, the VVD leaves this to the companies themselves (GroenLinks 2016; Volkspartij voor Vrijheid en Democratie 2016). As such, one would have expected the effects to be reversed.

The AME for Question 17 shows quite the opposite as the effects of question 16. Profiting most from the positive version of the question — that the coal plants in the Netherlands should remain open — are the VVD (−2.24), PVV (−2.52) and VNL (−2.64). Parties gaining most when the negative version — that the coal plants should be closed — is asked, are GL (+3.60), followed by D66 (+2.98) and the PvdD (+2.79). This would be as expected, given that the VVD and VNL are positive on coal plants, while GL has a negative stance on them.

Question 18, about the loan system for students, shows smaller effects. Here, PVV (−1.08), VNL (−2.20) and VVD (−2.53) profit when the question reads the loan system should be maintained, while 50Plus (+1.69), PvdD (+2.16) and DENK (+1.87) profit when it reads it should be abolished. This makes sense as the VVD supports the loan system, while DENK is against the system.

For Question 21 all parties would profit if the question exists in its negative form, with even the AME of the FvD (−0.91) being significantly not different from 0. This is logical given for this question — there can be no cuts in spending on social work — all parties agree that such cuts should not take place.

For Question 22, on the own risk in health care, parties profiting when the question reads the own risk should be maintained are the VVD (−4.38), D66 (−3.18) and VNL (−3.25), while parties profiting from the wording in which it should be abolished are GL (+4.83), PvdD (+4.01) and PvdA (+4.06). This is in line with expectations, as VVD, D66 and to a lesser degree FvD are proponents of the own risk, while GL, PvdA and the PvdD are proponents of abolishment.

For Question 25, the pattern is again as expected, with VNL (+2.71), FvD (+2.47) and PVV (+3.54) profiting when the question reads the multicultural society is not a good thing, which is consistent with their manifestos and general stance, while GL (−3.70), PvdA (−2.94) and D66 (−3.08), each supporters of the multicultural society, profit when the question states that the multicultural society is a good thing.

With regard to the explicit and implicit negatives, there are no discernable differences. In two cases (questions 22 and 25), the explicit negatives are significant, while in two others (questions 1 and 15), they are not. The highest negative and positive effect for the explicit negatives (−3.70 and +5.50) is different from the highest negative and positive effect for the implicit negatives (−5.51 and +5.35), but not overly so.

We can explain the observed effect best by focusing on the salience of the questions they represent. In general, as soon as the salience of the issue is high, the wording effect increases. This means that when the issue is salient for a voter, they are more affected by *how* the issue is worded than when the issue is not salient to them. We can thus accept H1: there is a difference in the match between versions of questions are either positive or negative. Yet, these differences for both parties and questions in different ways that do not appear to show any kind of coordinated behaviour. In other words, it cannot be predicted how the wording of certain questions will affect a party. For H2 we have to reject the hypothesis that both are different for there is no difference between the implicit and explicit negatives. We can

accept H3a and H3b as we find differences in the effect of the question wording between both parties and questions.

6.5.2 Conditional Effects of Political Sophistication

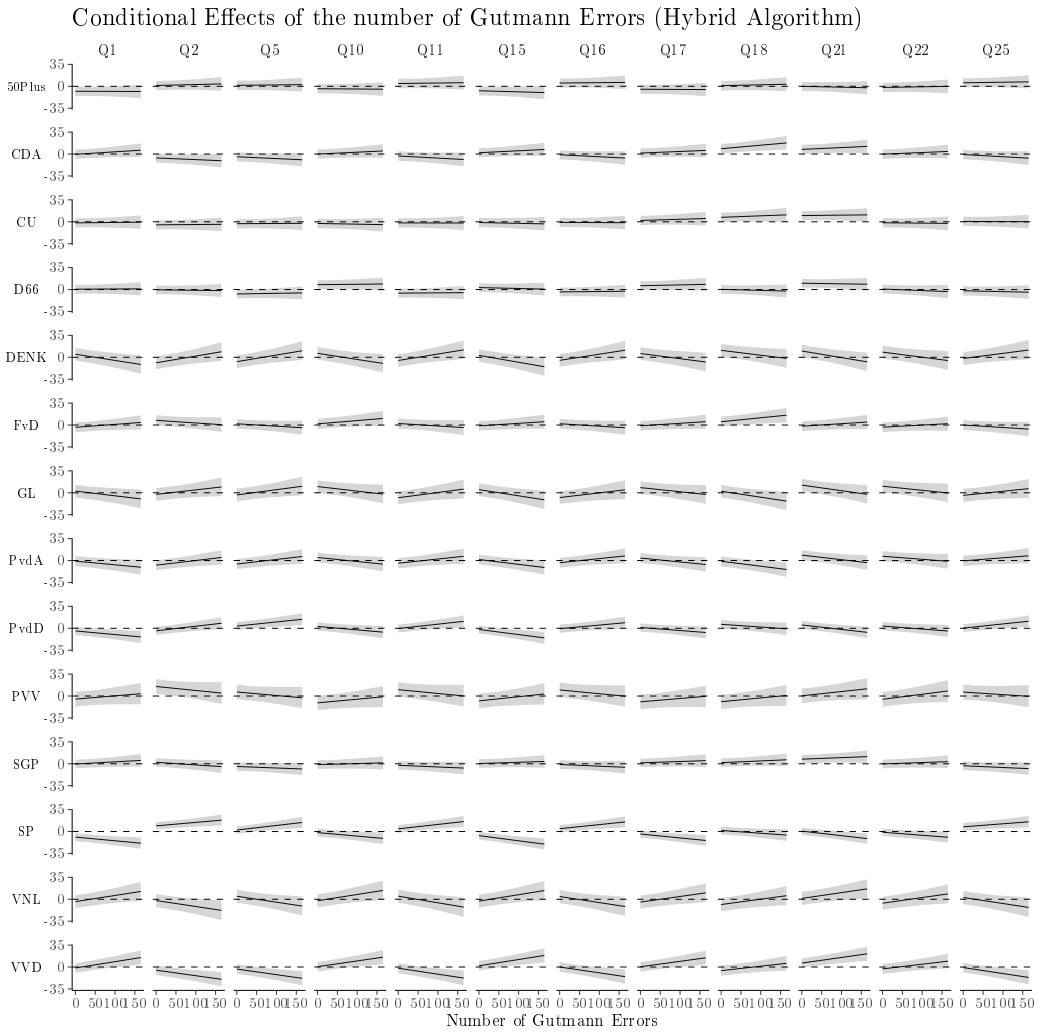


Figure 6.5: Conditional Effect of the # of Gutmann errors on the size of the coefficient of the positive-negative condition on the match, with 95% confidence interval

Figure 6.5 shows whether the marginal effect of the number of Gutmann errors is conditional on the respondent's level of political sophistication as measured by the number of Gutmann errors. In the figure, the x-axis shows the number of Gutmann errors, with a higher number of Gutmann errors indicating lower political sophistication, and the y-axis shows the size of the coefficient of the experimental condition on the match. The y-axis runs from -35 to 35. A unit increase here indicates that the match between the user and the party increases 1% if the question changes from positive to negative. Similarly, a unit decrease means that the match between the user and the party decreases 1% if the question changes from positive to negative.

As with Figure 6.4, each of the effects is different for each party and for each question. Still, the strength of the association seems to differ by party, and not by the question. As

such, DENK, PVV and VNL show rather large increases and decreases, while for the other parties the conditional effect is rather small. This means that in the cases of the CU, SGP, and D66, there is no difference in the effect the positive-negative condition has on the match with regard to the number of Gutmann errors. In other words, the level of political sophistication is unconditional on whether the wording changes: for each user the effect will be roughly the same.

An example of a negative effect is the case of question 15 for SP, the coefficient of the positive-negative condition is -6.57 when the number of Gutmann errors is 0, meaning that the average match with DENK for those users who have 0 Gutmann errors will be 6.57 less when the question changes from "on environmental measures taxes may be raised" to "on environmental measures taxes may *not* be raised". Yet, as the level of political sophistication goes up, so increases the negative effect. At 65 Gutmann errors, the size of the coefficient is -11.92, and finally, at 165 Gutmann errors, it has reached -20.1. This means that the lower one's political sophistication, the greater will be the effect of the question-wording. Equally, we find a positive effect for the PVV in the case of question 21. Here, at 0 Gutmann errors, the coefficient is 0.40, going through 4.80 at 65 errors, until at 165 errors it reaches 11.58. Thus, the less politically sophisticated one is, the higher the match will be when the question 21 changes from positive to negative. In other words, less politically sophisticated one is the more likely is one to be matched to the PVV when the questions change from "there can be cuts in spending on social work" to "there can be no cuts in spending on social work".

In this case, the coefficient started negative and remained that way. More interesting is the case when the sign switches. This is, for example, the case for the DENK in question 21. At 0, with no Gutmann errors, the coefficient is 10.03 and keeps decreasing until 95 Gutmann errors. Here, the sign switches and the decreasing continues until reaches -7.34 at 165 Gutmann errors. This indicates that until 95 Gutmann errors, changing the question from positive to negative increases the match between user and party, while after 95 Gutmann errors, the match will decrease instead.

Given that political sophistication does influence the effect of the question wording, we can there accept H3c. Again, also referring to H2, if we look at the differences between the implicit negatives and the explicit negatives, we find no significant differences between them.

6.5.3 Conditional Effects of Age, Political Interest, Education, and Sex

Additionally, I tested the conditional effects of age, political interest, education, and sex on the marginal effect. Similar plots like the one shown of the number of Gutmann errors can be found in Appendix B. Starting with age, we find little to no effect except for question 22. Here, for the question on whether the own risk in healthcare should be abolished or maintained, there was a strong effect for all parties. The effect was positive for $50Plus$, DENK, GL, PvdA, PvdD, PVV and SP and was negative for CDA, CU, D66, FvD, SGP, VNL, and VVD. This means that as soon as the age of the user increases, the effect of switching from a question that reads the own risk should be maintained to one that should be abolished increases, for the first group of parties. This is logical insofar as these parties all completely disagree with maintaining the own risk - in other words, they want to abolish it. As a result, when the question reads it should be abolished, especially older users will show

a higher match with this party.

On political interest, I find that in no case is there a significant difference between the 5 different levels of political interest, and even in cases where the conditional effects are different from 0, such as for the PVV in case of question 2, the effect is constant over all levels of political interest. This is interesting insofar that while political sophistication, of which political interest is supposed to be a part, does influence the marginal effect, political interest itself does not. A similar conclusion can be drawn for education, where there is little significant difference between the categories, and the change between them is minimal. The main exception to this is question 11. This question, on whether soft drugs should be legalized or forbidden, shows a decreasing pattern except for the FvD, PVV and VNL. This means that except for those parties, the match between user and party decreased as soon as the question reads soft drugs should be forbidden. However, I could find no sources that would explain why this would be related to these parties. For sex, I find no conditional effects for any of the questions or parties.

6.6 Conclusion

In this chapter, we saw that similar issues worded in a different way lead to different levels of match, though the differences are small. Here, we showed this was the case for a simple change in wording: whether we worded the questions positive or negative. Besides, by separating parties, questions and users during the analysis, I was able to show that these differences vary between parties and questions. Some questions show large effects, while other questions show only small effects. Yet, I have not found any consistency why certain questions show larger effects than others. Within the questions, there are variations per party. These variations depend on which version of the question is most related to the party. As such, a party in favour of leaving the European Union will receive a higher match when the question reads leave instead of remain.

Moreover, levels of political sophistication influence the effect of the question-wording. The lower the political sophistication of the user, the higher the effect of the question-wording. This is problematic because designers develop VAAs to help these kinds of users to find their way around the political landscape. The prominent effect of question-wording means that it might well be that the positions of the user as calculated by the VAA depend as much on the way we word the questions as on the actual content of the questions. For the broader literature on congruence, it might well turn out that current findings on a low match between parties and voters with a low political sophistication might be dependent on how the designers arrived at the user positions.

From here, there are several avenues for further research. One avenue is to establish whether other ways of wording questions have similar effects. One could think of short questions versus long questions. Or of different ways of formulating questions in a grammatical sense. Another avenue might be to also position the parties on both questionnaires. Given that we saw party effects here, it might well be that coders position parties in a different way when the questions have different wording. These differences might then annul any of the effects seen here. For VAA developers the implications of this chapter relate to whether to include negative questions at all. Problematic here is it is impossible for them to not include

any negative questions at all. Some questions only exist in their negative form in the debate and including them in their positive form could be confusing for the users. Also, there is no reason to assume that negative questions are problematic. They are only different from their positive counterparts. The best designers could do is at least to consider which questions they should make negative and why. They could then make these reasonings available on the website of the VAA together with the other design choices the designers have made. Moreover, it would be advisable to prevent using a negative or positive wording if it would favour a certain party.

7 | Visualization in Theory

In this chapter and the next, I focus on *how the visualization influences the way in which the user perceives their match with the parties.* In this chapter, I will discuss the underlying theory, while the next chapter reports on the results of an experiment. Here, I start by discussing how users look at visualizations by drawing on the theory of graphical comprehension. Based on this, I will give several recommendations for designing visuals. Finally, I will discuss what VAA aim to visualize and give examples of various VAA visualizations.

7.1 Graphical Comprehension

Graphical comprehension[1] is the way in which users understand graphs (or other kinds of figures) and assign meaning to it (Glazer 2011). Figure 7.1 visualizes the process of graphical comprehension (Shah and Hoeffner 2002; Shah, Freedman, and Vekiri 2005; Carpenter and Shah 1998). This process goes as follows. First, we encode the graph into a clear description. We do this with bottom-up (**A**) and top-down (**B**) processes. During the bottom-up processes, we extract the information from the graph. This can be explicit information like numbers or relations between objects such as one bar being higher than the other. Or, it can be inexplicit information where we have to do some calculations first. This happens when we want to see a relationship between two lines in a line graph, which is not shown, but which we have to infer (Shah and Carpenter 1995). Bottom-up processes are thus dependent on the type of graph we look at. Different graphs lead to different descriptions, even if they contain the same information (Shah 2002). Top-down processes are dependent on previous knowledge or expectations we have about the graph. This influences the way we view the graph (Shah 2002, p.180). When we see a graph, such as a bar chart, we have certain ideas of what it should look like. We know we need to look at the axes for the value of the bar and compare the bars with each other to know which one is larger or smaller. This experience steers us in looking for things in the graph that someone who has never seen a bar chart before would not look at. These top-down and bottom-up processes work alongside each other to convert the graph into a description. In this description, we separate all the aspects of the graph and categorize them according to the function. Thus, sizes go with sizes, shapes with shapes, distances with distances, etc. (Pinker 1990, p.77). This description is then matched to a

[1] Besides graphical comprehension there are the concepts of graphical perception and graphical cognition. Graphical perception is the ability of the user to extract information from graphs without a conscious effort. Graphical cognition refers to the ability to extract information with a conscious effort (Cleveland and McGill 1986). Here I use the more encompassing concept of graphical comprehension to denote any type of graph reading.

graph schema. A graph schema contains all the conceptual relations that the graph tries to communicate (Pinker 1990). We construct these graph schemas based on experience and previous encounters with graphs. The match-process (**C**) matches the description to each of the schemas and assesses which one fits best. This is the time when we first assign a label to the graph (such as bar graph, spider plot, or line graph)(Pinker 1990, p.101-102).

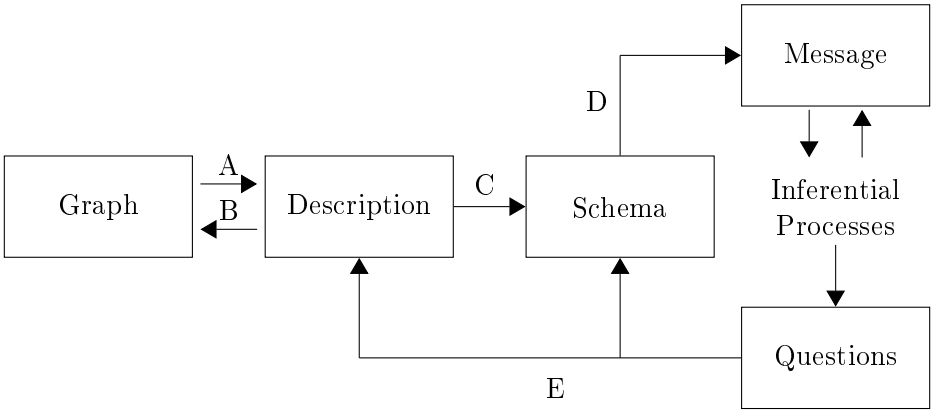


Figure 7.1: Model of Graphical Comprehension based on Carpenter and Shah (1998), Freedman and Shah (2002), Glazer (2011), Lee et al. (2016), McMahon, Stauffacher, and Knutti (2015), Pinker (1990), Shah (2002), and Shah, Freedman, and Vekiri (2005). A: Bottom-Up Encoding, B: Top-Down Encoding, C: Matching, D: Message Assembly, E: Interrogation

With the graph schema, we can then assess the graph. First, we establish the message of the graph. We do so during the message assembly process (**D**). This process translates the visual information into conceptual information (Pinker 1990, p.102). It does so by scanning over the graph and seeking items described in the graph schema. As soon as it finds them, it adds them to the message. Besides scanning, the message is also formed by inferential processes. This is when we use the explicit information in the graph to derive inexplicit information. This happens when we calculate the increasing distance between two lines to see which one is growing faster. Finally, the inferential processes help us answer the questions we have about the graph such as "with which parties do I have the highest match?". If we cannot answer these questions the interrogation process (**E**) goes back to the graph schema and description to check for any information that we did not observe the first time. The total process of graphical comprehension repeats itself many times. Each time the information in the process uses becomes more complex until we arrive at a complete and coherent interpretation of the graph (Carpenter and Shah 1998). So, the more complex a graph, the more often we need to repeat the process, and the longer it takes to read the graph.

Five factors influence the process of graphical comprehension: domain knowledge, graphical knowledge, data complexity, task demands, and graph characteristics (Freedman and Shah 2002; Shah, Freedman, and Vekiri 2005; Glazer 2011). *Domain knowledge* refers to all the knowledge we have about the domain the graph belongs to. For example, we use our knowledge of the weather when we look at a graph indicating temperature. Or we use our knowledge of elections when we look at a graph showing electoral turnout. This knowledge then defines what we see and influences the top-down processes that transform the graph into a description. The effects of this knowledge can be both positive and negative. On the

positive side, more knowledge means that we can make more accurate estimates of the relationship in the graph. This way we can get more out of the graph than someone without the knowledge. On the negative side, this knowledge makes us less willing to accept information that challenges our prior beliefs (Shah, Freedman, and Vekiri 2005, p.458-461; Glazer 2011). Moreover, if certain relations we expect are not present in the graph, we sometimes project these relations unto the graph ourselves (Shah 2002). *Graphical knowledge* refers to more general knowledge about the graph. The more types of graphs we know and the more often we see them, the higher our graphical knowledge. This knowledge is what makes the schemas: the more knowledge we have, the more schemas we can make. If we have low graphical knowledge and lack schemas, the process of interpretation becomes difficult. This makes graphical comprehension demanding, causing misconceptions to arise (Baker, Corbett, and Koedinger 2001). If we have a high graphical knowledge, the interpretation is easier, and we can provide "deeper" descriptions of the graph (Glazer 2011, p.194). Yet, as with domain knowledge, this can lead us to misrepresent graphs to make them fit with our existing schemas (Freedman and Shah 2002; Shah, Freedman, and Vekiri 2005). *Data complexity* is the amount of information stored in the graph. It refers to both the information that we can observe (such as points) and the relations we can draw between different pieces of information. More information and more relations make a graph more complex. Noise, which is all the unnecessary information in the graph, also increases complexity (Shah, Freedman, and Vekiri 2005). For complex graphs, we thus need a higher level of graphical and domain knowledge to understand them (Carpenter and Shah 1998). As a result, most users prefer simple visualizations over complex ones (Shamim, Balakrishnan, and Tahir 2015). But we cannot always avoid complex graphs as complex decision-making often involves complex data. *Task demands* depend on the task the user wants to perform. These tasks can be different, even for the same graph. For example, one task might be to read off the values of a certain bar in a bar chart. Another task might be to find out the relative differences between all the bars. The second task requires inferential processes and thus increases the task demands. Even more demanding is it when the user does not know what their task is. This happens when we see a certain type of graph for the first time. We then have to discover what we want to know and what our task is. *Visual characteristics* are what the graph looks like. These characteristics can be very general, such as the position, length, area, volume, and shading of the information in the graph (Cleveland and McGill 1984, 1985). They can also be more specific, like whether the graph has a legend, which aspect ratio it has and if it uses colour (Shah, Freedman, and Vekiri 2005; Shah and Hoeffner 2002). These visual characteristics influence all the stages of graphical comprehension (Freedman and Shah 2002; Shah, Freedman, and Vekiri 2005). This causes them to be the most investigated (Shah, Freedman, and Vekiri 2005).

7.2 Designing Visuals

We can use the process of graphical comprehension to design better graphs. Table 7.1 shows several suggestions made by scholars. Most important of these is to "above all else show the data" (Tufte 2001, p.92). This means the graph should show the data, and not anything else. Linked to this is the advice to maximize the data-ink ratio. This is the ratio between the ink spent on displaying the data and the total amount of ink used in the graph. We should

erase both non-data ink and redundant data-ink as far as is reasonable. The result of this is a graph which draws the viewer towards the data and not to anything else (Tufte 2001, p.91-106). To make this process easier, the amount of data in the graph should be as small as possible. In practical terms, this means not showing more than one kind of data at the same time. These advices reduce the task demands placed on the user and thus increase graphical comprehension (Shah and Hoeffner 2002).

Recommendation	Sources
Show the data	Tufte (2001), Shah and Hoeffner (2002), Freedman and Shah (2002), and Cleveland (1985)
Maximise data-ink ratio	Tufte (2001)
Minimise information	Shah and Hoeffner (2002)
Choose the right format	Peebles and Cheng (2003) and Shah and Hoeffner (2002)
Use multiple formats	Cleveland and McGill (1985), Robbins (2005), Peebles and Cheng (2003), and Shah and Hoeffner (2002)
Consider visual objects	Cleveland (1985), Cleveland and McGill (1984), and Pinker (1990)
Consider legends	Carpenter and Shah (1998) and Shah and Carpenter (1995)
Clear scales	Cleveland and McGill (1985)
Consider aspect ratio	Cleveland (1993) and Shah and Hoeffner (2002)
Consider colours	Cleveland (1993) and Shah and Hoeffner (2002)
Explain the graph	Shah and Hoeffner (2002)
Consider individuals	Shah (2002), Shah and Freedman (2011), and Ziemkiewicz et al. (2012)
Revise and Edit	Tufte (2001), Cleveland and McGill (1985), and Robbins (2005)

Table 7.1: Overview of recommendations for good visuals

When we have decided which information we include, we have to decide on the right format. If we want to show change over time, a line chart is the best option, while if we want to show quantities, we should use bar plots. Often there is more than a single way to visualize the data. For example, we can visualize quantities not only in a bar chart but also in a box plot. If we do so, we give the user a different perspective on the same information. This makes it easier for them to get out all the information. Then, we have to consider the visual objects. Table 7.2 shows these visual objects. Cleveland (1985) suggests a ranking of them based on the principles of Gestalt psychology[2]. The higher in rank, the easier it is to for us to read the information encoded in the object. Thus, it is easier to read a position of

[2]Gestalt psychology is a branch of psychology that tries to understand how we can make meaningful perceptions in a chaotic world. It proposes that the perception of objects does not depend on light and dark, but on how similar and close they are to each other (Das 2009). It influenced but is very different from, Gestalt therapy, which is a form of psychotherapy.

a dot along a scale than to determine the length of a bar. Area and colour are the hardest to interpret, and we should avoid them if possible. This explains why spider graphs are so problematic, as we will see later.

1	Position along a common scale
2	Position on identical but non-aligned scales
3	Length
4	Angle, Slope
5	Area
6	Volume, Density, Colour saturation
7	Colour hue

Table 7.2: Accuracy of visual objects (Cleveland 1985)

Next, we have to consider the layout of the graph. This refers to the inclusion of a legend, scales, the aspect ratio, and colours of the graph. Carpenter and Shah (1998) advise designers to avoid legends and label the lines and bars instead. Placing labels next to their objects makes them easier to relate to each other, which makes reading the graph simpler. But we should not put too many labels on the graph as this will make the graph too complex. Thus, simple letters and numbers (such as M/F or different ages) can go in the graph, while longer text, such as country names, belong in a legend. Scales are important as they are the basic elements of most graphs. Important here is the number of tick marks and where to start the scale. Most scales start at 0, but this is not necessary[3]. Also, the scale should ensure the data does not show more white space than necessary. Related to this is the decision on which aspect ratio we should use. These aspect ratios are especially relevant in a line graph, where a change from a 4:3 to 16:9 aspect ratio makes a decreasing line seem less decreasing. Besides, not all aspect ratios are easy to compare trends. Cleveland, McGill, and McGill (1988) find those aspect ratios where the average slope is around 45° generate the best results. Finally, we can use colour to group variables together. While Table 7.2 warned us to not use it to distinguish objects, we can use it to separate groups of objects (Kosslyn 1994).

Finally, we should ensure the graph is self-explanatory. If it is not, and we cannot find a way to improve, we have to explain it. It is better to provide a longer explanation than to risk users misunderstanding the graph. Also, we have to consider who these users are. Ziemkiewicz et al. (2012, p.91) note that "there is no single, representative user". As a result, some visualizations will work better for some than for others. To improve the visualization, we should thus allow users to give feedback or allow them to change the graph. We can then use this feedback to revise and edit the graph.

7.3 The Political Space

Now I have discussed how users understand graphs and how designers should design them, there is one more step we need to consider. This is what it is designers want to visualize. In the case of Voting Advice Applications, this is the *political space*. We use this political space often when we walk about politics. When we say "you are more left-wing than I am" or "party

[3] Though, as Huff (1954) shows, this can lead to misleading graphs.

A moved to the right", we locate ourselves and political parties in a political space. What this political space looks like depends on the number of dimensions and what they mean (Benoit and Laver 2006, 2012; Gabel and Hix 2002). For example, we can have a political space with two dimensions: an economic dimension and a social dimension. As the number of possible dimensions is infinite, the number we use depends on how much we are willing to simplify. There are no set rules for this decision and thus no "one true dimensionality" (Benoit and Laver 2012, p.216).

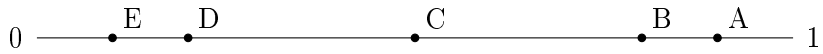
We distinguish between high-dimensional spaces and low-dimensional spaces. In the first, the dimensions are distinct issues such as opposition to abortion or support for the death penalty. In the second, the dimensions are combinations of interrelated issues (Gabel and Hix 2002; Gabel and Huber 2000). Doing this lowers the degrees of freedom of the space making it easier to understand (Benoit and Laver 2012). We can construct these dimensions in two ways. In the first, we assign issues to those dimensions they belong to according to a certain theory. For example, if we want to construct a left-right dimension, we assign to it issues related to left and right issues. In the second, we assign issues to dimensions using data-reduction methods. We then interpret these dimensions based on the issues they contain. We call the first type ex-ante dimensions and the second type ex-post dimensions. Ex-ante dimensions have the advantage that we can improve them by including the knowledge we have about the political space. Yet, if this knowledge is wrong or out-dated, the dimensions become problematic. This is the critique raised against the left-right dimension (rile) used by the Manifesto Project. Budge and Laver (1992) designed this dimension based on the political space in Western Europe at that time. But while more than twenty years later this political space has changed, the dimension stays the same. Mölder (2016) argues this makes the dimension flawed. Yet, Budge and Meyer (2013) argue this does not matter as long as the dimension is useful - a notion debatable at the least. Ex-post dimensions do not have this problem as we construct them from the data. Here, the problem is we are neither in control of the number of dimensions nor always agree on what they mean. Different interpretations can lead to different labels for the same dimension.

In the next two sections, I will discuss how VAAs use high-dimensional and low-dimensional visualizations and discuss their advantages and disadvantages.

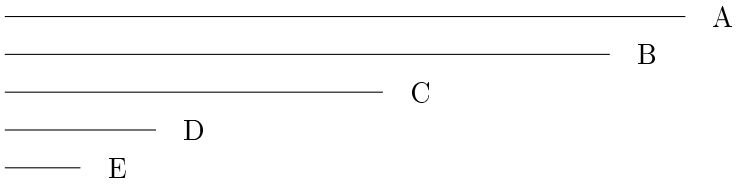
7.3.1 High-Dimensional Visualizations

Each VAA starts out as high-dimensional. This is because each of the questions is a dimension on its own (Louwerse and Rosema 2014). For mathematicians, the problem would end here. For VAA designers, the problem is how to visualize all these dimensions. We saw that one solution is to cluster them together to make the visualization low-dimensional. Yet, this assumes that certain questions are dependent on each other. Our opinion on one question then correlates with our opinion on another question. What if we cannot or do not want to, assume this and want to keep all the questions independent? The solution here is to compress all the dimensions into a single dimension. To see how this works, imagine that there would be 30 dimensions ranging from 0 to 1. If I agree with a party this will give the party a score of 1. If I disagree with a party this will give the party a score of 0. To get the position of the party on the dimension we add all these scores and divide them by the number of questions.

Thus, if I agree with a party C on 15 questions and disagree on 15 others, the position of that party on the dimension would be $\frac{(0*15)+(1*15)}{30} = 0.5$. We can visualize this follows[4]:



As 1 indicates that we and the party agree on all questions, the figure shows that we are closest to party A and farthest from party E. But, while this allows us to compare the parties to ourselves, we cannot compare the parties with each other. Moreover, if we position the parties on a single line, we claim that parties positioned next to each other are somehow related. This does not need to be the case. Thus, we split the line, which results in a bar plot:



In this graph the information is the same while the suggestion of a relation between the parties is gone. In most VAAs, bar plots often also show a value (such as a percentage) to give the user more accurate information. Also, many VAAs order the bars from highest to lowest and use colours to distinguish between low and high matches. While very clear, this also leads to a rather explicit voting advice: that of the party on top (Germann and Mendez 2016). In part because of this reason, designers of VAAs turned to low-dimensional visualizations.

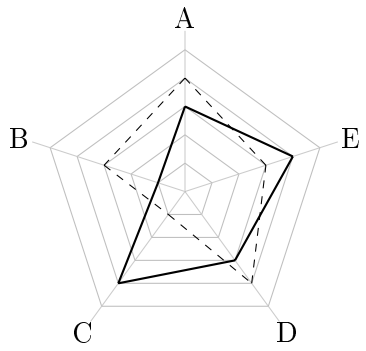
7.3.2 Low Dimensional Visualisations

In low-dimensional visualizations, we cluster together questions that relate to each other. The number of dimensions this results in defines which visualization we use. Most low-dimensional visualizations have two dimensions and show a map. Sometimes, we have more than two dimensions, such as five or six. In this case we can use a *spider graph* [5]. Spider graphs have an origin at their centre with axes radiating out from them (Draper, Livnat, and Riesenfeld 2009, p.763). Each of these axes represents a dimension. Note that these dimensions are different from the additive dimension I discussed earlier. There, a position on the dimension indicated the level of agreement between the user and the party. Here, a position on the dimension indicates the level of agreement with the dimension for both the party and the user. In a spider graph, we plot the positions of the user or the party on each of the dimensions and connect them with lines to form a polygon. These lines themselves do not signify anything and are there to aid visual interpretation. Also, we can draw grid lines to connect the different axes and create the web-like look.

[4] Note that this is a simplification as most VAAs use Likert scales. Also, we can calculate the agreement between the party and the user in different ways. The way we choose depends on how we understand political competition (cf. Mendez 2014b, 2017).

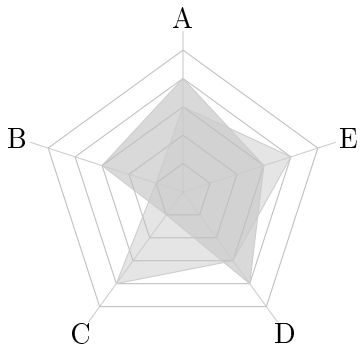
[5] Other names for this type of graph are: *web chart*, *spider chart*, *star chart*, *star plot*, *cobweb chart*, *irregular polygon*, *polar chart*, and *Kiviat diagram*.

While spider graphs look appealing, they have some problems. First, if we show more than two shapes, the graph can become cluttered. This means that we can only show one party (and the user) at the same time. Thus, to compare parties, we have to switch between all these graphs. We also have to remember what the shape of the other parties looked like. This is more complex than when we see these shapes next to each other and thus gives a higher chance of drawing the wrong conclusion (Munzner 2015, p.131). Second, while the idea is that users compare *shapes*, they might start to compare areas as well. Yet, as the axes mean different things, a similar area does not mean that the shapes are the same, as the following figure shows:

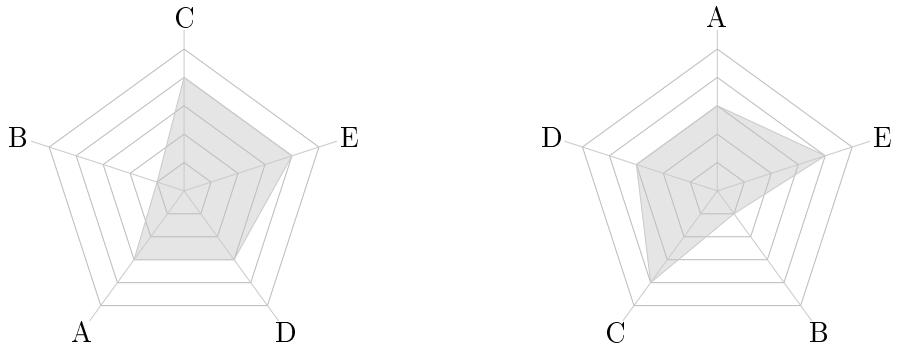


Here, the party (dashed line) and the user (solid line) have polygons with a similar area, but with different positions on each of the axes. Only looking at the area thus leads to wrong conclusions (Albo et al. 2016).

Third, changes in the values on the axes are not proportional to the changes in areas shown. As Huff (1954) notes in his seminal *How to Lie with Statistics*, when we want to visualize a twice as large value, it would be logical to double the size of the diameter. Yet, this would double the actual area of the object in a two-dimensional visualization and triple it in a three-dimensional visualization. For spider graphs, we can imagine a plot in which a party has values of 2 on all four axes of the plot. The result is square with an area of $2 \times 2 = 4$. If we then double the values of the scale to 4, the resulting area is $4 \times 4 = 16$. This is four, instead of two times as large. Fourth, users might assume there is a relation between axes that are next to each other. But, in the example above, A and E are as much related to each other as A and C. Fifth, as the axes are not on a straight line, the user has to mentally rotate them to compare them. This not only requires higher attention but is also more prone to error than when all axes would be parallel to each other. Sixth, there is the problem of occlusion. This means that one feature of the graph hides or obscures another. This often happens in spider graphs as many designers do not only show a simple line but fill up the complete area. The result is something like this:

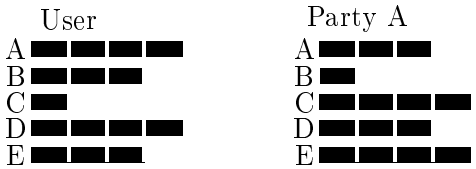


This gives us not two, but three areas. One area covered by the user, one area covered by the party, and an area covered by both the user and the party. This latter area does not signify anything but is an artefact of the visualization. Also, while the area does show what both the user and the party have in common, it is difficult to compare this area to the other areas. Moreover, the darker colours resulting from the overlapping can make the grid-lines and axis labels harder to make out, resulting in further complications. Seventh, when we compare areas, we need to realize the areas we are looking at are dependent on the way in which we position the axes:

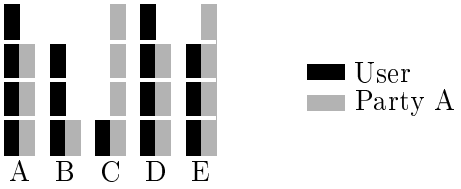


Here, the areas of both polygons are different, while their positions on the dimensions are the same. This is because the area of a polygon is at its greatest when all its sides are equal in length (Feldman 2013). But this depends on how we position the axes. The area covered by a polygon in a spider graph is thus not only dependent on the data, but also on how we order the axes.

As a result of these problems, two-dimensional visualizations perform better than spider graphs (Burch et al. 2008). They are also harder to memorize and take longer to understand (Diehl, Beck, and Burch 2010). Also, Albo et al. (2016) found that users used a tool-tip option provided in their study (which showed the actual value on a given dimension) more often for spider graphs than for the other graphs in their study. This suggests that users found it difficult to estimate the values in the spider graph. Moreover, we can replace spider graphs with a simpler solution. We can do this using the idea of *small multiples* (Tufte 1990, 2001). These small multiples are a series of small plots that users can compare. Thus, we can plot the information from the spider graph also as two bar plots:



This makes it easy to compare the user with the party using the height of the bars (with white lines replacing the grid-lines). It also solves the problems with areas and angles, and we can include other parties without any problem. We can make the visual even more clear when we group the user and the party together:



This visualization allows us to see in one glance the dimension on which the user and party agree or disagree. Besides, we can include more parties using different colours or shadings. This is possible, as we are comparing *within* and not *between* the different dimensions.

Thus, while spider graphs are often used in VAAs when designers want to show more than 2 dimensions, they are not without problems. As they are open to misinterpretation in many ways, they may lead to users making incorrect conclusions, thus undermining the purpose of the VAA. Moreover, if needed, we can replace them with bar plots, which are easier to comprehend and less confusing.

While bar graphs and spider graphs can show us certain aspects of the data, what they cannot do is show the relationships between certain points of the data. Notions of "further away" or "on the same position" have no real meaning in bar graphs or spider graphs. This, while we use the political space to talk about relationships. We can do so using two-dimensional graphs. Figure 7.2 shows an example of what a low-dimensional graph might look like. Here, two dimensions define the two different axes (left-right and progressive-conservative). The points A, B, and C define the positions of actors on these dimensions. This set-up allows us to say several things about these actors we cannot do with bar graphs and spider graphs. For example, we can say actor A is the most left-wing of all the actors and is more progressive than conservative. Actors B and C are both positioned on the right-wing side, with actor B being more progressive than C and A.

Another thing we can talk about is the distances between the actors. For example, the distance between actors A and C is 8 on the left-right axis and 3 on the progressive-conservative axis. Thus, we can say that the total distance between the two actors is: $8+3=11$. This is what we call the *city-block* or *Manhattan* distance. This type of distance looks at the distance between two actors on each of the axes and adds this up for the result. Another approach is to calculate the smallest distance between the points. In that case we would use the Pythagorean theorem to calculate the *Euclidean* distance: $\sqrt{3^2+8^2}=8.54$. This is shorter than the *city-block* distance of 11. Which one we choose depends on whether the dimensions relate to each other or not. A case in which this is relevant is when we want to

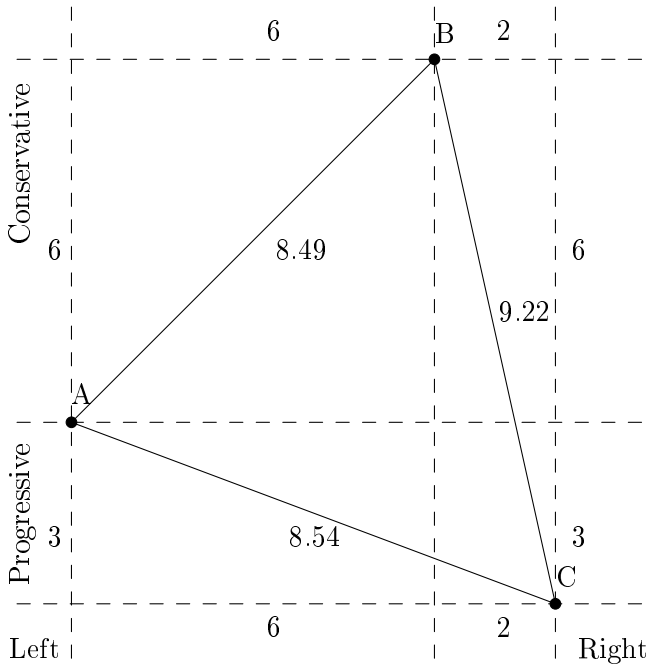


Figure 7.2: Example of a political space, adapted from Benoit and Laver (2006)

know whether B or C is closest to A. If we would use city-block distance this would be C, because the distance between A and C ($8 + 3 = 11$) is shorter than the distance between A and B ($6 + 6 = 12$). If we would use the Euclidean distance the closest actor would be B as $8.49 < 8.54$. Closeness is thus a matter of perspective.

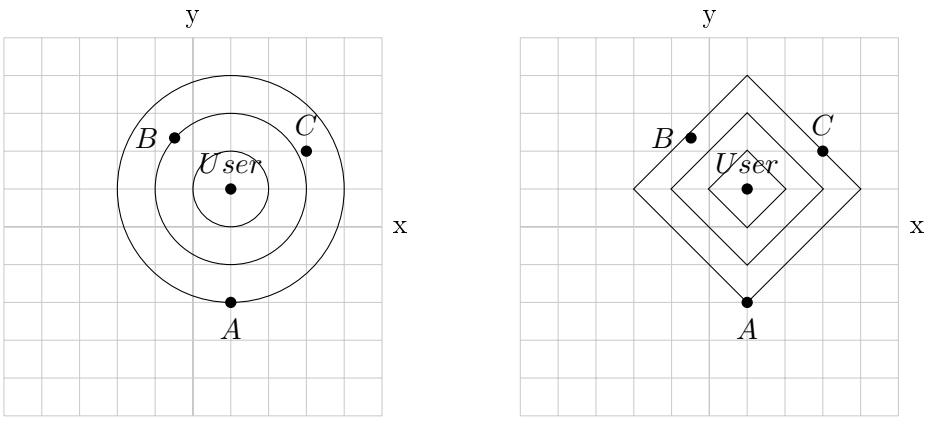


Figure 7.3: Equidistances of 1,2 and 3 under a city block and Euclidean logic. Adapted from Gärdenfors (2000, pp.15-21).

Figure 7.3 shows what happens when we view distances either a city-block or Euclidean. Here, all points on either the circles or the squares are either 1,2 or 3 units away from the user. Only in four cases, on the angles of the squares, do both squares and circles intersect and are the distances between both methods the same. In all other circumstances, the distances of the city-block method are closer to the user. In other words, each of the squares is always inside the circle of a corresponding distance. The result is that for a user using the city-block logic, party A is as far as party C (3 units), while for a user using Euclidean logic, C is closer

than A. Also, point B is close by for a user who uses Euclidean distances, but far for a user who uses city-block distances.

Whether we use Euclidean or city-block logic depends on how we view the dimensions. If we view the dimensions as unrelated we say that the position on the left-right axis is independent of the position on the progressive-conservative axis. In that case, it would be logical to use the city-block distance, as each dimension carries the same importance. If we view the dimensions as related we say that positions on the left-right dimension go together with positions on the conservative-progressive dimension. In this case, the Euclidean distance is more useful (Humphreys and Laver 2010). Which type of distance VAA users use is still unclear. Benoit and Laver (2006) report that while most scholars use Euclidean distances, the psychological literature suggests that city-block more often used. Whether users use a city-block or Euclidean logic, may also depend on the visualization itself. For example, in the *Kieskompas* VAA developed for the 2017 elections in the Netherlands there were circles around the location of the user. The VAA thus seems to assume and strengthen in the user, the Euclidean view.

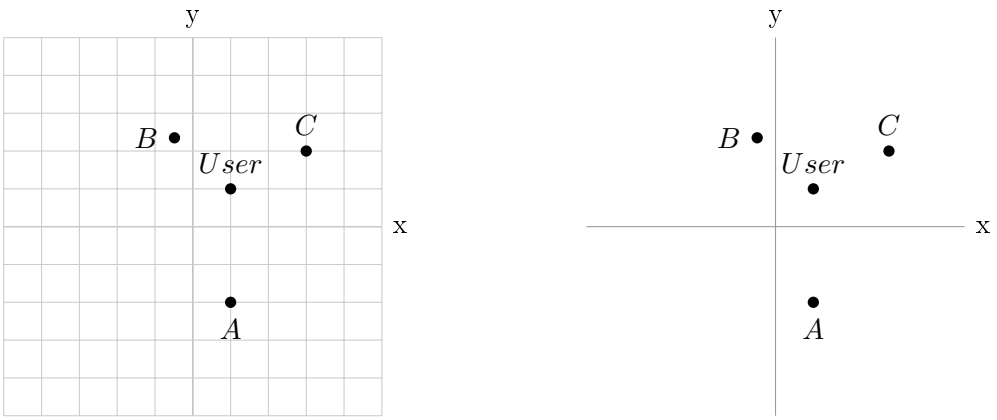


Figure 7.4: Plots with and without a grid. The grid itself makes it easier for users to add the horizontal and vertical distances, possibly making it more likely that a City-Block logic is used.

Also, the presence of a grid can be of influence. Figure 7.4 shows two times the same figure, once with and once without an underlying grid. In the right-hand figure, the lack of the grid makes it harder for the user use city-block logic. This is because they would have to visualize straight lines from the point of the user to mentally calculate the distance. In the left-hand figure, they can use the grid lines to make the calculations easier. Here, it also matters if the points in the figure are on the nodes of the grid or between them. As such, the city-block distance from the User to point C is easier to calculate than the same distance between the User and point B.

Another complication that can arise is that of importance. Figure 7.5 shows what would happen if we make the progressive-conservative dimension only half as important as the left-right dimension. As this makes the progressive-conservative dimension smaller, differences are thus less noticeable. Moreover, the differences between the city block distance and the Euclidean distances are now reversed. Here, A and C are farthest from each other using city-block distance ($8 + 1.5 = 9.5$), but closest using Euclidean distance (6.18). A and B

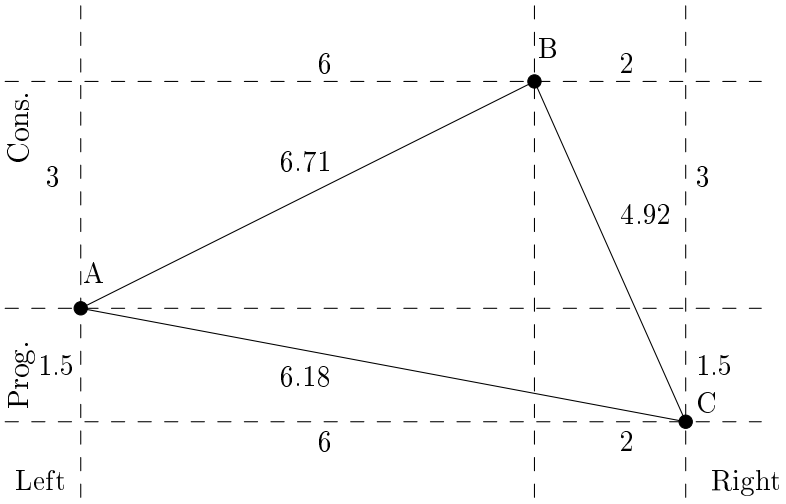


Figure 7.5: Example of Political Space with different importance, adapted from Benoit and Laver (2006)

are the closest on the city-block distance ($6 + 3 = 9$) and farthest on the Euclidean distance (6.71). Thus, importance can change our perception of which actors are the furthest away.

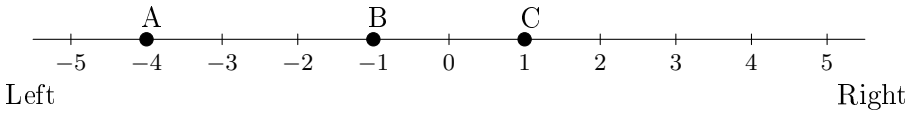
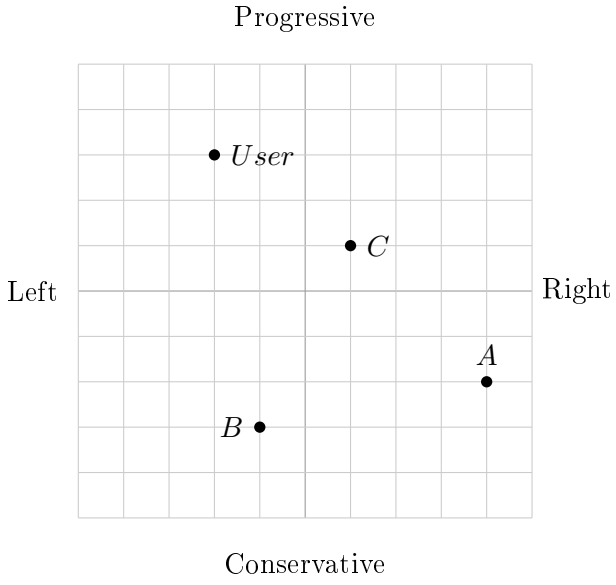


Figure 7.6: The Left-Right Dimension on an absolute scale

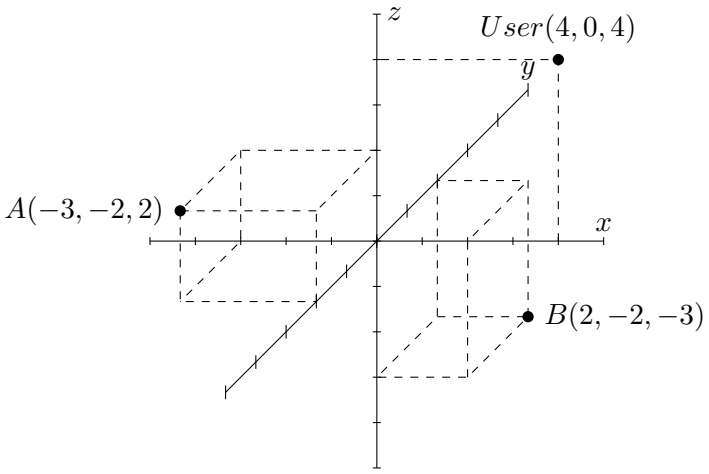
Another complication can arise when looking at directions. We can do this using either a proximity model or a directional model (Mendez 2012, 2014a,b). Figure 7.6 shows the actors A, B, and C on a left-right scale with A at -4, B at -1 and C at 1. If we would ask ourselves which of the actors A or C is closest to B, there are two possible answers we can give. Either we say C is closest to B as the distance of 2 is shorter than the distance of 3 between B and A. This would be using a proximity logic. Or we say that B and A are closer together because they are *on the same side* of the issue. This would be using directional logic. This situation becomes even more complicated when there is more than a single dimension on which the actors can differ.

Before we move on, we have to consider zero points. These points show the middle of the dimensions and allow us to calculate absolute distances in the graph:

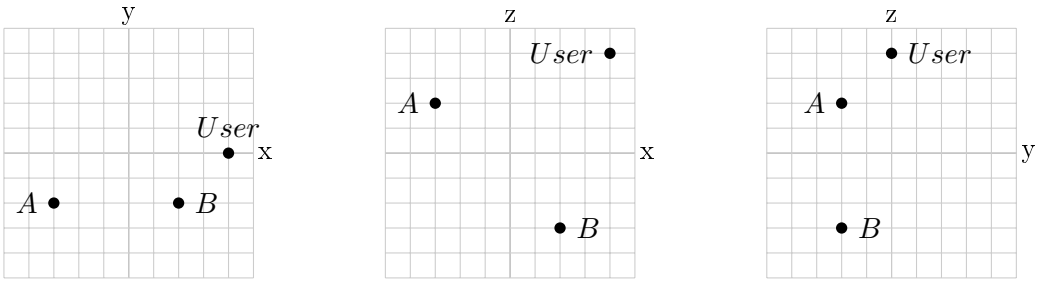


This figure, also known as a *political compass*, not only allows us to make relative statements but absolute ones as well. If a party would be at the edge of the right side of the compass, we could say the party was as right-wing as the dimension allows for. We can also say C has a central position: it is in the middle of both the left-right and the conservative-progressive scale. Yet, note that which party is closest to the user still depends on the way one looks at the visual. Using Euclidean logic, party C is closest - using city-block logic, both party C and B are closest. Also, using directional logic, we can say that while party B is closer because it is on the same side of the left-right dimension, we can say the same for C on the conservative-progressive dimension. In other words, the advice from a political compass is less absolute and more subtle than the advice from the high-dimensional bar-plots.

Sometimes, two dimensions may not be enough to show the political space we want to visualize. This was the case with the *euandi* VAA designed for the 2014 elections for the European Parliament. To show its three dimensions — European integration, economic left-right and conservative-progressive — it required a 3-dimensional visualization. While they may appear appealing, most of the visualization literature argues against them (Robbins 2005; Tufte 2001; Zacks et al. 1998). Both John et al. (2001) and Munzner (2015) found 2D views to be superior to 3D views with regard to understanding the relative positions of two objects. Cockburn and McKenzie (2004) and Cockburn (2004) meanwhile, find that 3D visualizations do not enhance spatial memory and may even make reading the graph more difficult due to the different focal depths of the items. Also, as soon as there are many parties or users, occlusion becomes inevitable. Still, Levy et al. (1996) find users themselves prefer three-dimensional graphs on certain occasions. Even so, they perform worse on these visualizations then on their two-dimensional equivalents. So, Munzner (2015, p.129) notes that there is a dissociation between the preference of users for three-dimensional graphs and their actual task performance. An example of a 3-dimensional visual would be the following, with coordinates and visual aids shown:



It is clear that in this visualization it is hard to see which of the parties are closer to the user and how the parties relate to each other. While the most such 3-dimensional visualizations do allow the user to rotate the visual, this only guarantees that the user spends more time looking at the visual than that they will better understand it. A simpler and more elegant solution would be to translate the 3-dimensional plot into 3 2-dimensional plots:



Here, we can see that while the user is closer to party B on the x dimension, they are closer to party A on the z dimension. Also, separating these plots allows the user to not take into account the dimensions they are uninterested in, reducing the complexity of the task.

Now we have seen the different types of low-dimensional and high-dimensional visualizations, we can wonder whether it matters which one we choose. To give a real-life example, we look at two of the visualizations Stem-Consult employed: a two-dimensional political map and a bar plot.

Figure 7.7 shows both kinds of graph. In the bar graph, the match runs between -100 and 100. This is because Stem-Consult used a hybrid algorithm that includes both directional and proximity logic (Mendez 2017). Here, -100 means that the user completely disagreed and the party completely agreed for all questions or the user completely agreed the party completely disagreed; $+100$ means that the user completely agreed and the party completely agreed for all questions or user completely disagreed and the party also completely disagreed; 0 means that either the party or the user were neutral while the other completely agreed or disagreed (cf. Mendez 2014b). At intervals of 50, the graph shows a grey dot and the strength of the match is further indicated by colour. Green shows a strong match above 50, orange shows an average match between 0 and 50, and red shows a weak match between 0 and -100. Moreover, the graph also shows the actual percentage match. The political map

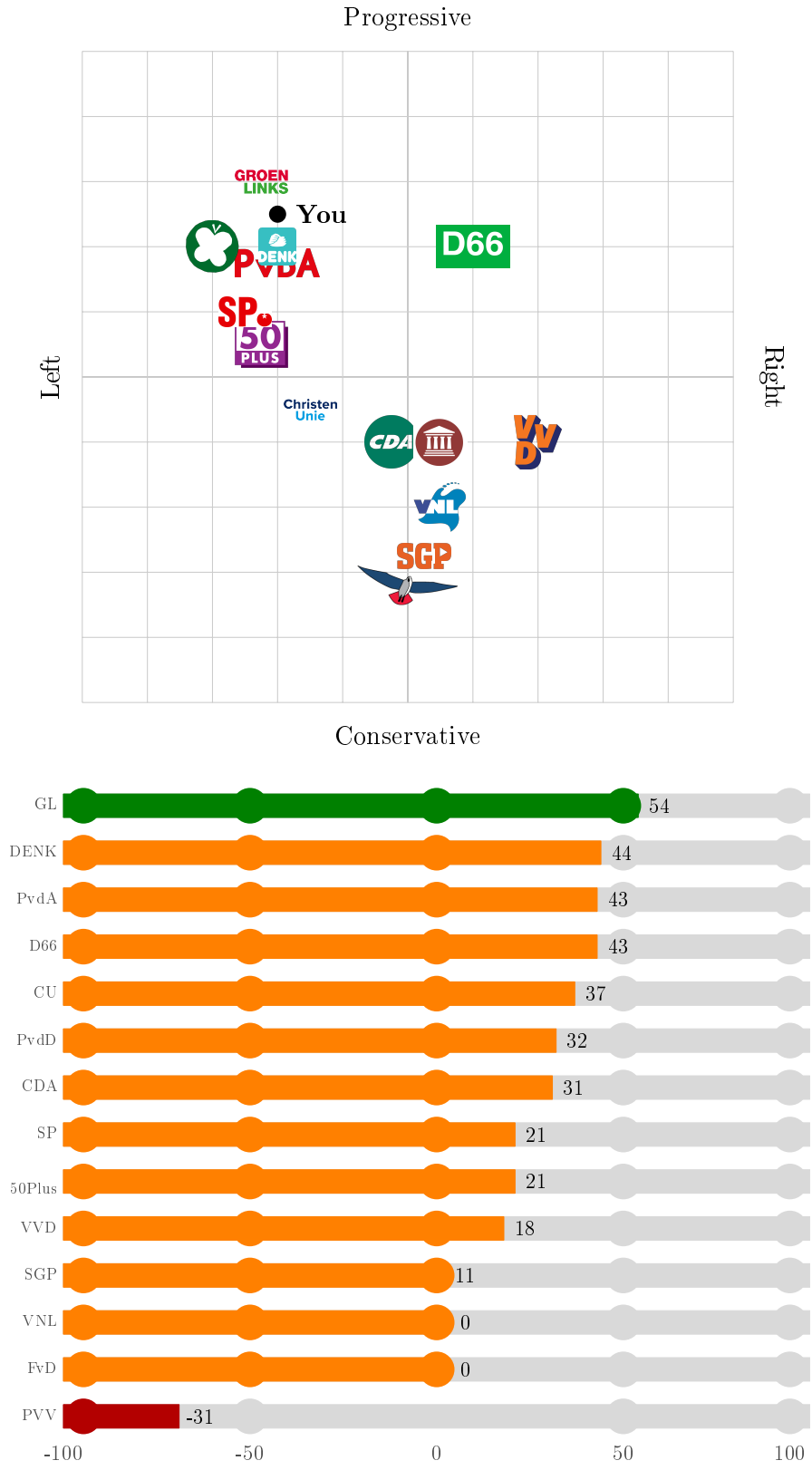


Figure 7.7: Two-dimensional and bar graph visualization in Stem-Consult

indicates the extremes of both dimensions and shows the logos of the parties. Moreover, it has a grid helping the user to calculate distances.

The bar plot shows us three things: there is one party we are the closest to (GL), many parties where the match is about the same, and one party we are farthest away from (PVV). The political map shows us many of the parties are very close together on both dimensions. We can identify a cluster in the left-progressive corner of the map, a cluster in the right-conservative corner of the map, and a single party (D66) in the right-progressive corner of the map. Also, we find ourselves located between two parties: DENK and GL, which are at equal distance from us. Also, we can identify our position as left-progressive, and more progressive than left.

This means that while in the bar-plot the implicit advice is GL, in the political map both GL and DENK are possible. Other interesting points is that the graph positions 50Plus, which has a 21% agreement with the user closer than the CU, which has a 37% agreement. Also, the graph locates D66 far away from the user in the political map but stands on place 4 in the bar graph. This is interesting as it shows that even though the matching algorithm does take directionality into account, the direction is still better visible on the map than in the bar graph. Note also that PvdA and D66 have identical matches (43%). Yet, the PvdA is closer to us than D66. We can say a similar thing can for FvD (symbolized by the small Greek temple) and VNL. Both have a match of 0%, though the FvD is closer to us than VNL. Moreover, while VNL and CDA are close together on the map, in the bar graph they are 5 places apart.

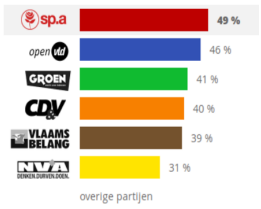
Which results we show thus matters. Using the bar graph, we would have chosen for GL, while the political map gives us more options. They also show us a different picture of how far and close to the other parties are. So, what party is closest to us depends on which visualization we use.

7.4 Examples of Visualizations

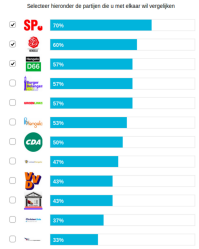
I now turn to a discussion of the visualizations that VAAs have used until now. I divide them into four groups: bar graphs, spider graphs, two- and three-dimensional maps, and other types of visualizations. For each category, I will show examples and discuss how each example varies on the main theme.

7.4.1 Bar Graphs

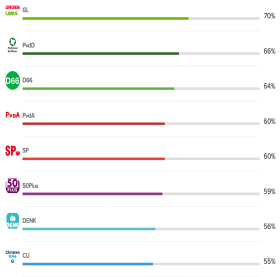
Figure 7.8 shows 9 different kinds of bar graphs. Between them, they differ in whether they are: a) horizontal or vertical, b) only positive or also negative, c) order, d) use different colours for each party, e) show the party logo or the party name, and f) show the percentages. Figure 7.8a is typical of most bar graphs as it orders the parties from highest to lowest. Also, it has different colours for each party, shows the party logos, and the gives the percentages. This makes it the same as Figure 7.8b and 7.8c, with the exception that Figure 7.8b does not use colour. Figures 7.8d, 7.8e and 7.8f differ from them as they do not only show positive but also negative agreement. This is the result of a different matching method. Also, Figure 7.8f uses colour for the bars. Yet, while in the previous figures colour indicated the party, here



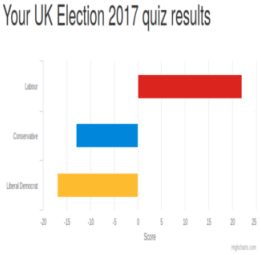
(a) Educatieve Stemtest (Belgium)



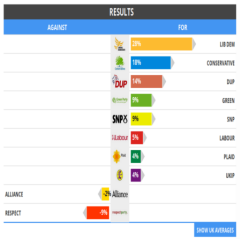
(b) Stemwijzer (Netherlands)



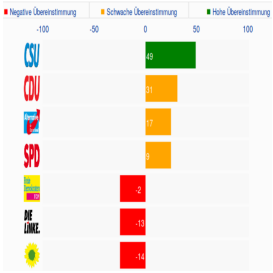
(c) Kieskompas (Netherlands)



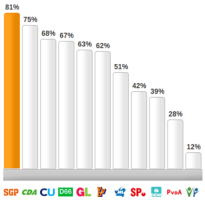
(d) Who Should You Vote For? (United Kingdom)



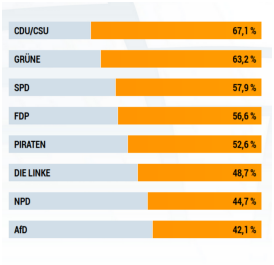
(e) Who Shall I Vote For? (United Kingdom)



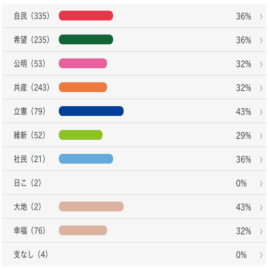
(f) Parteienavi (Germany)



(g) Kiesbalans (Netherlands)



(h) Wahl-O-Mat (Germany)



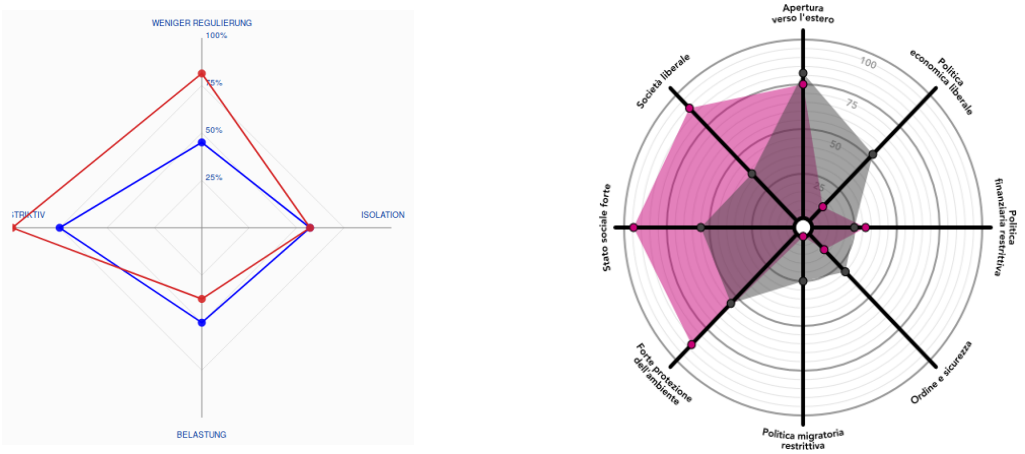
(i) Mainichi (Japan)

Figure 7.8: Variations of the Bar Plot

it indicates the strength of the match. Figure 7.8g, meanwhile, is the same as Figure 7.8b but uses vertical instead of horizontal bars. Figure 7.8h shows the match ordered, horizontal and with percentages but does so from right-to-left. This makes it confusing to interpret as one can think that the grey bars and not the orange ones show the match. Figure 7.8i, does show the colours for the parties but is unordered. When we compare these graphs to the recommendations in Table 7.1, we find few of them minimize the data-ink ratio or the info in the graph. For example, Figure 7.8h shows unnecessary grey bars. This is not only unnecessary but also confusing. Besides, one can wonder if including the exact percentage is necessary. Bar graphs are most useful when used to compare lengths. Adding the exact percentage should thus not be a necessity. Also, Figures 7.8g and 7.8e use unnecessary effect. Figure 7.8e adds a little triangle to the end of the bar, making it seems longer than it is, while Figure 7.8g uses a three-dimensional effect that adds no extra information.

About scales, only Figures 7.8d and 7.8f show any kind of scale. In both cases, there is no explanation of what the scale indicates. This is especially a problem with Figure 7.8d as the scale is uneven. This raises the question if there is an endpoint to the scale and what it is. Also, while the scale carries the label *Score*, what this means, or how we should interpret it is unclear. A similar problem occurs in Figure 7.8e, where again it is unclear what the scores mean. This is especially problematic with the negative score. Does this mean there is disagreement? Or opposite agreement? And if so, what does a score of 0 mean then? While there might be an explanation in the FAQ on the website, this does reduce the usefulness of the graph. Figure 7.8f addresses this problem by using a simple colour scheme to show that anything to the left of 0 indicates negative agreement. Other VAAs use colours to refer to parties. This can be practical when parties often use these colours. Yet, it becomes a problem when two colours are much alike, like the PvdA and SP in Figure 7.8c.

7.4.2 Spider Graphs



(a) Parteienavi (Germany)

(b) Smartvote (Switzerland)

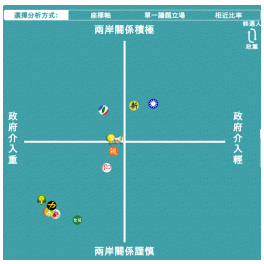
Figure 7.9: Variations of the Spider Graph Visualisations

Figure 7.9 shows two examples of spider graphs. Spider graphs are less used in VAAs, leading to fewer examples. The graph used by *Parteienavi* on the left in is simple, with only four axes and the profile of the user in blue and the party in red. We can choose different parties using a drop-down many (not shown in the Figure). The percentages are only on the vertical axis but are well visible. There are grey lines connecting the different axes, though these are not well visible against the background. The *Smartvote* spider graph is more complex, with 8 different axes. Also, the main axes are thicker than the connecting lines, and the resulting shapes look more like spider webs. Both the area of the party and the user are in colour and show the overlapping colour mentioned earlier.

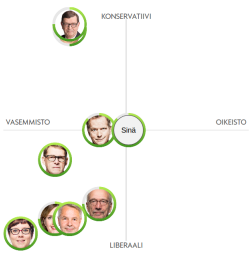
Especially for *Parteienavi*, it is doubtful whether a spider graph is the best way to visualize the result. With only four different axes, four bar plots or four two-dimensional maps could show the same information. Not only would this address the problems that come with spider graphs, but they would also show all parties at once. Also, while *Parteienavi* uses grid-lines, there is no grid-line for the 100% mark. For the *Smartvote* graph, the opposite seems to be the problem, with an unnecessary multitude of grid-lines at every 5%. This makes the

graph looks fuller than necessary. Also, both do not include any information on how the user should read the graph. As spider graphs are less obvious than the other graphs, this might be a welcome addition.

7.4.3 Two and Three Dimensional Visualisations

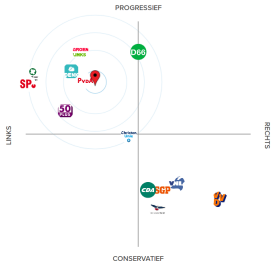


(a) iVoter (Taiwan)



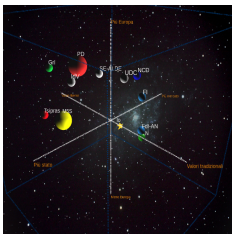
(b) Vaalikone (Finland)

(c) Parteienavi (Germany)



(d) Kieskompas (Netherlands)

(e) Vote Compass (United States)



(f) euandi (European Union)

Figure 7.10: Variations of the 2 and 3-Dimensional Political Landscapes

Figure 7.10 shows six variations of the political map. Five are two-dimensional, and one is the (so far) only visualization that uses a three-dimensional map. This map belongs to the *euandi* VAA launched for the European Parliament elections of 2014. It is interactive and allows for 360° rotation in each of the three dimensions. Also, users can zoom in and out. To show the parties, it uses spheres. The size of these spheres is proportional to the share of the votes the party obtained in the previous national election. This is unclear from the visual, and the three dimensions make bigger parties seem even bigger than they are. Also, it might look like some parties are close together, while actually, they are not. The only reason they do look together is that the large spheres touch each other. The *iVoter* shows the logos of the parties without any grid lines and uses a green background. The Finnish *Vaalikone* (Figure 7.10b) does the same but uses pictures of the candidates instead. *Kieskompas* (Figure 7.10d) uses party logos and shows concentric rings around the user. As we saw, this indicates the designers suppose a comparison in the Euclidean sense makes the most sense. The *Parteienavi* omits this but shows a raster instead. This makes it easier to calculate a city-block distance. *Vote Compass* in Figure 7.10e does the same, but shows the names of the candidates instead of their picture. It also uses colour to show which party the

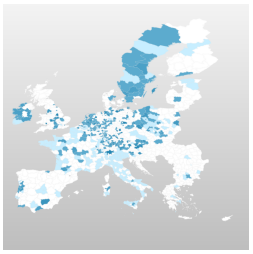
candidate belongs to.

On the data-ink criterion, most graphs perform better than the bar-plots, except for the Taiwanese iVoter, which uses a turquoise background, and the *Vaalikone*, which uses green circles around the portraits. While both *Vote Compass* and *Parteienavi* use grid lines they are more successful in the latter than in the former. As Tufte (1990, p.112) notes: "[a grid should be] muted or completely suppressed so that its presence is only implicit". This is to prevent it from competing with the data for the attention of the user. In *Vote Compass*, the grid is visible, while for its actual use a more greyed-out version would have worked as well. Yet, when we compare *Parteienavi* to *Kieskompas* we find none of the information is missing as we can still see the relative distances. Also, *Vote Compass* uses a confusing labelling system. Instead of labelling the end of each of the axes, the labels are on top of the graph. This makes it more difficult to link them to their respective axes.

Returning to the *euandi* graph, there are we find several more problems. First, is that reading the graph is difficult. It requires several rotations and zooms to get a good idea of the space. Also, not only is the size of the planets obscuring information, but it is also unnecessary information. The size of the party in parliament hardly matters for our match. Another unnecessary feature are the stars in the background. While they are most likely there to suggest "space", they do not contribute anything. Also, the indicator for where the user stands (a star) is small and often disappears behind the party spheres. Finally, the colours for the planets seem random - sometimes they correspond to party colours, other times they do not. All this makes the graph confusing and replacing it with three two-dimensional graphs would make it easier to understand.

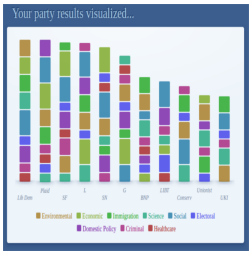
7.4.4 Other Types of Visualisations

Figure 7.11 shows several examples of alternative visualizations VAAs use. Figure 7.11a comes from *euandi* and shows a map of Europe showing how much the results of the user compare with other users. The darker an area, the higher the similarity. Figure 7.11b from *isidewith* addresses the question of how to deal with many low-dimensional dimensions without using a spider graph. Besides the visual, the VAA also gives the user a description of how to interpret their result. The dimensions themselves include some common ones like left-right, but also dimensions like tough-tender and politically correct - politically incorrect. From the same VAA, Figure 7.11c shows a bar graph with the questions grouped together. The result is a confusing overview of how often a user agreed with that party on these groups. Figures 7.11e and 7.11d show the visualizations that *PositionDial* used. Figure 7.11d shows a figure which has different attributes based on the answers the user provides. In this case, the figure wears a helmet with a peace-symbol to show the anti-war stance, wears a shield labelled "tradition", holds chains to showcase the regulatory stance, and stands above a fence to show an immigration-restricting position. This figure starts out as neutral, but gains attributes over the course of the VAA. The result, while interesting, is hard to interpret, with the description below it the most helpful. The other visual is also updates over the course of the VAA and shows three segmented rings. The middle ring shows the number of questions the user is in the middle on, the first ring shows the statements where they have a position on, and the outer ring shows the number of statements where they have a strong position on.



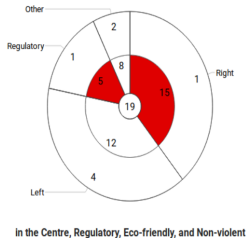
(a) euandi
(European Union)

(b) isidewith
(United Kingdom)

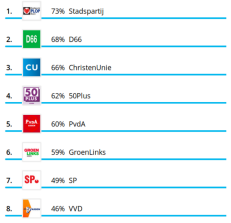


(c) isidewith
(United Kingdom)

(d) PositionDial
(United Kingdom)

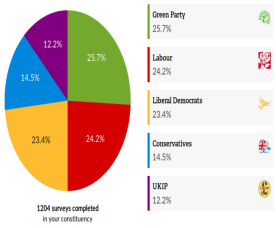


(e) PositionDial
(United Kingdom)

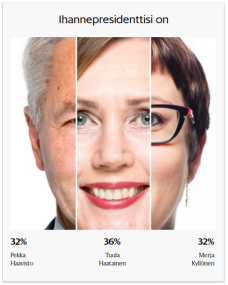


(f) MijnStem
(Netherlands)

(g) Stemwijzer
(Netherlands)



(h) Vote for Policies
(United Kingdom)



(i) Vaalikone
(Finland)

Figure 7.11: Visualisations - Other

Here, while both the *left* and *right* categories have the same number of questions, the user answered only a single statement using an extreme response option (e.g. "completely agree") for the *right* category, while this has happened 4 times for the *left* category. The red shading shows which of the categories contained the "stronger" positions. *Stemwijzer* places the graph from Figure 7.11g on the same screen as the bar graph discussed earlier. It not only tells the user which party is closest but also shows the three highest matches in a circular bar graph with a written percentage. *Vote for Policies* in Figure 7.11h lists the parties in order with a percentage and presents the same information in a pie chart. The percentages here refer to how often the policies the user chose where the same as those of the parties. Thus, the total percentage can be 100. Figure 7.11i from *Vaalikone* VAA shows a face built up of the faces of the candidates. The size of their faces depends on the level of agreement with that candidate.

While all the visuals look interesting, their main problem is that it is often not clear how they work. The most problematic are the coloured bars of *isidewith* and both visualizations of *PositionDial*. Users are unlikely to often see these graphs, which undermines their purpose. Also, it is unclear what the graph exactly wants to show. A similar problem occurs for the *Vote for Policies* visual. Not only is the pie-chart unnecessary and is it violating the data-ink principle, from the visualization it is unclear why the percentages should sum to 100%. This is only explained elsewhere on the website. Another graph violating the data-ink principle is *Helsingin Sanomat* which looks appealing but is hard to read. Replacing it with a bar plot would have made interpreting it easier. With regard to unnecessary data, Figure 7.11b contains interpretations of the bars that are not needed. Not only does it force an interpretation on the user, but it also lacks a good description of what the dimension is about. Besides, *Stemwijzer* points out the party with the highest agreement, while this is already clear from the bar graph (that is below it). Also, by mentioning the party another time in a text, this enforces the idea that it gives real advice.

On colour, the *euandi* map is a good example of colour as it shows darker colours for a higher degree of similarity. *isidewith* also uses colours but these seem to be random and do not add any information. Finally, the *MijnStem* graph uses blue bars that do not have any function, a numbering that is already clear from the order and names of the parties that are also in the logos. Removing these and replacing it with a bar graph would improve the visual.

7.5 Reflection

If we follow the literature, we can conclude that the simpler the graph is, the better. For VAAs, this seems especially relevant as their users are very different from each other and do not have similar backgrounds. Thus, bar graphs and two-dimensional visualizations are the clearest to use. This is most likely also the reason VAAs so often use these. Yet, all suffer from some problems, with lacking a clear message and overuse of non-data ink the most frequent. The reason for this is VAA visualizations have rarely been the topic of discussion in the VAA community[6]. This is a worry as the visualization is the reason why most users fill out a VAA. While the questions might already help them orient, the visual is their real goal. As we design VAAs to aid users who would have trouble to position themselves in the elections, visualizations should be as simple and clear as possible. There should be no need for explanation or possibilities for misunderstanding. Yet, many visualizations seem to focus more on visual attractiveness than usefulness. This is at odds with the idea that visualizations in VAAs should be "valid and instructive" and that VAAs should have a "simple and understandable" design (Garzia and Marschall 2014, p.228).

To address this problem, the VAA research field would do well to pay closer attention to the visual aspects of their VAAs. Instead of being a separate choice, the visual should be part of the main design process. To that effect, a wider debate on the visuals in VAAs and the way they can influence users would be welcome. To begin with, designers can use

[6] The study by Alvarez, Levin, Trechsel, et al. (2014) is the only example of this. Here, the authors looked at the usefulness of the visualizations. For this, they used the rating users gave to different aspects of the visual. They found that the visualizations they looked at — bar graphs, spider graphs, and political compass — were all useful to the user in the same degree (Alvarez, Levin, Trechsel, et al. 2014, p.98).

the experiences from the field of information visualization to improve their graphs. Also, the theory of graphical perception can help to improve their designs. This will not only improve the quality of the data the VAA generates but also help the user.

8 | Visualization in Practice

In the previous chapter, I showed how designers use different visualizations to show the match between the user and the parties in the VAA. We saw the choice for a visualization depends on how the designers understand the political space. We also saw that the way in which users react to these visualizations depends on the process of graphical comprehension. Yet, we know little about this process with regard to VAAs. Most designers seem to suppose users take the visualization "as-is". But we saw in the previous chapter, that this seems unlikely. Thus, in this chapter, I will look closer at this process of comprehension. To do so, I will run a brief online experiment in which I show the respondents several types of visualization and answer them some questions. This analysis aims to address each of the five influences on the process of graphical comprehension: domain knowledge, graphical knowledge, data complexity, task demands, and graph characteristics.

Domain and Graphical Knowledge. For VAAs, domain knowledge is equal to political knowledge. This means, for example, the knowledge to know what a political map is. Or the knowledge to know how to interpret "left" and "right". Graphical knowledge is, as we saw in the previous chapter, the knowledge to know how to read the visualization. Thus, knowing to compare the distances between the points on the political map. Or to read the values of the axis in a bar graph. Users with a high domain and graphical knowledge will thus more often interpret the visualizations in the "correct" way. Correct here means: as intended. Note that this assumes the visualization itself is correct. This assumption is most likely to be incorrect. Yet, we assume it here as we are not so much interested in the correctness of the visualization but more in if the interpretation is in line with our expectations. Our first hypothesis thus is:

H1 A correct interpretation of the visualization is more likely if the user has a high domain and graphical knowledge.

Graph characteristics and Task Demands. The graph characteristics describe what the graph looks like, while task demands describe what the user has to do to understand the graph. Here, I compare three different types of graphs: a bar graph, a closed spider graph and an open spider graph. Closed and open refers to if the area in the spider graph is opaque or not. Each of the graphs presents the same information in a different way. This means the characteristics of each graph are different. Also, it means the task users have to carry out is different. Note that this assumes the visualization itself is correct. There should thus be no difference in the responses of the users to these visualizations unless there is an effect of the graph characteristics. Thus, we expect:

H2 The responses to the bar graph and the open and closed spider graphs are similar.

Data complexity. The information the graph holds defines the data complexity. The more data in the graph, the more complex it is. Higher complexity makes a graph more difficult to read and understand. Here, I give the users two political maps: one with 3 parties and one with 6 parties. As the graph with 3 parties is less complex, I expect that:

H3 The political map with 3 parties has more correct responses than the political map with 6 parties

Besides the five influences, I will look at two other points of interest. First is to see if users look at the political map using either a City-Block or Euclidean logic. I will test this by asking the users to describe their position on a given political map. If their description contains references to both axes, I will view this as a City-Block view. If their description contains references to distances between actors, I will view this as a Euclidean view. Second, I will see what users think about the usability of the visualizations. Usability is a central concept in visualization studies and is "the extent to which a product can be used by specified users to achieve specified goals with *effectiveness*, *efficiency* and *satisfaction* in a specified context of use" (Forsell 2014, p.178, emphasis added). Testing for the usability thus allows us to see if users find the visualization useful to find out their match between them and the parties.

8.1 Measurements

In total, 67 respondents filled out the online survey. I designed and ran the survey using PsyToolkit (Stoet 2010, 2017) and recruited the respondents using Amazon's Mechanical Turk.

To measure graphical knowledge, I use the *subjective graph literacy* (SGL) measure (Garcia-Retamero et al. 2016). This measure replaces the standard *objective graph literacy* measure (OGL). SGL has respondents judge their own literacy whereas in OGL the researcher does this. Despite being subjective, SGL provides a reliable and valid measure of graphical knowledge (Garcia-Retamero et al. 2016). SGL contains five questions, each asking the user their ability on a certain type of graph. These are "How good are you at: a) working with bar charts, b) working with line plots, c) working with pie charts, d) finding out what the size of a bar is in a bar chart, and e) determining the difference between 2 bars in a bar chart". Responses to the questions range from 1 (Not At All Good) to 6 (Extremely Good). I calculate the final SGL score by summing the responses and dividing them by the number of questions. To measure domain knowledge, I ask the respondents to state their political interest using a seven-point slider. The reason I choose political interest and not political knowledge is that Mechanical Turk makes it difficult to select from which country users originate. Thus, it is difficult to construct a measure of political knowledge that I could use for all respondents.

To measure the responses to the bar graph and the open and closed spider graph, I first split the respondents into two groups. One group receives three closed spider graphs, while the other group receives three open spider graphs. Assignment to either of the groups was

random. I then asked the respondents in which of the three visualizations did they consider the shape of the user to be most like the shape of the party. Then, I showed all users three bar graphs which showed the same information as the spider plots and asked the same question. Note that the users were not made aware of the fact that the information in the plots was exactly the same. Figure 8.1 shows the visualizations that the users got to see. In each of the three columns in the figure, the different graphs have the same values on each of the dimensions. The users see either the three graphs of the first or the second row. All see the three graphs from the third row. Both the graphs in the first and third column have the same amount of difference on all dimensions combined (8). The amount of difference in the middle column is the lowest 6.

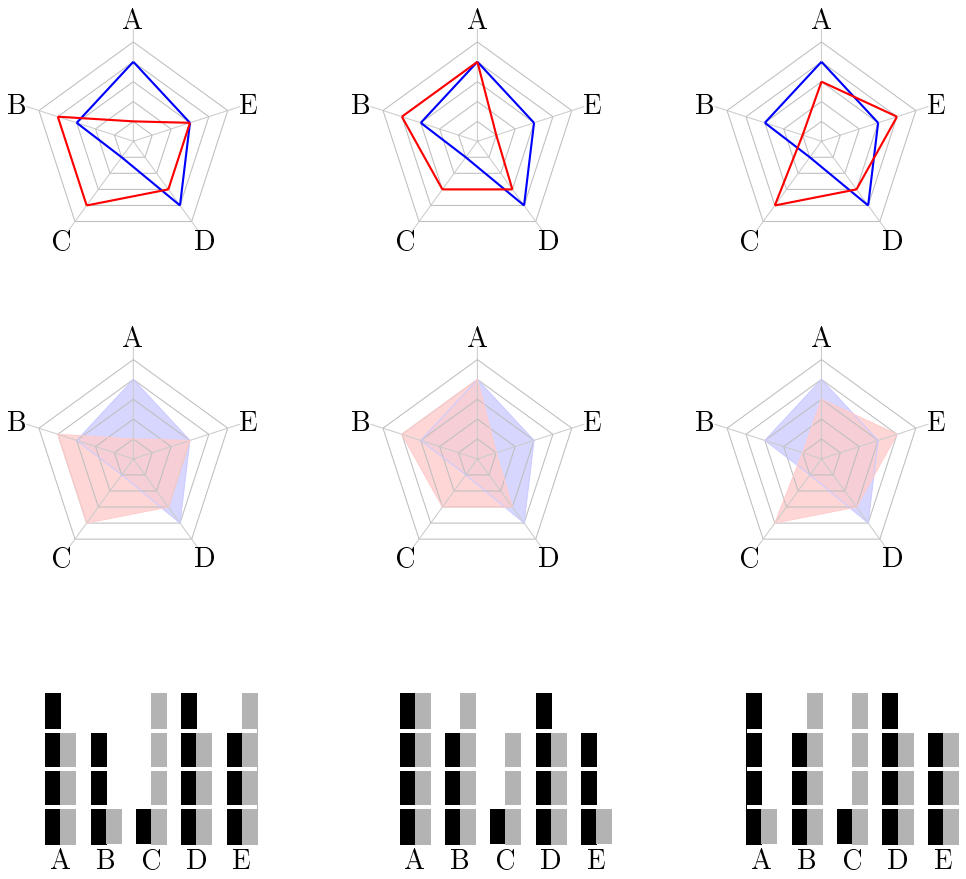


Figure 8.1: Figures for the Spider graph and bar graph experiments.

To measure whether a political map with 3 parties has more correct responses than the political map with 6 parties, I show users the political maps from the upper row of Figure 8.2. I ask them to select the point that they deem to be closest to the point labelled "You". I constructed the political maps in such a way that it is possible to interpret it using either City-Block or Euclidean logic. In the political map in the top-left, point A is the closest with Euclidean distance (4.8), while the point B is closest with City-Block distance (6.0). In the political map in the top-right, this is similar though this map contains three more parties. Also, point A and C are now labelled as points E and F. This allows us to observe whether users will recognize these points are the same. To see whether users use City-Block or Euclidean logic to describe their position I show the political map from the lower row of

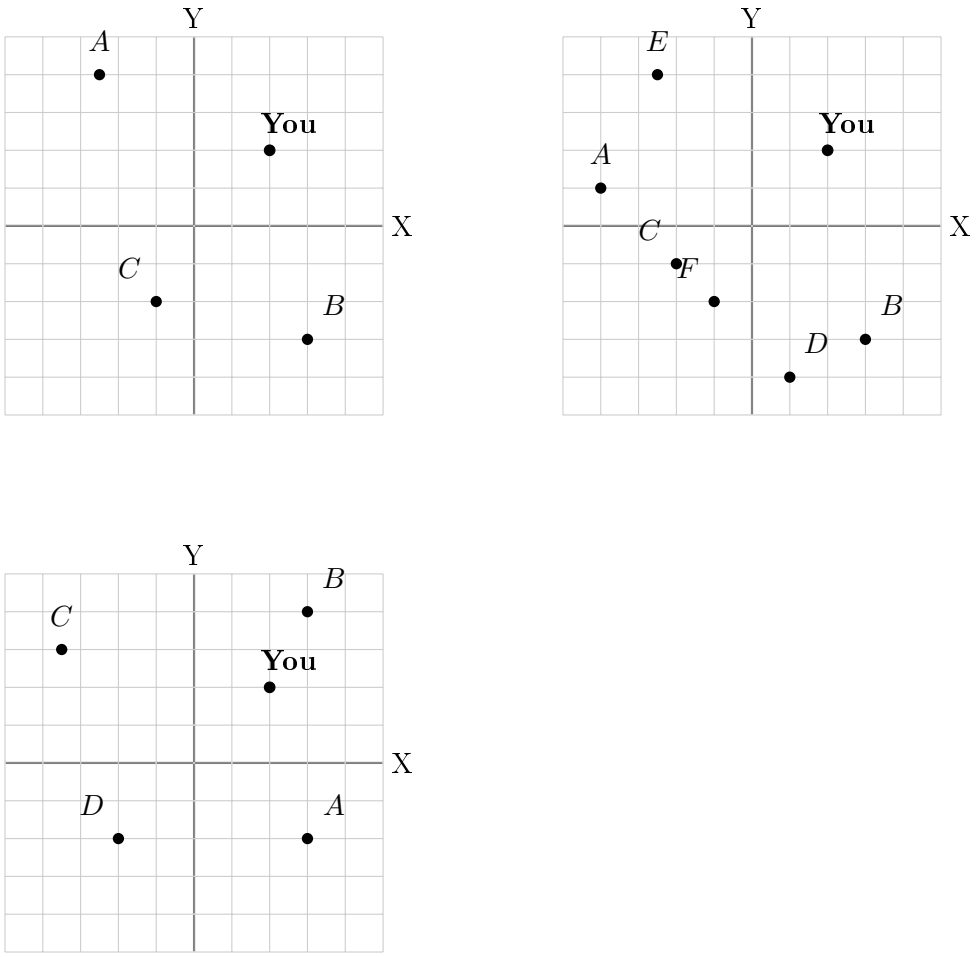


Figure 8.2: Visuals for the Political Maps

Figure 8.2 and ask them to describe their position.

To measure the usability, I use a set of 10 items based on those used in Shamim, Balakrishnan, and Tahir (2015). For each visualization, I ask the users whether:

Q1 The visualization is eye pleasing

Q2 The visualization is easy to understand

Q3 The visualization is user-friendly

Q4 The visualization is informative

Q5 The visualization is intuitive

Q6 The visualization is useful

Q7 The data in the visualization is comprehensive

Q8 The data can easily be compared in the visualization

Q9 The data is well represented in the visualization

Q10 Pre-knowledge is required to understand the visualization

I present these after the questions on the spider graph, bar graph and the three questions on the political map. Responses are on a 5-point Likert scale and the usability score is the average of all the items.

8.2 Results

	Open	Closed	Bar
I	5	6	12
II	19	14	42
III	9	14	13

Table 8.1: Data for the two different types of spider graphs and the Bar plots. Responses are to the question *In which graph is the profile of the political party the most similar to you?*

Before I address the hypotheses, I will first discuss the results. Table 8.1 shows the choices the users made. For the open spider, 19 users choose visualization II, while 14 did so in case of the closed spider graph. Besides, in that case, a similar number of users considered visualization III as being the most similar. Most interesting is, that for the three different visualizations there seems to be no agreement between the users on what the closest is. This is also the case for the bar plot, though in this instance 42 of the 67 users choose the second visualization.

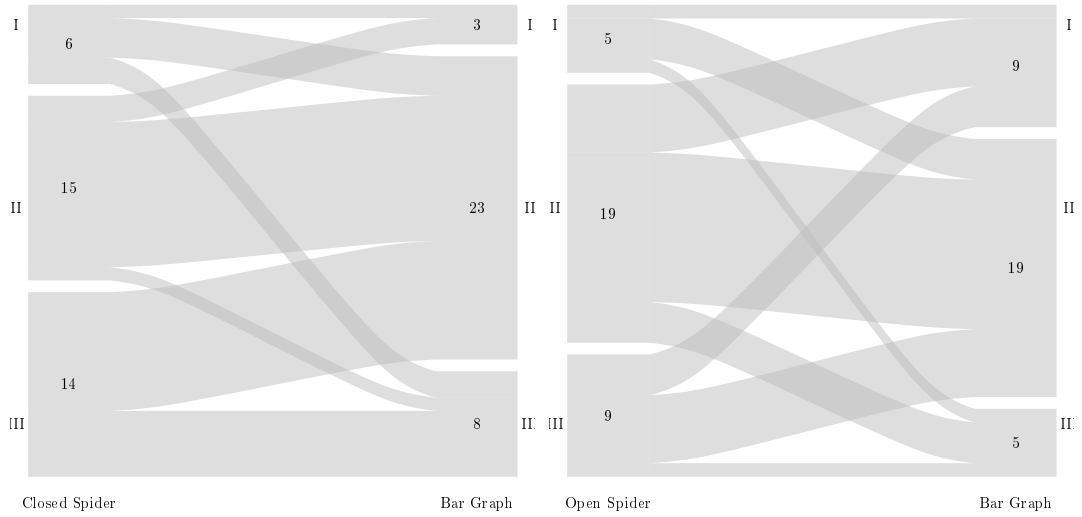


Figure 8.3: Users switching between either version of the spider plot and the bar plots.

Another point of interest is to see whether users noted the spider graphs contained the same information as the bar plots. To see whether this is the case, I look if users switched between either version of the spider plot and the bar plot. From Figure 8.3 we see that for the closed spider graph almost all users who choose visualization II remained with this visualization in the bar graphs. Users who chose visualization III also switched to visualization II though some remained with number III and none switched to number I. This was not the case for the open visuals where the group that chose visualization II switched more often to

visualization number I, while only one of those that chose visualization III remained with their choice. In both cases, this indicates a significant group of users did not recognize that the spider graphs and bar graphs showed the same information.

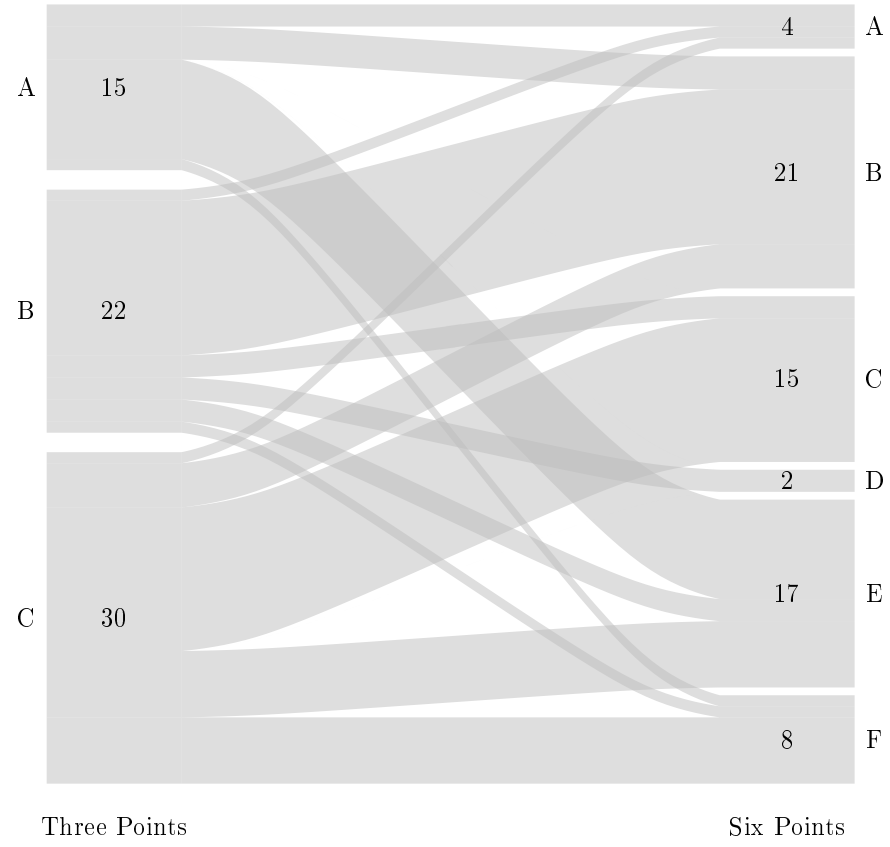
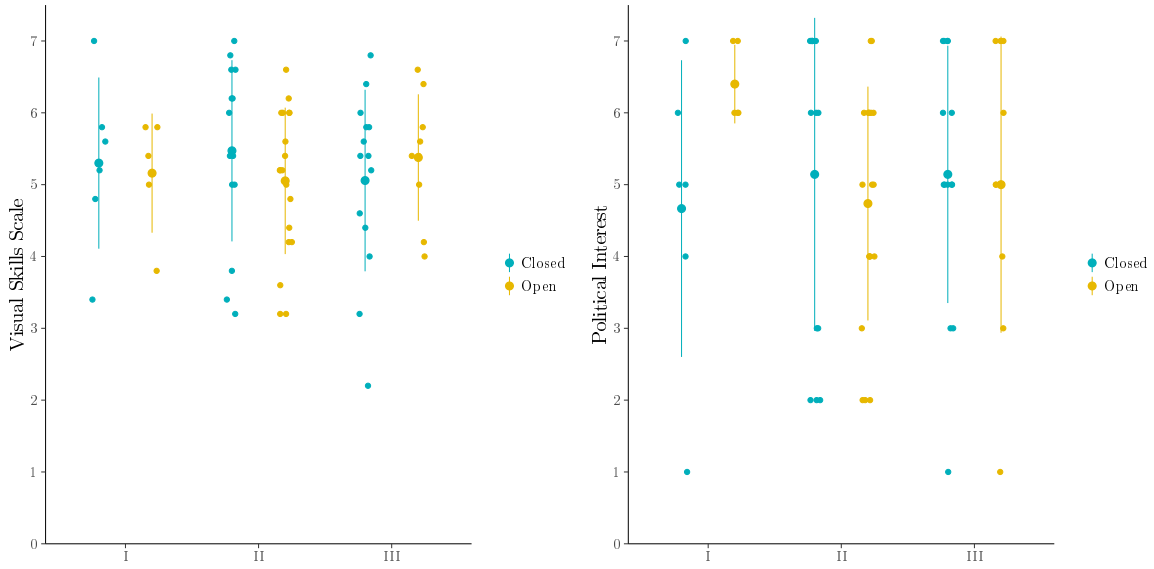


Figure 8.4: Flow plot of users switching between the different labelled points. Note that in the visual for three points, I labelled point E as A, point B as B and point F as C. In both cases, point B was closest in a City-Block distance, while point E was the closest in a Euclidean distance.

For the political map, Figure 8.4 shows the choices the users made and whether they switched their positions. Of interest is that in the graph with the three points, most users opted for the point C as the closest point. Yet, that point is neither the closest in a Euclidean nor in a City-Block fashion. Also, when I added more points in the second visualization, users changed their position, even if the two closest points had not changed. Besides, some remained at their point, in particular, point C, even if I moved the point. Again, as with the bar and spider graphs, the message here is that there seems no agreement among the users on the which point is the closest.

I will now turn to the hypotheses, starting with the graphical and domain knowledge. For the political interest, I find a mean of $\bar{x} = 5.06$ and a standard deviation of $s = 1.82$. For graphical knowledge the mean is $\bar{x} = 5.21$ with a standard deviation of $s = 1.09$. As the political interest is on a scale of seven and graphical knowledge on a scale of six, this means that the users have high levels of both. There is no relation between the level of political interest and the level of graphical knowledge ($\tau_b = 0.17$,$p = 0.06$). This means that one's



(a) Graphical Knowledge (b) Political Interest

Figure 8.5: Relation between the choice in spider graphs and the degree of visual skills of the user.

interest in the subject does not tell anything about how well one can read the visualization.

Starting with the spider graphs, Figure 8.5b shows the relation between the choice of the user and their degree of graphical knowledge and political interest. The different colours show whether the user received the open or closed spider graph. We can see there is no difference between the responses for both versions of the spider graph. Also, there is no relationship between the correct interpretation and the level of graphical knowledge or political interest for both versions of the graph. From the perspective of a designer visualization II would be correct (adding the distances together on each of the dimensions gives 8 for visuals I and III and 6 for II). Yet, the level of graphical knowledge and political interest for these users was not different from the other options. Thus, I conclude that a correct interpretation of the visualization is not more likely if the user has a high domain and graphical knowledge. Also, I find no difference in the choices between the closer and open spider graphs.

Turning to data complexity, I find that most users (30) consider point C the closest in the 3 party version, and point B (21 users) in the 6 party version. Yet, as noted earlier, for the 3 party version, either A or B was correct. For the 6 party version, either E or B was correct. Thus, were in the 3 party version 37 users were correct, in the 6 party version 38 users were correct. This means that increasing the number of points on the map does not lead to less correct responses. The only effect it has is that it generates more different responses.

Another way to look at the data complexity is by considering how long it took the users to respond. This time approach is a simple and effective way to measure how easy to understand the visualization was. Starting again with the bar and spider graph, I find that there is no difference in the scores for the response time for the open version ($\bar{x} = 18.73$, $s = 14.90$) and the closed version of the spider graph ($\bar{x} = 23.43$, $s = 21.41$) ($t(58.99) = -1.04$, $p = 0.30$). Besides, there is no difference between the bar graph ($\bar{x} = 23.32$, $s = 19.91$) and the open spider graph ($t(98) = -1.17$, $p = 0.24$) and the bar graph and the closed spider graph

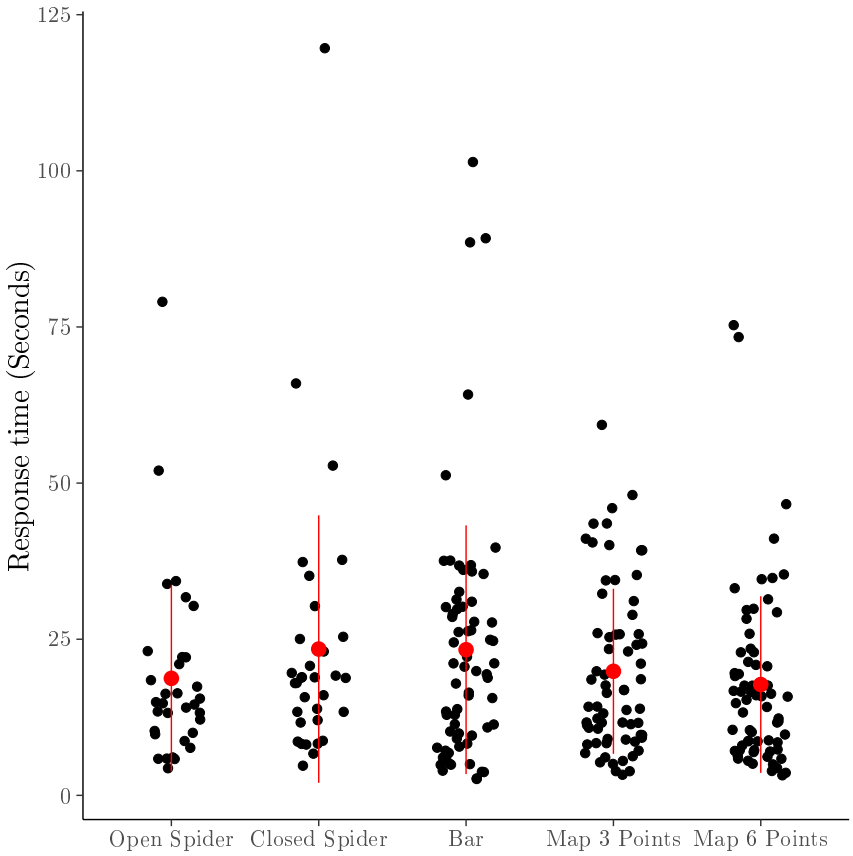


Figure 8.6: Time in seconds spent on answering the question with the spider graphs - for both the open and the closed version - the bar graph and both of the political maps. The red line and point represent the mean and the standard deviation respectively.

($t(99) = 0.02$, $p = 0.98$)[1]. For the political maps, I find there was no significant difference between the mean response time for the graph with the 3 ($\bar{x} = 19.87$, $s = 13.22$) and 6 points (($\bar{x} = 17.75$, $s = 14.14$)) ($t(132) = 0.90$, $p = 0.37$). Figure 8.6 visualizes all the response times. Here, we can see most of the users responded between 0 and 40 seconds, with several outliers in each of the categories. The longest outlier was a single user who needed 119 seconds, or around 2 minutes, to give a reply. Disregarding this user, the response times for the open and closed spider are similar, with the deviation of the closed spider being larger. A similar story goes for the bar graph, where 5 users required more time than the others. Apart from them, the distribution is like the one of the closed spider graph. The same goes for both of the political maps, which show fewer outliers but very similar distributions.

I now turn to how the users interpret the political map. Based on their descriptions[2], I divide them into four categories. I do so with two qualifiers: a) if the user considers their position in the whole graph or in relation to one of the points, and b) whether the user uses relative or absolute terms. Users use relative terms most often, in 48 of the cases, of which 25

[1] I used Welch's unequal variances t-test for the comparison between the open and the closed spider plot as the variances were unequal ($F(32, 33) = 0.48$, $p = 0.04$). I used a regular t-test for the comparisons between the bar graph and the open spider graph ($F(32, 66) = 0.56$, $p = 0.07$) and the closed spider graph ($F(32, 66) = 1.16$, $p = 0.60$).

[2] All these descriptions are available in Appendix E. From all descriptions, I dropped two responses as they were unuseful. These included "I'm third one to forward" and "Don't know".

in combination in relation to a point and 23 in relation to the whole graph. Absolute terms were only used in 16 cases, of which 5 when related to a point and 11 when related to the full graph.

When users considered their relation towards a point in relative terms, this was most often for the point B. This was the point closest in both the City-Block and Euclidean logic towards the user. We can see this in responses like "My position is closest to B" or "Nearest to B". Others are more descriptive, and also take the other points into account, such as in "I am closest to party B and farthest to party D" or "Very similar to B, pretty dissimilar to A, D, and C". Note that in the latter opinion, the user uses the word "similar" to mean the same as "close". For the absolute terms about a point, most users again used the point B, though now the descriptions related to positions such as "2 points from b vertically, -1 point from y horizontally" or "two blocks below B". In this case, the user has additionally considered the position of both the point B and the point labelled "You" on the axes.

When users considered their position in view of the whole graph in relative terms, they used a wide range of descriptions. Some used the fact that their position was in the upper-right quadrant, like "on the right" or "in the upper right quadrant between the origin and point B". Others deemed the position to be central or "close to the middle". Besides, some also took the fact that the points represented political parties into consideration and noted that their position was "Fairly centrist in both ways, not in an extreme position". Users using the absolute terms often gave a short description and noted their position either as "2 points above x and 2 points right of y", "two points away in the positive side of X and Y axis from the origin", or as "2,2"[3].

	Spider	Bar	Map
Eye pleasing	3.37 (1.13)	3.25 (1.21)	3.52 (0.96)
Easy to understand	3.04 (1.27)	3.67 (1.15)	3.66 (1.05)
User-friendly	2.99 (1.31)	3.66 (1.15)	3.54 (1.09)
Informative	3.25 (1.22)	3.57 (1.10)	3.55 (1.03)
Intuitive	2.94 (1.13)	3.42 (1.20)	3.39 (1.11)
Usefulness	3.12 (1.16)	3.64 (1.01)	3.64 (0.95)
Comparison utility	3.10 (1.20)	3.76 (1.10)	3.73 (1.02)
Pre-knowledge	3.66 (1.20)	3.18 (1.31)	3.42 (1.24)
Comprehensiveness	3.03 (1.17)	3.52 (1.01)	3.43 (1.12)
Representation style	3.15 (1.14)	3.61 (0.98)	3.58 (0.94)
Mean	3.17 (0.88)	3.53 (0.83)	3.55 (0.76)

Table 8.2: Opinion of the users about the three different graphics. The numbers in brackets are the standard deviations. Answers range from 1=Completely Disagree to 5=Completely Agree.

Let us now turn towards usability. Table 8.2 shows the means for each of the 10 questions for each of the visualizations. I find there are significant differences in the opinions on

[3]One other hypothesis I could put forward from this is whether a description using absolute terms is more common for users who chose the closest point in the earlier two points based on city-block distance. Testing this reveals no significant relation between using an absolute description and choosing the closest point in the city-block sense either for the graph with 3 points ($\chi^2(1) = 0.06$,$p = 0.81$) and 6 points ($\chi^2(1) = 2.05 * 10^{-31}$,$p = 1$).

the users for the three different types of graphs ($F(2, 198) = 4.55$, $p < 0.05$). A post-hoc comparison using the Tukey "honestly significant difference" test reveals that while the mean score for the spider graph ($\bar{x} = 3.17$, $s = 0.88$) was significantly different at a 95% family-wise confidence level from the mean scores from both the bar graph ($\bar{x} = 3.53$, $s = 0.83$) and the plot ($\bar{x} = 3.55$, $s = 0.76$), the mean scores for the plot and the bar were not significantly different from each other. When considered individually, there are significant differences between the three visuals for items 2($H(2) = 11.13$, $p < 0.01$) ,3 ($H(2) = 10.40$, $p < 0.01$),5 ($H(2) = 7.54$, $p < 0.05$),6 ($H(2) = 9.96$, $p < 0.01$),7 ($H(2) = 13.51$, $p < 0.01$),9 ($H(2) = 7.38$, $p < 0.05$),10 ($H(2) = 7.44$, $p < 0.05$), while the items 1 ($H(2) = 1.34$, $p = 0.51$) 4 ($H(2) = 3.10$, $p = 0.21$) and 8 ($H(2) = 4.55$, $p = 0.10$) showed no significant differences. To find out where the difference lies, I carry out a multiple comparison test for each of the items following the method as described by Siegel and Castellan Jr. (1988, pp 213-214)[4]. Doing so, I find that in all cases, the observed differences between the bar graphs and the plots were lower than the critical difference (24.06). Considering only the difference between either the bar graph or the political map and the spider graph[5], I find that for items 2, 3, 6 ,7 both the bar and the plot are significantly different from the spider plot based on the critical difference (22.53), while for items 5, 9, and 10 there is only a difference between the spider plot and the bar plot.

8.3 Conclusion

What can we conclude from this? First, that a higher graphical knowledge does not contribute to a "correct" choice. Second, that those with a higher graphical knowledge are no different from those with low graphical knowledge. Third, users seem to generally dislike spider graphs and seem to prefer bar graphs. Fourth, political maps do not seem to show many problems when they contain more parties. The only effect is that more choices lead to a wider spread of which party users consider the closest. Fifth, that in describing their position, users use both absolute and relative descriptions for the political map. This is while most VAA designers assume users see their position as relative (Evans 2004). Yet, the most interesting of all is the low degree of agreement users seem to have. Even in the simple visualizations in the survey, in no case did the users agree on which party was closest to them. This is relevant as it would mean that VAAs with a similar outcome will have a different effect based on who is looking at the visualization.

This raises the question of whether and how VAA designers should address such a level of disagreement. To begin with, they could argue it is not in their power to change the way users think and interpret the results. Their responsibility ends with a visualization of the validated and constructed match. Even so, it seems inherent to the idea of a voting "advice" application that it provides more than something as simple as visualized information. Indeed, most VAAs

[4] I calculate this using the **pgirmess** package in R. In effect, the method takes the difference between the mean ranks of each of the groups and compares this to to a value z which the method then corrects for the number of comparisons and a constant based on both the total sample size and the sample size of the two groups that are being compared (Field, Miles, and Field 2012, p.681-684). When interpreting the results, the groups are significantly different when the observed difference between the groups is larger than the calculated critical difference.

[5] I do consider only two of the three possibilities to ensure that the Type I errors are not inflated (cf. Field, Miles, and Field 2012, p.683)

are selling themselves by highlighting that the match generated is a kind of personal advice to the user instead of the more generic advice they receive in the media. But when users are unable to recognize the advice, the designer wants to give them - what VAAs aim to achieve then becomes problematic. Addressing these problems is a different matter, however. To start with, it is clear VAA designers have to visualize the result between user and party in some way. As even the simplest graph studied here — the bar graph — showed problems, it seems inevitable that there will never be a perfect visualization. Yet, designers could try to make their visualizations less complex — such as the spider graphs — and think more about how to visualize the result in as simple a way as possible. Indeed, above all else, the result of the VAA should show the match between the user and each party. One way to verify this is to engage in small pre-checks before the launch of the VAA to verify if there is disagreement between what the user considers their match is and what the VAA designer intends it to be.

9 | Conclusions

Voting Advice Applications could not have come at a more opportune time. Their goal is simple: to inform voters about the issues at play during the elections and show them their own position and the positions of the political parties on those issues. This way VAAs allow voters to answer a simple question: *what party should I vote for?* Before VAAs first appeared at the end of the 1980*s* the answer to this question depended on social class. Nowadays, we can no longer rely on this but have to "choose" our party instead (Rose and McAllister 1986). At the same time, a true "explosion" of information faces us (Fuller 2010). Where before communication between parties and voters was a one-way flow of information structured by the party, now have lost their monopoly (Alvarez 1996). Voters can demand, and comment upon, parties during and outside the elections. This forces parties to consider their positions on more issues than ever and communicate these positions to voters. This new flow of information, coupled with the appearance of new cleavages transforms elections from a clear choice into a complex conundrum.

It was this conundrum that VAAs can to entangle. After a slow start, they gained speed during the early 2000*s* and have since then established themselves. In line with their success, the first scholars started paying attention to them and found VAAs could be successful in their aim to improve the political knowledge of the voters and the electoral turnout (e.g. Kamoen, Holleman, Krouwel, et al. 2015; Marschall and Schmidt 2010; Garzia, Trechsel, and De Angelis 2017; Germann and Gemenis 2018). At the same time, they found the design of VAAs had an impact on the advice the VAAs provided (Walgrave, Aelst, and Nuytemans 2008b; Louwerse and Rosema 2014). More problematic was this design did not always live up to the required standards (Gemenis 2013a; Camp, Lefevere, and Walgrave 2014; Germann and Mendez 2016). This, coupled with both media and political parties becoming more aware (and critical) of VAAs, has made VAA vulnerable.

My motivation for this thesis was to see whether I could address this vulnerability. To be more specific, I wanted to see whether it was *possible to improve VAAs so that they can fulfil their promises.* This also required that I first established how well VAAs are actually doing and to what degree they stand up to scrutiny. I realized the only way I could do this with success was by focusing on VAA design as a process, instead of as of a series of loose steps. Then, I first assessed the quality of the VAA questionnaire. For this, I used the data from a trans-national VAA, EUVox. I did so as this allowed me to compare the quality of the questionnaire between various versions of a VAA supposed to be similar. To start with, I estimated the quality, reliability and unidimensionality of these scales used in the EUVox VAA. Here, I found these scales are often lacking, regardless of the way I constructed them.

Besides, I found large differences in the quality, reliability and unidimensionality. And not only between the topics of the scales but between the different countries as well. A closer look showed the cause of these differences was users adopting response patterns or not using all the response options available. In a second study, I delved deeper into the wording of the questions and found that altering the wording of the questions not only changes the responses but also influences the match between the user and the various parties. To verify this, I ran an experiment during the 2017 elections in the Netherlands. The message here is that the wording of the questions is by no means neutral and that VAA designers might (without intention) favour certain parties above others.

In the second part, I looked at the visualizations VAAs use to visualize the calculated match between the user and the parties included in the VAA. This is an area that scholars until now have not much explored. While scholars know how to calculate the match (e.g. Louwerse and Rosema 2014; Germann and Mendez 2016; Mendez 2017), they know little about how users *perceive* this match. Thus, I first discussed how different visualizations depend on different conceptions of what the VAA should be. I then found that an increasing number of VAAs seem to opt for impressive visuals instead of those that are clear and precise. Finally, by conducting an online experiment, I found that some visualizations confuse users. Especially when they have to decide which of the parties is closest to them. In the next section, I will go into each of these findings in more detail.

9.1 The Questionnaire of the VAA

The questionnaire was the focus of the first VAA design studies (e.g. Walgrave, Aelst, and Nuytemans 2008b). This is logical as the questionnaire decides not only what the VAA is about, but also what the position of both the user and the party is. Here, I focused on two issues: a) the quality of the scales that VAAs construct from the questions, and b) the effect question wording had on the match between the user and the parties in the VAA. This led to the following conclusions.

The majority of the scales in VAAs does not live up to the desired standards. VAA designers use scales to construct the two-dimensional political map. Despite the risk that including poorly constructed scales make the political map incomprehensible, scholars found some scales in VAAs lacking (Gemenis 2013a; Germann, Mendez, et al. 2015; Germann and Mendez 2016). Here, I assessed this problem by using data from the EUVox. This was a VAA launched for the 2014 elections for the European Parliament. EUVox had versions in 28 countries, which all shared a questionnaire on which 21 core questions were similar. EUVox had three dimensions: economic left to economic right, social progressive to social conservative and pro-EU integration to anti-EU integration. This allowed for a comparison of the quality of the scales not only over countries but also over different dimensions. Also, I constructed three different versions of each scale, based on different ideas of scale construction. These versions included the original version of the scale which included items based on theory, a version based on dynamic scale validation were I recalibrated the scales using Mokken scaling analysis, and a version based on a quasi-inductive approach in which I constructed new scales from the full pool of items. This led to a comparison between 142 scales. The three criteria on which I assessed the scales were unidimensionality, reliability, and quality. Unidimensionality

assesses the degree to which the scale measures a single underlying concept. Reliability measures if the scale is consistent. Quality measures if users used the response categories of the items as intended. I found the original scales to lack in reliability, unidimensionality, and quality. The scales based on dynamic scale validation fared better, though the level of improvement differed per country. The quasi-inductive scales scored best, though their method of construction means that in some countries only a single scale was available. This is problematic if VAA designers want to construct a two-dimensional political map.

We can trace problems with scales can back to respondents "simplifying" their response options. Using Multiple Correspondence Analysis (MCA), I investigated scales from five countries — Lithuania, Ireland, Hungary, Estonia, and the United Kingdom. This technique allowed me to visualize the underlying structures of the data. By decomposing the total variation of the scale into multiple dimensions, MCA can give a visual representation of the scale. If the scale is of high quality, this representation shows a horseshoe-like shape, in which the response categories run from *completely agree* to *completely disagree* on the first dimension, and from *neutral* to *completely agree/completely disagree* on the second dimension. The presence of ties, where two response options share a similar or position on either of the dimensions or the absence of a horseshoe shape, indicates the data is problematic. In Lithuania, I found none of the three versions of the scale could provide an economic left-right dimension that was without problems. Both in its original and DSV forms, the scale was multidimensional. Also, users simplified the 5-point Likert scale and either chose an extreme response (completely agree or completely disagree) or a non-extreme response (agree, disagree and neutral). Ireland exhibited similar problems for the original EU scale but showed a noticeable improvement after DSV. More problematic was Estonia, in which all three scales scored low on quality, reliability and unidimensionality. This was due to users simplifying their responses options into a binary agree-disagree. Also, they used the *Neutral* response as a way to hide non-opinion and did not notice the reversal of the formulation of some items. The other two countries, the United Kingdom and Hungary did show a coherent structure. Especially the United Kingdom showed both consistent dimensions and a clear horseshoe shape.

Question wording can influence the match between the user and the parties in the VAA. In an earlier study, Holleman, Kamoen, Krouwel, et al. (2016) showed that using positive or negative variants of the same questions can influence the response the user gives. Here, I followed this line of argumentation further and looked if this wording had any effect on the match between the user and the party. Any change in responses is interesting but irrelevant if the outcome would be the same. Following Holleman, Kamoen, Krouwel, et al. (2016), I choose to focus on wording statements either in a positive or negative way. So, a positive statement like "for environmental measures, taxes may be raised" also had a negative version: "For environmental measures, taxes may *not* be raised". To measure this, I designed a VAA which included both versions of the statement, and in which users were randomly shown one of them. I launched this VAA during the 2017 elections for the House of Representatives in the Netherlands and reached 6283 users. The results of this experiment were in line with expectations: whether a statement is positive or negative does matter. Yet, the effect depends both on the content of the statement, and with which party we calculate the match. For example, a party known for promoting the country to *leave* the European Union, will

profit when the statement reads "the country should leave the European Union" and will be worse off when it reads "the country should remain in the European Union" - despite the question measuring the same sentiment. In addition, the effect of the wording of the statement becomes stronger when the political sophistication is lower. This means that design choices especially affect those users for which VAAs could be most beneficial.

9.2 The Visualization of the VAA

Contrary to the questionnaire, visualizations in Voting Advice Applications have received little attention. This is interesting, given that the visualization is what displays the advice of the VAA to the user. Based on a discussion of the literature and a pilot study, I arrived at the following conclusions:

Spider plots are unsuited for VAA visualizations. While spider plots are appealing, they are more complex than bar graphs or political maps. Most important is that the shape of the area that signifies the user is often difficult to compare with that of the political parties included in the VAA. Moreover, by not maintaining a logic in the ordering of their axes, the shapes that it shows are dependent on the arbitrary choice by the designer on how to order the axes. Thus, most users in the experiment preferred the bar graphs and two-dimensional maps above the spider graph, though there was no difference in the time they took to view them.

Users can reach different conclusions based on a similar visualization. Most VAA visualizations suppose the interpretation of the VAA visualization is neutral. That is, while the effect of the visualization might be different for different users, the interpretation of the visualization is the same for everyone. Here, we found this is not the case. Showing users some parties and then adding several others showed that some noticed that the distances did not change, while others supposed they did. Also, given the same visualization users not only drew incorrect conclusions about their position and that of the parties but also differed among each other.

Users use both absolute and relative qualifiers to describe their position on a two-dimensional map. While most VAAs seem to suppose that users only apply relative qualifiers (Evans 2004), like *closer* or *farther away*, the experiment showed that they use as often absolute qualifiers in which they positioned themselves on an absolute point on the political map. This is interesting as this makes the positioning of the parties and the user as relevant as the relative distances between them.

9.3 Limitations and Future Research

While I attempted to be as complete as possible, there were some limitations I ran into during this thesis. I will discuss these in brief and also provide some ideas for further research.

The first limitation lies in the use of the EUVox data-set. Whereas most VAAs are constructed for national elections, the EUVox VAA was constructed for the elections for the European Parliament. These elections did not only have a lower electoral turnout, but they are also farther away from day-to-day politics than national elections. As a result, users might be unaware of the issues included in the VAA or their own positions on it. As such,

it would be interesting to observe similar VAAs during the national elections in the same countries as those studied here to establish whether the VAAs score similarly on reliability, unidimensionality and quality. I especially expect the level of quality to increase as this is dependent on the user understanding what the question is about. A greater degree of familiarity with the topic of the question, which seems to be more likely in the case of national elections, will increase the user's understanding.

The second limitation lay in the small set-up of the visualization study. Here, I administered an online survey to test how users reacted to certain visualizations. Even so, this leads to the results coming from a small group of users. Also, by running the survey online, I was unable to identify *what* users were looking at, only what they *said* they were looking at. For the future, it is advisable to carry out such surveys in a laboratory in which the behaviour and reaction of the respondents can be more accurately observed. Also, in this case, I constructed the visuals of the VAAs myself. In future studies, these visualizations could be made more realistic by drawing upon actual visualizations as they occur in VAA. This prevents the study from focusing visualizations in which the parties positions are rather extreme.

Up until now, the literature on VAAs has been growing at a steady pace. This prompts the question of where the field should be heading next. For this, I would like to make a few suggestions. To start with, I would consider it fruitful when VAA designers would expand the debate on the philosophical foundations of the VAA. Apart from the first attempts by Fossen and Anderson (2014), Fossen and Brink (2015) and Anderson and Fossen (2014), there has been little debate on what it is VAA do and should do. Given that the awareness of the impact of VAAs seems to be increasing not only amongst scientists but also amongst voters and politicians, it would benefit VAA research if there would be more awareness on the purpose of VAAs. Besides, the research on VAA visualizations should be expanded. Until now, visualizations in VAAs have received limited attention. This is interesting, especially given that citizens know some VAAs - like *Kieskompas* - through their visualization. Here, significant advances can be made when this area is linked with the burgeoning literature on *information visualization*, which alike VAAs aims to structure complex information in a visual form.

10 | Bibliography

50Plus (2016). *Verkiezingsprogramma 2017-2021 - Omdat ouderen het niet meer pikken!* Den Haag: 50Plus Partij.

Abdi, H. (2007a). *Singular value decomposition (SVD) and generalized singular value decomposition (GSVD).* In: *Encyclopedia of Measurement and Statistics.* Ed. by N. Salkind. Thousand Oaks, CA: Sage, pp. 907–912. DOI: 10.4135/9781412952644.

— (2007b). *The Eigen-Decomposition: Eigenvalues and Eigenvectors.* In: *Encyclopedia of Measurement and Statistics.* Ed. by N. Salkind. Thousand Oaks, CA: Sage, pp. 304–308. DOI: 10.4135/9781412952644.

Abdi, H. and L. J. Williams (2010). "Principal Component Analysis". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 433–459. DOI: 10.1002/wics.101.

Albo, Y., J. Lanir, P. Bak, and S. Rafaeli (2016). "Off the Radar: Comparative Evaluation of Radial Visualization Solutions for Composite Indicators". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 569–578. DOI: 10.1109/TVCG.2015.2467322.

Alvarez, R. M. (1996). *Information and Elections.* Ann Arbor, MI: University of Michigan Press. DOI: 10.3998/mpub.15100.

Alvarez, R. M., I. Levin, P. Mair, and A. H. Trechsel (2014). "Party preferences in the digital age: The impact of voting advice applications". In: *Party Politics* 20.2, pp. 227–236. DOI: 10.1177/1354068813519960.

Alvarez, R. M., I. Levin, A. H. Trechsel, and K. Vassil (2014). "Voting Advice Applications: How Useful and for Whom?" In: *Journal of Information Technology & Politics* 11.1, pp. 82–101. DOI: 10.1080/19331681.2013.873361.

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.* Ed. by D. Garzia and S. Marschall. Colchester: ECPR Press, pp. 217–226.

Andreadis, I. (2014). "Data Quality and Data Cleaning". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective.* Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 79–92.

Ansolabehere, S., J. M. Snyder, and C. Stewart (2001). "The Effects of Party and Preferences on Congressional Roll Call Voting". In: *Legislative Studies Quarterly* 26.4, pp. 533–72. DOI: 10.2307/440269.

Ark, L. A. van der (2007). "Mokken Scale Analysis in R". In: *Journal of Statistical Software* 20 (11), pp. 1–19. DOI: 10.18637/jss.v020.i11.

— (2012). "New Developments in Mokken Scale Analysis in R". In: *Journal of Statistical Software* 48 (5), pp. 1–27. DOI: 10.18637/jss.v048.i05.

Ark, L. A. van der, D. W. van der Palm, and K. Sijtsma (2011). "A latent class approach to estimating test-score reliability". In: *Applied Psychological Measurement* 35.5, pp. 380–392. DOI: 10.1177/0146621610392911.

Austin, E. J., I. J. Deary, and V. Egan (2006). "Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items". In: *Personality and Individual Differences* 40.6, pp. 1235–1245. DOI: 10.1016/j.paid.2005.10.018.

Baka, A., L. Figgou, and V. Triga (2012). "'Neither agree, nor disagree': a critical analysis of the middle answer category in Voting Advice Applications". In: *International Journal of Electronic Governance* 5.3, pp. 244–263. DOI: 10.1504/IJEG.2012.051306.

Baker, R. S., A. T. Corbett, and K. R. Koedinger (2001). "Toward a model of learning data representations". In: *Twenty-Third Annual Conference of the Cognitive Science Society*. Ed. by J. D. Moore and K. Stenning. Mahwah, NJ: Erlbaum, pp. 45–50.

Baron-Epel, O., G. Kaplan, R. Weinstein, and M. S. Green (2010). "Extreme and acquiescence bias in a bi-ethnic population". In: *European Journal of Public Health* 20.5, pp. 543–548. DOI: 10.1093/eurpub/ckq052.

Bartels, L. M. (1986). "Issue Voting Under Uncertainty: An Empirical Test". In: *American Journal of Political Science* 30 (4), pp. 709–728. DOI: 10.2307/2111269.

Bassili, J. N. and J. A. Krosnick (2000). "Do Strength-Related Attitude Properties Determine Susceptibility to Response Effects? New Evidence From Response Latency, Attitude Extremity, and Aggregate Indices". In: *Political Psychology* 21.1, pp. 107–132. DOI: 10.1111/0162-895X.00179.

Baumgartner, H. and J.-B. E. M. Steenkamp (2001). "Response Styles in Marketing Research: A Cross-National Investigation". In: *Journal of Marketing Research* 38.2, pp. 143–156. DOI: 10.1509/jmkr.38.2.143.18840.

Beller, J. and S. Kliem (2013). *GetR: Calculate Guttman error trees in R*. R package version 0.1. URL: https://CRAN.R-project.org/package=GetR.

Benoit, K. and M. Laver (2006). *Party Policy in Modern Democracies*. London, United Kingdom: Routledge.

— (2012). "The dimensionality of political space: Epistemological and methodological considerations". In: *European Union Politics* 13.2, pp. 194–218. DOI: 10.1177/1465116511434618.

Benzécri, J.-P. (1973a). *Analyse des Données 1: La Taxonomie*. Paris: Dunod.

— (1973b). *Analyse des Données 2: Analyse des Correspondances*. Paris: Dunod.

Blais, A. (2000). *To vote or not to vote? The merits and limits of rational-choice theory*. Pittsburgh, PA: University of Pittsburgh Press.

Blasius, J. and J. C. Gower (2005). "Multivariate Prediction with Nonlinear Principal Components Analysis: Application". In: *Quality & Quantity* 39 (4), pp. 373–390. DOI: 10.1007/s11135-005-3006-0.

Blasius, J. and M. Greenacre (2006). "Multiple Correspondence Analysis and Related Methods in Practice". In: *Multiple Correspondence Analysis and Related Methods*. Ed. by M. Greenacre and J. Blasius. Boca Raton, FL: Chapman & Hall/CRC, pp. 3–40.

Blasius, J., O. Nenadić, and V. Thiessen (2017). "The Dirty Data Index - Assessing the Quality of Survey Data in International Comparison". In: *Statistica Applicata - Italian Journal of Applied Statistics* 29 (2–3), pp. 137–152. DOI: 10.26398/IJAS.0029-007.

Blasius, J. and V. Thiessen (2001a). "Methodological Artifacts in Measures of Political Efficacy and Trust: A Multiple Correspondence Analysis". In: *Political Analysis* 9 (1), pp. 1–20. DOI: 10.1093/oxfordjournals.pan.a004862.

— (2001b). "The Use of Neutral Responses in Survey Questions: An Application of Multiple Correspondence Analysis". In: *Journal of Official Statistics* 17.3, pp. 351–367.

— (2012). *Assessing the Quality of Survey Data*. London/Thousand Oaks/New Delhi/Singapore: SAGE.

— (2015). "Should we trust survey data? Assessing response simplification and data fabrication". In: *Social Science Research* 52, pp. 479–493. DOI: 10.1016/j.ssresearch.2015.03.006.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York,NY: John Wiley.

Bostic, T. J., D. McGartland Rubio, and M. Hood (2000). "A Validation of the Subjective Vitality Scale Using Structural Equation Modeling". In: *Social Indicators Research* 52 (3), pp. 313–324. DOI: 10.1023/A:1007136110218.

Boulianne, S. (2009). "Does Internet Use Affect Engagement? A Meta-Analysis of Research". In: *Political Communication* 26 (2), pp. 193–211. DOI: 10.1080/10584600902854363.

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

Brennan, J. (2011). "The Right to a Competent Electorate". In: *The Philosophical Quarterly* 61 (245), pp. 700–724. DOI: 10.1111/j.1467-9213.2011.699.x.

Bruinsma, B. and K. Gemenis (2019). "Validating Wordscores: The Promises and Pitfalls of Computational Text Scaling". In: *Communication Methods and Measures*.

Budge, I. (2001). "Validating Party Policy Placements". In: *British Journal of Political Science* 31.1, pp. 210–223. DOI: 10.1017/S0007123401230087.

Budge, I. and M. Laver (1992). *Party Policy and Government Coalitions*. Basingstoke: Macmillan.

Budge, I. and T. Meyer (2013). "Understanding and Validating the Left-Right Scale (RILE)". In: *Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts*. Ed. by A. Volkens, J. Bara, I. Budge, M. D. McDonald, and H.-D. Klingemann. Oxford, United Kingdom: Oxford University Press, pp. 85–106.

Budge, I. and P. Pennings (2007). "Do they work? Validating computerised word frequency estimates against policy series". In: *Electoral Studies* 26.1, pp. 121–129. DOI: 10.1016/j.electstud.2006.04.002.

Burch, M., F. Bott, F. Beck, and S. Diehl (2008). "Cartesian vs. Radial – A Comparative Evaluation of Two Visualization Tools". In: *Advances in Visual Computing*. Ed. by G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, F. Porikli, J. Peters, J. Klosowski, L. Arns, Y. K. Chun, T.-M. Rhyne, and L. Monroe. Berlin/Heidelberg, Germany: Springer Berlin Heidelberg, pp. 151–160. DOI: 10.1007/978-3-540-89639-5_15.

Burden, B. C. (2004). "Candidate Positioning in US Congressional Elections". In: *British Journal of Political Science* 34.2, pp. 211–227. DOI: 10.1017/S000712340400002X.

Camp, K. van, J. Lefevere, and S. Walgrave (2014). "The Content and Formulation of Statements in Voting Advice Applications. A Comparative Analysis of 26 VAAs". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 11–32.

Caplan, B. (2007). *The Myth of the Rational Voter: Why Democracies choose bad Policies*. Princeton, NJ: Princeton University Press.

Carmines, E. G. and R. A. Zeller (1979). *Reliability and Validity Assessment*. Quantitative Applications in the Social Sciences 07-017. Beverly Hills, CA/London, United Kingdom: SAGE Publications.

Carpenter, P. A. and P. Shah (1998). "A Model of the Perceptual and Conceptual Processes in Graph Comprehension". In: *Journal of Experimental Psychology: Applied* 4.2, pp. 75–100. DOI: 10.1037/1076-898X.4.2.75.

Cattell, R. B. (1966). "The Scree Test For The Number Of Factors". In: *Multivariate Behavioral Research* 1.2, pp. 245–276. DOI: 10.1207/s15327906mbr0102_10.

Cedroni, L. and D. Garzia (2010). *Voting Advice Applications in Europe: The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napels, Italy: ScriptaWeb.

Chessa, A. G. and B. C. Holleman (2007). "Answering attitudinal questions: modelling the response process underlying contrastive questions". In: *Applied Cognitive Psychology* 21.2, pp. 203–225. DOI: 10.1002/acp.1337.

ChristenUnie (2016). *Hoopvol realistisch - Voorstellen voor een samenleving met toekomst - Verkiezingsprogramma 2017-2021*. Amersfoort: ChristenUnie.

Clark, H. H. (1976). *Semantics and Comprehension*. Den Haag, The Netherlands: Mouton.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

— (1993). *Visualizing Data*. Murray HIll, NJ: AT&T Bell Laboratories.

Cleveland, W. S., M. E. McGill, and R. McGill (1988). "The Shape Parameter of a Two-Variable Graph". In: *Journal of the American Statistical Association* 83.402, pp. 289–300. DOI: 10.2307/2288843.

Cleveland, W. S. and R. McGill (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387, pp. 531–554. DOI: 10.2307/2288400.

— (1985). "Graphical Perception and Graphical Methods for Analyzing Scientific Data". In: *Science* 229.4716, pp. 828–833. DOI: 10.1126/science.229.4716.828.

— (1986). "An Experiment in Graphical Perception". In: *International Journal of Man-Machine Studies* 25.5, pp. 491–500. DOI: 10.1016/S0020-7373(86)80019-0.

Clinton, J., S. Jackman, and D. Rivers (2004). "The Statistical Analysis of Roll Call Data". In: *American Political Science Review* 98.2, pp. 355–370. DOI: 10.1017/S0003055404001194.

Cockburn, A. (2004). "Revisiting 2D vs 3D Implications on Spatial Memory". In: *Proceedings of the Fifth Conference on Australasian User Interface*. Vol. 28. AUIC '04. Dunedin, New Zealand: Australian Computer Society, Inc., pp. 25–31. URL: http://dl.acm.org/citation.cfm?id=976310.976314.

Cockburn, A. and B. McKenzie (2004). "Evaluating Spatial Memory in Two and Three Dimensions". In: *International Journal of Human-Computer Studies* 61.3, pp. 359–373. DOI: 10.1016/j.ijhcs.2004.01.005.

Converse, P. E. (1964). "The nature of belief systems in mass publics". In: *Ideology and Discontent*. Ed. by D. E. Apter. New York, NY: The Free Press of Glencoe, pp. 206–261.

— (1974). "Comment: The Status of Nonattitudes". In: *American Political Science Review* 68.2, pp. 650–660. DOI: 10.1017/S0003055400117447.

Corbetta, P. (2003). *Social Research Theory. Theory, Methods and Techniques*. Trans. by B. Patrick. London, United Kingdom: SAGE Publications.

Couper, M. P., R. Tourangeau, F. G. Conrad, and E. Singer (2006). "Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment". In: *Social Science Computer Review* 24 (2), pp. 227–245. DOI: 10.1177/0894439305281503.

Cronbach, L. J. and P. E. Meehl (1955). "Construct validity in psychological tests". In: *Psychological Bulletin* 52.4, pp. 281–302. DOI: 10.1037/h0040957.

Dalege, J., D. Borsboom, F. van Harreveld, and H. L. J. van der Maas (2017). "A Network Perspective on Political Attitudes: Testing the Connectivity Hypothesis". In: *CoRR* abs/1705.00193. arXiv: 1705.00193.

Dalkey, N. (1969). "An experimental study of group opinion: the Delphi method". In: *Futures* 1.5, pp. 408–426. DOI: 10.1016/S0016-3287(69)80025-X.

Dalrymple, K. E. and D. A. Scheufele (2007). "Finally Informing the Electorate? How the Internet Got People Thinking about Presidential Politics in 2004". In: *Harvard International Journal of Press/Politics* 12 (3), pp. 96–111. DOI: 10.1177/1081180X07302881.

Dalton, R. J. (1988). *Citizen Politics in Western Democracies: Public Opinion and Political Parties in the United States, Great Britain, West-Germany, and France*. Chatham, NJ: Chatham House Publishers.

— (2000). "The Decline of Party Identifications". In: *Parties without Partisans: Political Change in Advanced Industrial Democracies*. Ed. by M. Wattenberg and R. J. Dalton. Oxford: Oxford University Press, pp. 19–37.

— (2014). *Citizen Politics: Public Opinion and Political Parties in Advanced Industrial Democracies*. Thousand Oaks, CA: CQ Press.

Das, A. (2009). *Contextual Interactions in Visual Processing*. In: *Encyclopedia of Neuroscience*. Ed. by L. R. Squire. Oxford: Academic Press, pp. 145–158. DOI: 10.1016/B978-008045046-9.00209-6.

Delli Carpini, M. X. and S. Keeter (1997). *What Americans Know about Politics and Why It Matters*. New Haven, CT: Yale University Press.

Diehl, S., F. Beck, and M. Burch (2010). "Uncovering Strengths and Weaknesses of Radial Visualizations—an Empirical Approach". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 935–942. DOI: 10.1109/TVCG.2010.209.

Dimara, E., A. Bezerianos, and P. Dragicevic (2017). "The Attraction Effect in Information Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1, pp. 471–480. DOI: 10.1109/TVCG.2016.2598594.

Dimitrova, D. V., A. Shehata, J. Strömbäck, and L. W. Nord (2014). "The Effects of Digital Media on Political Knowledge and Participation in Election Campaigns: Evidence From Panel Data". In: *Communication Research* 41 (1), pp. 95–118. DOI: 10.1177/0093650211426004.

Dinas, E. and K. Gemenis (2010). "Measuring Parties' Ideological Positions With Manifesto Data: A Critical Evaluation of the Competing Methods". In: *Party Politics* 16.4, pp. 427–450. DOI: 10.1177/1354068809343107.

Dinas, E., A. H. Trechsel, and K. Vassil (2014). "A look into the mirror: Preferences, representation and electoral participation". In: *Electoral Studies* 36, pp. 290–297. DOI: 10.1016/j.electstud.2014.04.011.

Djouvas, C., F. Mendez, and N. Tsapatsoulis (2016). "Mining online political opinion surveys for suspect entries: An interdisciplinary comparison". In: *Journal of Innovation in Digital Ecosystems* 3.2, pp. 172–182. DOI: 10.1016/j.jides.2016.11.003.

Dodou, D. and J. C. F. de Winter (2014). "Social desirability is the same in offline, online, and paper surveys: A meta-analysis". In: *Computers in Human Behavior* 36, pp. 487–495. DOI: 10.1016/j.chb.2014.04.005.

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

Draper, G. M., Y. Livnat, and R. F. Riesenfeld (2009). "A Survey of Radial Methods for Information Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 15.5, pp. 759–776. DOI: 10.1109/TVCG.2009.23.

Dunteman, G. H. (1989). *Principal Components Analysis*. Quantitative Applications in the Social Sciences 07-069. Newbury Park/London/New Delhi: SAGE Publications.

Eijk, C. van der and J. Rose (2015). "Risky Business: Factor Analysis of Survey Data – Assessing the Probability of Incorrect Dimensionalisation". In: *PLOS ONE* 10 (3), pp. 1–31. DOI: 10.1371/journal.pone.0118900.

Eisinga, R. and P. H. Franses (1996). "Testing for convergence in left-right ideological positions". In: *Quality & Quantity* 30.4, pp. 345–359. DOI: 10.1007/BF00170141.

Evans, G., A. Heath, and M. Lalljee (1996). "Measuring Left-Right and Libertarian-Authoritarian Values in the British Electorate". In: *The British Journal of Sociology* 47.1, pp. 93–112. DOI: 10.2307/591118.

Evans, J. A. J. (2004). *Voters & Voting. An Introduction*. London/Thousand Oaks/New Delhi: SAGE Publications.

Feldman, R. (2013). "Filled Radar Charts Should not be Used to Compare Social Indicators". In: *Social Indicators Research* 111.3, pp. 709–712. DOI: 10.1007/s11205-012-0028-6.

Field, A., J. Miles, and Z. Field (2012). *Discovering Statistics Using R*. London, United Kingdom: SAGE Publications.

Forsell, C. (2014). "Heuristic Evaluation: A Guide for Use in Information Visualization". In: *Information Visualisation: Techniques, Usability and Evaluation*. Ed. by E. Banissi, F. T. Marchese, C. Forsell, and J. Johansson. Newcastle upon Tyne, United Kingdom: Cambridge Scholars Publishing, pp. 177–199.

Forum voor Democratie (2016). *Verkiezingsprogramma 2017-2021*. Amsterdam: Forum voor Democratie.

Fossen, T. and J. Anderson (2014). "What's the point of voting advice applications? Competing perspectives on democracy and citizenship". In: *Electoral Studies* 36, pp. 244–251. DOI: 10.1016/j.electstud.2014.04.001.

Fossen, T., J. Anderson, and W. Tiemeijer (2012). "Wijzer stemmen? StemWijzer, Kieskompas en het voorgeprogrammeerde electoraat. Hoe internet ons leven leidt". In: *Voorge-*

programmeerd. Ed. by C. C. G. van 't Hof, J. Timmer, and R. van Est. Den Haag, The Netherlands: Boom Lemma, pp. 163–188.

Fossen, T. and B. van den Brink (2015). "Electoral Dioramas: On the Problem of Representation in Voting Advice Applications". In: *Representation* 51.3, pp. 341–358. DOI: 10.1080/00344893.2015.1090473.

Francia, P. L. and P. S. Herrnson (2007). "Keeping it Professional: The Influence of Political Consultants on Candidate Attitudes toward Negative Campaigning". In: *Politics & Policy* 35.2, pp. 246–272. DOI: 10.1111/j.1747-1346.2007.00059.x.

Franklin, M. N., T. Mackie, and H. Valen (1992). *Electoral Change : Responses to evolving social and attitudinal structures in western countries*. New York, NY: Cambridge University Press.

Freedman, E. G. and P. Shah (2002). "Toward a Model of Knowledge-Based Graph Comprehension". In: *Diagrammatic Representation and Inference*. Ed. by M. Hegarty, B. Meyer, and N. H. Narayanan. Berlin/Heidelberg, Germany: Springer Berlin Heidelberg, pp. 18–30.

Fuller, J. (2010). *What Is Happening to News: The Information Explosion and the Crisis in Journalism*. Chicago, IL: The University of Chicago Press.

Gabel, M. and S. Hix (2002). "Defining the Eu Political Space: An Empirical Study of the European Elections Manifestos, 1979-1999". In: *Comparative Political Studies* 35.8, pp. 934–964. DOI: 10.1177/001041402236309.

Gabel, M. and J. Huber (2000). "Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifestos Data". In: *American Journal of Political Science* 44.1, pp. 94–103. DOI: 10.2307/2669295.

Galesic, M. and M. Bosnjak (2009). "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey". In: *Public Opinion Quarterly* 73.2, pp. 349–360. DOI: 10.1093/poq/nfp031.

Garcia-Retamero, R., E. T. Cokely, S. Ghazal, and A. Joeris (2016). "Measuring Graph Literacy without a Test: A Brief Subjective Assessment". In: *Medical Decision Making* 36.7, pp. 854–867. DOI: 10.1177/0272989X16655334.

Gärdenfors, P. (2000). *Conceptual Spaces - The Geometry of Thought*. Cambridge, MI: The MIT Press.

Garry, J., N. Matthews, and J. Wheatley (2017). "Dimensionality of Policy Space in Consociational Northern Ireland". In: *Political Studies* 65.2, pp. 493–511. DOI: 10.1177/0032321716658917.

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 a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 105–114.

Garzia, D. and S. Marschall (2014). "The Lausanne Declaration on Voting Advice Applications". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 227–228.

Garzia, D., A. H. Trechsel, and A. De Angelis (2017). "Voting Advice Applications and Electoral Participation: A Multi-Method Study". In: *Political Communication* 34.3, pp. 424–443. DOI: 10.1080/10584609.2016.1267053.

Garzia, D., A. H. Trechsel, K. Vassil, and E. Dinas (2014). "Indirect Campaigning: Past, Present and Future of Voting Advice Applications". In: *The Internet and Democracy in Global Perspective*. Ed. by B. Grofman, A. H. Trechsel, and M. Franklin. Cham, Switzerland: Springer International Publishing, pp. 25–41.

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". In: *The Public Opinion Quarterly* 58.2, pp. 241–254. DOI: 10.1086/269420.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gelman, A., C. Pasarica, and R. Dodhia (2002). "Let's Practice What We Preach". In: *The American Statistician* 56 (2), pp. 121–130. DOI: 10.1198/000313002317572790.

Gemenis, K. (2012). "Proxy documents as a source of measurement error in the Comparative Manifestos Project". In: *Electoral Studies* 31 (3), pp. 594–604. DOI: 10.1016/j.electstud.2012.01.002.

— (2013a). "Estimating parties' policy positions through voting advice applications: Some methodological considerations". In: *Acta Politica* 48.3, pp. 268–295. DOI: 10.1057/ap.2012.36.

— (2013b). "What to Do (and Not to Do) with the Comparative Manifestos Project Data". In: *Political Studies* 61.1suppl, pp. 3–23. DOI: 10.1111/1467-9248.12015.

— (2015). "An iterative expert survey approach for estimating parties' policy positions". In: *Quality & Quantity* 49 (6), pp. 2291–2306. DOI: 10.1007/s11135-014-0109-5.

Gemenis, K., B. Bruinsma, C. Djouvas, V. Manavopoulos, and F. Mendez (2017). *StemConsult: Voting Advice Application data for the 2017 parliamentary election in the Netherlands*. English.

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 a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 33–47.

Gemenis, K. and M. Rosema (2014). "Voting Advice Applications and electoral turnout". In: *Electoral Studies* 36, pp. 281–289. DOI: 10.1016/j.electstud.2014.06.010.

Gerbing, D. W. and J. C. Anderson (1988). "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment". In: *Journal of Marketing Research* 25.2, pp. 186–192. DOI: 10.2307/3172650.

Germann, M. and K. Gemenis (2018). "Getting Out the Vote With Voting Advice Applications". In: *Political Communication* 0.0, pp. 1–22. DOI: 10.1080/10584609.2018.1526237.

Germann, M. and F. Mendez (2016). "Dynamic scale validation reloaded. Assessing the psychometric properties of latent measures of ideology in VAA spatial maps". In: *Quality & Quantity* 50 (3), pp. 981–1007. DOI: 10.1007/s11135-015-0186-0.

Germann, M., F. Mendez, J. Wheatley, and U. Serdült (2015). "Spatial maps in voting advice applications: The case for dynamic scale validation". In: *Acta Politica* 50 (2), pp. 214–238. DOI: 10.1057/ap.2014.3.

Gifi, A. (1980). *Niet-lineaire Multivariate Analyse*. Leiden, The Netherlands: Department of Data Theory FSW/RUL.

— (1990). *Nonlinear Multivariate Analysis*. New York, NY: Wiley.

Giger, N. and H. Klüver (2016). "Voting Against Your Constituents? How Lobbying Affects Representation". In: *American Journal of Political Science* 60.1, pp. 190–205. DOI: 10.1111/ajps.12183.

Gilbert, J. and L. Gilbert (1995). *Linear Algebra and Matrix Theory*. San Diego, CL: Academic Press.

Giora, R., O. Fein, K. Aschkenazi, and I. Alkabets-Zlozover (2007). "Negation in Context: A Functional Approach to Suppression". In: *Discourse Processes* 43.2, pp. 153–172. DOI: 10.1080/01638530709336896.

Glazer, N. (2011). "Challenges with graph interpretation: a review of the literature". In: *Studies in Science Education* 47.2, pp. 183–210. DOI: 10.1080/03057267.2011.605307.

Graaf, J. de (2010). "The Irresistible Rise of Stemwijzer". In: *Voting Advice Applications: The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napoli: ScriptaWeb, pp. 35–46.

Greenacre, M. (2017). *Correspondence Analysis in Practice*. 3rd ed. Boca Raton, FL: CRC Press.

Greenacre, M. and J. Blasius (2006). *Multiple Correspondence Analysis and Related Methods*. Boca Raton, FL: Chapman & Hall/CRC.

Greenleaf, E. A. (1992). "Measuring Extreme Response Style". In: *Public Opinion Quarterly* 56.3, pp. 328–351. DOI: 10.1086/269326.

Grimmer, J. and B. M. Stewart (2013). "Text as Data: The Promise and Pitfals of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, pp. 267–297. DOI: 10.1093/pan/mps028.

GroenLinks (2016). *Tijd voor Verandering - Verkiezingsprogramma GroenLinks 2017-2021*. Amsterdam: GroenLinks.

Groot, L. F. M. (2003a). *Criteria to Evaluate Voting Indicators and a Recipe for a New One*. Tech. rep. Amsterdam: SISWO/Netherlands Institute for the Social Sciences.

— (2003b). "Een kritische evaluatie van de StemWijzer 2002". In: *Beleid en Maatschappij* 30.1, pp. 20–30.

— (2003c). "Het verrassingseffect van de StemWijzer is niet verrassend". In: *Beleid en Maatschappij* 30.3, pp. 201–203.

— (2004). "De Voorspelkracht van Stemprogramma's". In: *Tijdschrift voor Politieke Ekonomie* 25 (3), pp. 88–115.

Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Gschwend, T. and S.-O. Proksch (2010). "Improving the measurement of policy preferences in surveys: Bringing the status-quo back in". In: *REPCONG-Workshop Policy Representation Europe*. Mannheim, Germany.

Hansen, M. E. (2008). "Back to the archives? A critique of the Danish part of the manifesto dataset". In: *Scandinavian Political Studies* 31.2, pp. 201–216. DOI: 10.1111/j.1467-9477.2008.00202.x.

Hayashi, C. (1950). "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of View". In: *Annals of the Institute of Statistical Mathematics* 2.1, pp. 35–47. DOI: 10.1007/BF02919500.

— (1952). "On the Prediction of Phenomena from Qualitative Data and the Quantification of Qualitative Data from the Mathematico-Statistical Point of View". In: *Annals of the Institute of Statistical Mathematics* 3.1, pp. 69–98. DOI: 10.1007/BF02949778.

— (1953). "Multidimensional Quantification. with the Applications to Analysis of Social Phenomena". In: *Annals of the Institute of Statistical Mathematics* 5 (2), pp. 121–143. DOI: 10.1007/BF02949809.

He, J., F. J. R. van de Vijver, V. H. Fetvadjiev, A. de Carmen Dominguez Espinosa, B. Adams, I. Alonso-Arbiol, A. Aydinli-Karakulak, C. Buzea, R. Dimitrova, A. Fortin, G. Hapunda, S. Ma, R. Sargautyte, S. Sim, M. K. Schachner, A. Suryani, P. Zeinoun, and R. Zhang (2017). "On Enhancing the Cross-Cultural Comparability of Likert-Scale Personality and Value Measures: A Comparison of Common Procedures". In: *European Journal of Personality* 31 (6), pp. 642–657. DOI: 10.1002/per.2132.

Hemker, B. T., K. Sijtsma, and I. W. Molenaar (1995). "Selection of Unidimensional Scales From a Multidimensional Item Bank in the Polytomous Mokken IRT Model". In: *Applied Psychological Measurement* 19 (4), pp. 337–352. DOI: 10.1177/014662169501900404.

Hippler, H.-J. and N. Schwarz (1986). "Not Forbidding Isn't Allowing: The Cognitive Basis of the Forbid-Allow Asymmetry". In: *Public Opinion Quarterly* 50.1, pp. 87–96. DOI: 10.1086/268961.

Holleman, B. (Nov. 1999). "The Nature of the Forbid/Allow Asymmetry". In: *Sociological Methods & Research* 28.2, pp. 209–244. DOI: 10.1177/0049124199028002004.

— (2000). *The forbid/allow asymmetry: On the cognitive mechanisms underlying wording effects in questions*. Amsterdam/Atlanta, GA: Rodopi.

Holleman, B. and N. Kamoen (2017). "Attitude strength as an explanation for wording effects in political opinion questions". In: *Conference of the European Survey Research Association*. Lisbon, Portugal.

Holleman, B., N. Kamoen, A. Krouwel, J. van de Pol, and C. de Vreese (2016). "Positive vs. Negative: The Impact of Question Polarity in Voting Advice Applications". In: *PLOS ONE* 11.10, e0164184, pp. 1–17. DOI: 10.1371/journal.pone.0164184.

Holleman, B., N. Kamoen, and C. de Vreese (2013). "Voting advice via internet: answers, attitudes and voting intentions". In: *Tijdschrift voor Taalbeheersing* September 2012, pp. 1–30.

Horn, J. L. (1965). "A rationale and test for the number of factors in factor analysis". In: *Psychometrika* 30.2, pp. 179–185. DOI: 10.1007/BF02289447.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Huff, D. (1954). *How to Lie with Statistics*. New York, NY: W.W. Norton & Company.

Hug, S. (2010). "Selection effects in roll call votes". In: *British Journal of Political Science* 40 (1), pp. 225–235. DOI: 10.1017/S0007123409990160.

Humphreys, M. and M. Laver (2010). "Spatial Models, Cognitive Metrics, and Majority Rule Equilibria". In: *British Journal of Political Science* 40.1, pp. 11–30. DOI: 10.1017/S0007123409990263.

Husson, F., J. Josse, and J. Pagès (2010). *Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?* Tech. rep. Agrocampus - Applied Mathematics Department.

Husson, F., S. Lê, and J. Pagès (2011). *Exploratory Multivariate Analysis by Example Using R*. Boca Raton, FL: CRC Press.

Irvin, R. A. and J. Stansbury (2004). "Citizen Participation in Decision Making: Is It Worth the Effort?" In: *Public Administration Review* 64 (1), pp. 55–65. DOI: 10.1111/j.1540-6210.2004.00346.x.

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

John, M. S., M. B. Cowen, H. S. Smallman, and H. M. Oonk (2001). "The Use of 2D and 3D Displays for Shape-Understanding versus Relative-Position Tasks". In: *Human Factors* 43.1, pp. 79–98. DOI: 10.1518/001872001775992534.

Johns, R. (2005). "One Size Doesn't Fit All: Selecting Response Scales For Attitude Items". In: *Journal of Elections, Public Opinion and Parties* 15.2, pp. 237–264. DOI: 10.1080/13689880500178849.

Jolliffe, I. T. (2004). *Principal Component Analysis*. 2nd ed. New York, NY: Springer.

Jong, M. G. de, J.-B. E. M. Steenkamp, J.-P. Fox, and H. Baumgartner (2008). "Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation". In: *Journal of Marketing Research* 45.1, pp. 104–115. DOI: 10.1509/jmkr.45.1.104.

Jürges, H. and J. Winter (2013). "Are Anchoring Vignettes Ratings Sensitive to Vignette Age and Sex?" In: *Health Economics* 22 (1), pp. 1–13. DOI: 10.1002/hec.1806.

Just, M. A. and P. A. Carpenter (1971). "Comprehension of negation with quantification". In: *Journal of Verbal Learning and Verbal Behavior* 10.3, pp. 244–253. DOI: 10.1016/S0022-5371(71)80051-8.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kaid, L. L. and C. Holtz-Bacha, eds. (2008). *Encyclopedia of Political Communication*. London, UK: SAGE.

Kamoen, N. (2012). "Positive versus Negative - A cognitive perspective on wording effects for contrastive questions in attitude surveys". PhD thesis. Utrecht, The Netherlands: Universiteit Utrecht.

Kamoen, N., B. Holleman, A. Krouwel, J. van de Pol, and C. de Vreese (2015). "The Effect of Voting Advice Applications on Political Knowledge and Vote Choice". In: *Irish Political Studies* 30.4, pp. 595–618. DOI: 10.1080/07907184.2015.1099096.

Kamoen, N., B. Holleman, P. Mak, T. Sanders, and H. van den Bergh (2017). "Why Are Negative Questions Difficult to Answer? On the Processing of Linguistic Contrasts in Surveys". In: *Public Opinion Quarterly* 81.3, pp. 613–635. DOI: 10.1093/poq/nfx010.

Kastellec, J. P. and E. L. Leoni (2007). "Using Graphs Instead of Tables in Political Science". In: *Perspectives on Politics* 5 (4), pp. 755–771. DOI: 10.1017/S1537592707072209.

Katsanidou, A. and S. Otjes (2016). "How the European debt crisis reshaped national political space: The case of Greece". In: *European Union Politics* 17 (2), pp. 262–284. DOI: 10.1177/1465116515616196.

Kaup, B., J. Lüdtke, and R. A. Zwaan (2006). "Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed?" In: *Journal of Pragmatics* 38.7. Special Issue: Processes and Products of Negation, pp. 1033–1050. DOI: 10.1016/j.pragma.2005.09.012.

King, G., R. O. Keohane, and S. Verba (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, G., C. J. L. Murray, J. A. Salomon, and A. Tandon (2004). "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research". In: *American Political Science Review* 98 (1), pp. 191–207. DOI: 10.1017/S000305540400108X.

Kleinnijenhuis, J., J. van de Pol, A. M. J. van Hoof, and A. P. M. Krouwel (2017). "Genuine effects of vote advice applications on party choice: Filtering out factors that affect both the advice obtained and the vote". In: *Party Politics* 0 (0), pp. 1–17. DOI: 10.1177/1354068817713121.

Klüver, H. and J.-J. Spoon (2016). "Who Responds? Voters, Parties and Issue Attention". In: *British Journal of Political Science* 46 (3), pp. 633–654. DOI: 10.1017/S0007123414000313.

Knowles, E. S. and C. A. Condon (1999). "Why people say "yes": A dual-process theory of acquiescence". In: *Journal of Personality and Social Psychology* 77 (2), pp. 379–386. DOI: 10.1037/0022-3514.77.2.379.

König, P. D. and D. Nyhuis (2018). "Assessing the applicability of vote advice applications for estimating party positions". In: *Party Politics*. DOI: 10.1177/1354068818790111.

König, P. D. and T. Waldvogel (2018). "Ni gauche ni droite? Positioning the candidates in the 2017 French presidential election". In: *French Politics*. DOI: 10.1057/s41253-018-0059-8.

Korthals, R. and M. Levels (2016). *Multi-attribute compositional voting advice applications (MacVAAs): A methodology for educating and assisting voters and eliciting their preferences*. Tech. rep. Maastricht, The Netherlands: Maastricht University - Researchcentrum voor Onderwijs en Arbeidsmarkt.

Kosslyn, S. M. (1994). *Elements of Graph Design*. New York, NY: W.H. Freeman.

Krippendorff, K. (2004). *Content Analysis - An Introduction to Its Methodology*. 2nd ed. Thousand Oaks - London - New Delhi: SAGE Publications.

Kroh, M. (2007). "Measuring Left-Right Political Orientation: The Choice of Response Format". In: *The Public Opinion Quarterly* 71.2, pp. 204–220. DOI: 10.1093/poq/nfm009.

Krosnick, J. A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys". In: *Applied Cognitive Psychology* 5 (3), pp. 213–236. DOI: 10.1002/acp.2350050305.

Krosnick, J. A. and R. E. Petty (1995). "Attitude Strength: An Overview". In: *Attitude Strength: Antecedents and Consequences*. Ed. by R. E. Petty and J. A. Krosnick. Mahwah, NJ: Erlbaum, pp. 1–24.

Krosnick, J. A. and S. Presser (2010). "Question and Questionnaire Design". In: *Handbook of Survey Research*. Ed. by P. V. Marsden and J. D. Wright. 2nd ed. Bingley, United Kingdom: Emerald, pp. 263–313.

Krosnick, J. A. and H. Schuman (1988). "Attitude intensity, importance, and certainty and susceptibility to response effects". In: *Journal of Personality and Social Psychology* 54.6, pp. 940–952. DOI: 10.1037/0022-3514.54.6.940.

Krouwel, A. and A. van Elfrinkhof (2014). "Combining strengths of methods of party positioning to counter their weaknesses: The development of a new methodology to calibrate parties on issues and ideological dimensions". In: *Quality & Quantity* 48.3, pp. 1455–1472. DOI: 10.1007/s11135-013-9846-0.

Krouwel, A. and J. van de Pol (2014). "Stemhulpen: ter lering ende vermaak". In: *Politieke Partijen: Overbodig of Nodig?* Ed. by S. L. de Lange, M. Leyenaar, and P. de Jong, pp. 151–162.

Krouwel, A., T. Vitiello, and M. Wall (2012). "The practicalities of issuing vote advice: a new methodology for profiling and matching". In: *International Journal of Electronic Governance* 5.3/4, pp. 223–243. DOI: 10.1504/IJEG.2012.051308.

Laaksonen, S.-M., M. Nelimarkka, and J. Haapoja (2016). "Telling citizens how to vote: voting advice applications as a boundary object for political influence and discussion". In: *IPP 2016 Conference Online Proceedings*. University of Oxford, Oxford, United Kingdom.

Ladner, A., G. Felder, and J. Fivaz (2010). "More than toys? A first assessment of voting advice applications in Switzerland". In: *Voting Advice Applications in Europe - The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napoli, Italy: ScriptaWeb, pp. 91–123.

Ladner, A., J. Fivaz, and J. Pianzola (2012). "Voting Advice Applications and Party Choice: Evidence From Smartvote Users in Switzerland". In: *International Journal of Electronic Governance* 5 (3/4), pp. 367–387. DOI: 10.1504/IJEG.2012.051303.

Laver, M. (2001). "How Should We Estimate the Policy Positions of Political Actors?" In: *Estimating the Policy Positions of Political Actors*. Ed. by M. Laver. London, United Kingdom and New York, NY: Routledge, pp. 239–244.

Laver, M., K. Benoit, and J. Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data". In: *The American Political Science Review* 97.2, pp. 311–331. DOI: 10.1017/S0003055403000698.

Laver, M. and J. Garry (2000). "Estimating Policy Positions from Political Texts". In: *American Journal of Political Science* 44.3, pp. 619–634. DOI: 10.2307/2669268.

Lavine, H., J. W. Huff, S. H. Wagner, and D. Sweeney (1998). "The Moderating Influence of Attitude Strength on the Susceptibility to Context Effects in Attitude Surveys". In: *Journal of Personality and Social Psychology* 75.2, pp. 359–373. DOI: 10.1037/0022-3514.75.2.359.

Lay, D. C., S. R. Lay, and J. J. McDonald (2016). *Linear Algebra and Its Applications*. 5th ed. London, United Kingdom: Pearson.

Le Roux, B. and H. Rouanet (2004). *Geometric Data Analysis - From Correspondence Analysis to Structured Data Analysis*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

— (2010). *Multiple Correspondence Analysis*. Quantitative Applications in the Social Sciences 07-163. Los Angeles/London/New Delhi/Singapore/Washington DC: SAGE Publications.

Lê, S., J. Josse, and F. Husson (2008). "FactoMineR: An R Package for Multivariate Analysis". In: *Journal of Statistical Software* 25.1, pp. 1–18. DOI: 10.18637/jss.v025.i01. URL: https://www.jstatsoft.org/v025/i01.

Lee, S., S. H. Kim, Y. H. Hung, H. Lam, Y. A. Kang, and J. S. Yi (2016). "How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novicesś Information Visualization Sensemaking". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1, pp. 499–508. DOI: 10.1109/TVCG.2015.2467195.

Leeper, T. J. (2014). "Cognitive Style and the Survey Response". In: *Public Opinion Quarterly* 78.4, pp. 974–983. DOI: 10.1093/poq/nfu042.

— (2018a). *Interpreting Regression Results using Average Marginal Effects with R's margins*. URL: https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf.

— (2018b). *margins: Marginal Effects for Model Objects*. R package version 0.3.23. URL: https://CRAN.R-project.org/package=margins.

Leeuw, J. de (2006). "Nonlinear Principal Component Analysis and Related Techniques". In: *Multiple Correspondence Analysis and Related Methods*. Ed. by M. Greenacre and J. Blasius. Boca Raton, FL: Chapman & Hall/CRC, pp. 107–133.

Leeuw, J. de and P. Mair (2009). "Gifi Methods for Optimal Scaling in R: The Package homals". In: *Journal of Statistical Software* 31 (4), pp. 1–20. DOI: 10.18637/jss.v031.i04.

Lefevere, J. and S. Walgrave (2014). "A perfect match? The impact of statement selection on voting advice applications' ability to match voters and parties". In: *Electoral Studies* 36, pp. 252–262. DOI: 10.1016/j.electstud.2014.04.002.

Levy, E., J. Zacks, B. Tversky, and D. Schiano (1996). "Gratuitous Graphics? Putting Preferences in Perspective". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '96. New York, NY: ACM, pp. 42–49. DOI: 10.1145/238386.238400.

Liao, D.-c. and B. Chen (2016). "Strengthening Democracy: Development of the iVoter Website in Taiwan". In: *Political Behavior and Technology - Voting Advice Applications in East Asia*. Ed. by D.-c. Liao, B. Chen, and M. J. Jensen. Houndmills, Basingstoke/Hampshire, NY: Palgrave Macmillan, pp. 67–89.

Linden, C. van der and Y. Dufresne (2017). "The curse of dimensionality in Voting Advice Applications: reliability and validity in algorithm design". In: *Journal of Elections, Public Opinion and Parties* 27.1, pp. 9–30. DOI: 10.1080/17457289.2016.1268144.

Linden, W. J. van der and R. K. Hambleton (1997). "Item Response Theory: Brief History, Common Models, and Extensions". In: *Handbook of Modern Item Response Theory*. Ed. by W. J. van der Linden and R. K. Hambleton. New York, NY: Springer-Verlag, pp. 1–28.

Linstone, H. A. and M. Turoff (1975). *The Delphi Method - Techniques and Applications*. Reading, MA: Addison-Wesley.

Linting, M. (2007). "Nonparametric Inference inNonlinear Principal Components Analysis: Exploration and Beyond". PhD thesis. Universiteit Leiden.

Linting, M., J. J. Meulman, P. J. F. Groenen, and A. J. van der Kooij (2007). "Nonlinear Principal Components Analysis: Introduction and Application". In: *Psychological Methods* 12.3, pp. 336–358. DOI: 10.1037/1082-989X.12.3.336.

Liu, M. and F. Keusch (2017). "Effects of Scale Direction on Response Style of Ordinal Rating Scales". In: *Journal of Official Statistics* 33.1, pp. 137–154. DOI: 10.1515/JOS-2017-0008.

Loevinger, J. (1947). "A systematic approach to the construction and evaluation of tests of ability". In: *Psychological Monographs* 61 (4), pp. 1–49. DOI: 10.1037/h0093565.

— (1948). "The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis." In: *Psychological Bulletin* 45 (6), pp. 507–529. DOI: 10.1037/h0055827.

Louwerse, T. and S. Otjes (2012). "Design challenges in cross-national VAAs: the case of the EU Profiler". In: *International Journal of Electronic Governance* 5.3/4, pp. 279–297. DOI: 10.1504/IJEG.2012.051305.

Louwerse, T. and M. Rosema (2014). "The design effects of voting advice applications: Comparing methods of calculating results". In: *Acta Politica* 49 (3), pp. 286–312. DOI: 10.1057/ap.2013.30.

Lupia, A. (1994). "Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections". In: *The American Political Science Review* 88 (1), pp. 63–76. DOI: 10.2307/2944882.

Mahéo, V.-A. (2016). "The Impact of Voting Advice Applications on Electoral Preferences: A Field Experiment in the 2014 Quebec Election". In: *Policy & Internet* 8 (4), pp. 391–411. DOI: 10.1002/poi3.138.

Mair, P. and J. de Leeuw (2017). *Gifi: Multivariate Analysis with Optimal Scaling*. R package version 0.3-8. URL: https://CRAN.R-project.org/package=Gifi.

Mair, P. (2001). "Searching for the Positions of Political Actors: A Review of Approaches and a Critical Evaluation of Expert Surveys". In: *Estimating the Policy Positions of Political Actors*. Ed. by M. Laver. London, United Kingdom and New York, NY: Routledge, pp. 10–30.

Marin, G., R. J. Gamba, and B. V. Marin (1992). "Extreme Response Style and Acquiescence among Hispanics". In: *Journal of Cross-Cultural Psychology* 23.4, pp. 498–509. DOI: 10.1177/0022022192234006.

Marks, G., L. Hooghe, M. R. Steenbergen, and R. Bakker (2007). "Crossvalidating data on party positioning on European integration". In: *Electoral Studies* 26 (1), pp. 23–38. DOI: 10.1016/j.electstud.2006.03.007.

Marschall, S. (2005). "Idee und Wirkung des Wahl-O-Mat". In: *Aus Politik und Zeitgeschichte* 51-52, pp. 41–46.

— (2014). "Profiling Users". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 93–104.

Marschall, S. and C. K. Schmidt (2010). "The Impact of Voting Indicators: The Case of the German Wahl-O-Mat". In: *Voting Advice Applications: The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napoli, Italy: ScriptaWeb, pp. 65–90.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: L. Erlbaum Associates.

McMahon, R., M. Stauffacher, and R. Knutti (Nov. 2015). "The unseen uncertainties in climate change: reviewing comprehension of an IPCC scenario graph". In: *Climatic Change* 133.2, pp. 141–154. DOI: 10.1007/s10584-015-1473-4.

Meijer, R. R. (1994). "The number of Guttman errors as a simple and powerful person-fit statistic". In: *Applied Psychological Measurement* 18.4, pp. 311–314. DOI: 10.1177/014662169401800402.

Mendez, F. (2012). "Matching voters with political parties and candidates: an empirical test of four algorithms". In: *International Journal of Electronic Governance* 5.3/4, pp. 264–278. DOI: 10.1504/IJEG.2012.051316.

— (2014a). "Modelling proximity and directional logic in VAAs". In: *ECPR General Conference*. Glasgow, United Kingdom.

— (2014b). "What's behind a matching algorithm? A critical assessment of how VAAs produce voting recommendations". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 49–66.

— (2017). "Modeling proximity and directional decisional logic: What can we learn from applying statistical learning techniques to VAA-generated data?" In: *Journal of Elections, Public Opinion and Parties* 27.1, pp. 31–55. DOI: 10.1080/17457289.2016.1269113. eprint: http://dx.doi.org/10.1080/17457289.2016.1269113.

Mendez, F., K. Gemenis, and C. Djouvas (2014). "Methodological Challenges in the Analysis of Voting Advice Application Generated Data". In: *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, pp. 142–148. DOI: 10.1109/SMAP.2014.32.

Mendez, F. and V. Manavopoulos (2018). *EUvox2014: Voting Advice Application data for the 2014 European Parliament elections*. Licences: CC BY-NC 4.0. GESIS Datorium.

Mendez, F. and J. Wheatley (2014). "Using VAA-generated data for mapping partisan supporters in the ideological space". In: *Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective*. Ed. by D. Garzia and S. Marschall. Colchester, United Kingdom: ECPR Press, pp. 161–173.

Meulman, J. J. (1986). *A Distance Approach to Nonlinear Multivariate Analysis*. Leiden: DSWO Press.

— (1998). "Optimal Scaling Methods for Graphical Display of Multivariate Data". In: *COMPSTAT*. Ed. by R. Payne and P. Green. Heidelberg, Germany: Physica-Verlag HD, pp. 65–76.

Meulman, J. J., W. J. Heiser, and S. P. S. S. Inc. (2004). *SPSS Categories 13.0*. Chicago, IL: SPSS Inc.

Meulman, J. J., A. J. van der Kooij, and W. J. Heiser (2004). "Principal Components Analysis With Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data". In: *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Ed. by D. Kaplan. Thousand Oaks, CL: SAGE Publications, pp. 49–70.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis - With Applications in Political Research*. Den Haag/Paris: Mouton.

Mokken, R. J. and C. Lewis (1982). "A Nonparametric Approach to the Analysis of Dichotomous Item Responses". In: *Applied Psychological Measurement* 6.4, pp. 417–430. DOI: 10.1177/014662168200600404.

Mokken, R. J., C. Lewis, and K. Sijtsma (1986). "Rejoinder to "The Mokken Scale: A Critical Discussion"". In: *Applied Psychological Measurement* 10 (3), pp. 279–285. DOI: 10.1177/014662168601000306.

Mölder, M. (2016). "The validity of the RILE left–right index as a measure of party policy". In: *Party Politics* 22.1, pp. 37–48. DOI: 10.1177/1354068813509525.

Molenaar, I. W. (1991). "A Weighted Loevinger H Coefficient Extending Mokken Scaling to Multicategory Items". In: *Kwantitatieve Methoden* 12 (37), pp. 97–117.

Molenaar, I. W. (1997). "Nonparametric Models for Polytomous Responses". In: *Handbook of Modern Item Response Theory*. Ed. by W. J. van der Linden and R. K. Hambleton. New York, NY: Springer-Verlag, pp. 367–380.

Molenaar, I. W. and K. Sijtsma (1988). "Mokken's approach to reliability estimation extended to multicategory items". In: *Kwantitatieve Methoden* 9.28, pp. 115–126.

— (2000). *MSP5 for Windows User's Manual*. Groningen, The Netherlands: Iec ProGAMMA.

Moors, G. (2008). "Exploring the effect of a middle response category on response style in attitude measurement". In: *Quality & Quantity* 42 (6), pp. 779–794. DOI: 10.1007/s11135-006-9067-x.

Mori, Y., M. Kuroda, and N. Makino (2016). *Nonlinear Principal Component Analysis and Its Applications*. JSS Research Series in Statistics. Singapore: Springer.

Mueller, K., T. Straatmann, K. Hattrup, and M. Jochum (2014). "Effects of Personalized Versus Generic Implementation of an Intra-Organizational Online Survey on Psychological Anonymity and Response Behavior: A Field Experiment". In: *Journal of Business and Psychology* 29.2, pp. 169–181. DOI: 10.1007/s10869-012-9262-9.

Munzner, T. (2015). *Visualization Analysis & Design*. Boca Raton, FI: CRC Press.

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". In: *The Journal of General Psychology* 142.2, pp. 71–89. DOI: 10.1080/00221309.2014.994590.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto, Canada: University of Toronto Press.

— (2007). *Multidimensional Nonlinear Descriptive Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Norris, P., ed. (1999). *Critical Citizens: Global Support for Democratic Government*. Oxford: Oxford University Press.

Nunnally, J. C. (1967). *Psychometric Theory*. New York, NY: McGraw Hill.

Nuytemans, M., S. Walgrave, and K. Deschouwer (2010). "Do the Vote Test: The Belgian Voting Aid Application". In: *Voting Advice Applications: The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napoli, Italy: ScriptaWeb, pp. 125–142.

Oppenheim, A. N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*. New Edition. London, United Kingdom/New York, NY: Continuum.

Otjes, S. and T. Louwerse (2014). "Spatial Models in Voting Advice Applications". In: *Electoral Studies* 36, pp. 263–271. DOI: 10.1016/j.electstud.2014.04.004.

Partij van de Arbeid (2016). *Een Verbonden Samenleving - Verkiezingsprogramma 2017*. Den Haag: Partij van de Arbeid.

Partij voor de Dieren (2016). *Plan B - Verkiezingsprogramma Partij voor de Dieren Tweede Kamerverkiezingen 2017*. Amsterdam: Partij voor de Dieren.

Partij voor de Vrijheid (2016). *Verkiezingsprogramma PVV 2017-2021*. Den Haag: Partij voor de Vrijheid.

Pasek, J. and J. A. Krosnick (2010). "Optimizing Survey Questionnaire Design in Political Science: Insights from Psychology". In: *The Oxford Handbook of American Elections and Political Behavior*. Ed. by J. E. Leighley. Oxford, United Kingdom: Oxford University Press, pp. 27–50.

Payné, S. L. (1951). *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.

Peebles, D. and P. C.-H. Cheng (2003). "Modeling the Effect of Task and Graphical Representation on Response Latency in a Graph Reading Task". In: *Human Factors* 45.1, pp. 28–46. DOI: 10.1518/hfes.45.1.28.27225.

Petty, R. E. and J. A. Krosnick (1995). *Attitude Strength: Antecedents and Consequences*. Mahwah, NJ: Erlbaum.

Pinker, S. (1990). "A Theory of Graph Comprehension". In: *Artificial intelligence and the future of testing*. Ed. by R. Freedle. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 73–126.

Pol, J. van de, B. Holleman, N. Kamoen, A. Krouwel, and C. de Vreese (2014). "Beyond Young, Highly Educated Males: A Typology of VAA Users". In: *Journal of Information Technology & Politics* 11 (4), pp. 397–411. DOI: 10.1080/19331681.2014.958794.

Polk, J., J. Rovny, R. Bakker, E. Edwards, L. Hooghe, S. Jolly, J. Koedam, F. Kostelka, G. Marks, G. Schumacher, M. Steenbergen, M. Vachudova, and M. Zilovic (2017). "Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data". In: *Research & Politics* 4.1, pp. 1–9. DOI: 10.1177/2053168016686915.

Popkin, S. (1991). *The Reasoning Voter*. Chicago, IL: The University of Chicago Press.

Power, T. J. and C. Zucco Jr. (2009). "Estimating Ideology of Brazilian Legislative Parties, 1990–2005: A Research Communication". In: *Latin American Research Review* 44.1, pp. 218–246. DOI: 10.1353/lar.0.0072.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rabinowitz, G. and S. E. MacDonald (1989). "A Directional Theory of Issue Voting". In: *The American Political Science Review* 83.1, pp. 93–121. DOI: 10.2307/1956436.

Ramonaitė, A. (2010). "Voting Advice Applications in Lithuania: Promoting Programmatic Competition or Breeding Populism?" In: *Policy and Internet* 2.1, pp. 117–147. DOI: 10.2202/1944-2866.1027.

Reuband, K.-H. (2003). "The Allow-Forbid Asymmetry in Question Wording - a New Look at an Old Problem". In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 80.1, pp. 25–35. DOI: 10.1177/075910630308000104.

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

Robbins, N. B. (2005). *Creating More Effective Graphs*. Hoboken, NJ: John Wiley & Sons.

Rose, R. and I. McAllister (1986). *Voters begin to choose: From closed class to open elections in Britain*. London: Sage Publications.

Rosema, M. and T. Louwerse (2016). "Response Scales in Voting Advice Applications: Do Different Designs Produce Different Outcomes?" In: *Policy & Internet* 8 (4), pp. 431–456. DOI: 10.1002/poi3.139.

Rovny, J. (2012). "Who emphasizes and who blurs? Party strategies in multidimensional competition". In: *European Union Politics* 13.2, pp. 269–292. DOI: 10.1177/1465116511435822.

Schaeffer, N. C. (1991). "Hardly Ever or Constantly? Group Comparisons Using Vague Quantifiers". In: *Public Opinion Quarterly* 55.3, pp. 395–423. DOI: 10.1086/269270. eprint: /oup/backfile/content_public/journal/poq/55/3/10.1086/269270/2/55-3-395.pdf.

Schriesheim, C. A., R. J. Eisenbach, and K. D. Hill (1991). "The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation". In: *Educational and Psychological Measurement* 51 (1), pp. 67–78. DOI: 10.1177/0013164491511005.

Schuldt, J. P., S. H. Konrath, and N. Schwarz (2011). ""Global warming" or "climate change"? Whether the planet is warming depends on question wording". In: *Public Opinion Quarterly* 75 (1), pp. 115–124. DOI: 10.1093/poq/nfq073.

Schultze, M. (2014). "Effects of Voting Advice Applications (VAAs) on political knowledge about party positions". In: *Policy and Internet* 6 (1), pp. 46–68. DOI: 10.1002/1944-2866.POI352.

Schuman, H. and S. Presser (1981). *Questions and Answers in Attitude Surveys - Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage Publications.

Schuszler, P., J. de Graaf, and P. Lucardie (2003a). "Reactie: Misverstanden blijven groot". In: *B en M: tijdschrift voor beleid, politiek en maatschappij* 30.3, p. 204.

— (2003b). "Zin en onzin over de StemWijzer 2002: een reactie". In: *B en M: tijdschrift voor beleid, politiek en maatschappij* 30.3, pp. 194–204.

Schuur, W. H. van (2003). "Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory". In: *Political Analysis* 11.2, pp. 139–163. DOI: 10.1093/pan/mpg002.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

Shah, P. (2002). "Graph Comprehension: The Role of Format, Content and Individual Differences". In: *Diagrammatic Representation and Reasoning*. Ed. by M. Anderson, B. Meyer, and P. Olivier. London, United Kingdom: Springer-Verlag London, pp. 173–185.

Shah, P. and P. A. Carpenter (1995). "Conceptual limitations in comprehending line graphs". In: *Journal of Experimental Psychology: General* 124 (1), pp. 43–61. DOI: 10.1037/0096-3445.124.1.43.

Shah, P. and E. G. Freedman (2011). "Bar and Line Graph Comprehension: An Interaction of Top-Down and Bottom-Up Processes". In: *Topics in Cognitive Science* 3.3, pp. 560–578. DOI: 10.1111/j.1756-8765.2009.01066.x.

Shah, P., E. G. Freedman, and I. Vekiri (2005). "The Comprehension of Quantitative Information in Graphical Displays". In: *The Cambridge Handbook of Visuospatial Thinking*. Ed. by P. Shah and A. Miyake. Cambridge Handbooks in Psychology. Cambridge, United Kingdom: Cambridge University Press, pp. 426–476.

Shah, P. and J. Hoeffner (2002). "Review of Graph Comprehension Research: Implications for Instruction". In: *Educational Psychology Review* 14.1, pp. 47–69. DOI: 10.1023/A:1013180410169.

Shamim, A., V. Balakrishnan, and M. Tahir (2015). "Evaluation of opinion visualization techniques". In: *Information Visualization* 14.4, pp. 339–358. DOI: 10.1177/1473871614550537.

Shikano, S. (2013). "Estimating ideological positions of political parties using a deliberative expert survey". In: *Voting Advice Applications Workshop, University of Twente*.

Siegel, S. and N. J. Castellan Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York, NY: McGraw-Hill.

Sijtsma, K. (2009). "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha". In: *Psychometrika* 74 (1), pp. 107–120. DOI: 10.1007/s11336-008-9101-0.

Sijtsma, K. and I. W. Molenaar (2002). *Introduction to nonparametric item response modeling*. Vol. 5. Measurement methods for the social sciences. Thousand Oaks, CA: Sage Publications.

Simon, H. A. (1957). *Models of Man*. New York, NY: Wiley.

Skocpol, T. and M. P. Fiorina, eds. (1999). *Civic Engagement in American Democracy*. Washington, DC: Brooking.

Škop, M. (2010). "Are the Voting Advice Applications (VAAs) Telling the Truth? Measuring VAAs' Quality. Case Study from the Czech Republic". In: *Voting Advice Applications: The State of the Art*. Ed. by L. Cedroni and D. Garzia. Napoli, Italy: ScriptaWeb, pp. 199–216.

Slapin, J. B. and S.-O. Proksch (2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts". In: *American Journal of Political Science* 52 (3), pp. 705–722. DOI: 10.1111/j.1540-5907.2008.00338.x.

Smits, I. A. M., M. E. Timmerman, and R. R. Meijer (2012). "Exploratory Mokken Scale Analysis as a Dimensionality Assessment Tool: Why Scalability Does Not Imply Unidimensionality". In: *Applied Psychological Measurement* 36.6, pp. 516–539. DOI: 10.1177/0146621612451050.

Socialistische Partij (2016). *PakDeMacht - Programma voor een Sociaal Nederland voor de Verkiezingen van 15 maart 2017*. Amersfoort: Socialistische Partij.

Solt, F. and Y. Hu (2018). *interplot: Plot the Effects of Variables in Interaction Terms*. R package version 0.2.1. URL: https://CRAN.R-project.org/package=interplot.

Steenbergen, M. R. and G. Marks (2007). "Evaluating expert judgments". In: *European Journal of Political Research* 46 (3), pp. 347–366. DOI: 10.1111/j.1475-6765.2006.00694.x.

Stochl, J., P. B. Jones, and T. J. Croudace (2012). "Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers". In: *BMC Medical Research Methodology* 12.1, pp. 74–90. DOI: 10.1186/1471-2288-12-74.

Stoet, G. (2010). "PsyToolkit: A software package for programming psychological experiments using Linux". In: *Behavior Research Methods* 42.4, pp. 1096–1104. DOI: 10.3758/BRM.42.4.1096.

— (2017). "PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments". In: *Teaching of Psychology* 44.1, pp. 24–31. DOI: 10.1177/0098628316677643.

Straat, J. H., L. A. van der Ark, and K. Sijtsma (2013). "Comparing Optimization Algorithms for Item Selection in Mokken Scale Analysis". In: *Journal of Classification* 30.1, pp. 75–99. DOI: 10.1007/s00357-013-9122-y.

Swain, S. D., D. Weathers, and R. W. Niedrich (Feb. 2008). "Assessing Three Sources of Misresponse to Reversed Likert Items". In: *Journal of Marketing Research* 45.1, pp. 116–131. DOI: 10.1509/jmkr.45.1.116.

Thiessen, V. and J. Blasius (2008). "Mathematics Achievement and Mathematics Learning Strategies: Cognitive Competencies and Construct Differentiation". In: *International Journal of Educational Research* 47.6, pp. 362–371. DOI: 10.1016/j.ijer.2008.12.002.

Thomassen, J. (2015). "What's gone wrong with democracy, or with theories explaining why it has?" In: *Citizenship and democracy in an era of crisis*. Ed. by T. Poguntke, S. Rossteutscher, R. Schmitt-Beck, and S. Zmerli. London: Routledge.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *The Psychology of Survey Response*. Cambridge, United Kingdom: Cambridge University Press.

Trechsel, A. H., D. Garzia, and L. De Sio (2014). *euandi (Expert Interviews)*. Data file. DOI: 10.4232/1.12138.

— (2015). *euandi (General Population Survey) - User Profiles in the 2014 European Elections*. Data file. DOI: 10.4232/1.12246.

Trechsel, A. H. and P. Mair (2011). "When Parties (Also) Position Themselves: An Introduction to the EU Profiler". In: *Journal of Information Technology & Politics* 8.1, pp. 1–20. DOI: 10.1080/19331681.2011.533533.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

— (2001). *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Tversky, A. and D. Kahneman (1981). "The framing of decisions and the psychology of choice". In: *Science* 211 (4481), pp. 453–458. DOI: 10.1126/science.7455683.

Van Vaerenbergh, Y. and T. D. Thomas (2013). "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies". In: *International Journal of Public Opinion Research* 25.2, pp. 195–217. DOI: 10.1093/ijpor/eds021.

Volkspartij voor Vrijheid en Democratie (2016). *Zeker Nederland - VVD verkiezingsprogramma 2017-2021*. Den Haag: Volkspartij voor Vrijheid en Democratie.

Walgrave, S., P. van Aelst, and M. Nuytemans (2008a). "'Do the Vote Test': The Electoral Effects of a Popular Vote Advice Application at the 2004 Belgian Elections". In: *Acta Politica* 43.1, pp. 50–70. DOI: 10.1057/palgrave.ap.5500209.

— (2008b). "Vote Advice Applications as New Campaign Players? The Electoral Effects of the Do the Vote Test during the 2004 Regional Elections in Belgium". In: *Non-party actors in electoral politics - The role of interest groups and independent citizens in contemporary election campaigns*. Ed. by D. M. Farrell and R. Schmitt-Beck. Baden-Baden, Germany: Nomos, pp. 237–258.

Walgrave, S., M. Nuytemans, and K. Pepermans (2009). "Voting Aid Applications and the Effect of Statement Selection". In: *West European Politics* 32.6, pp. 1161–1180. DOI: 10.1080/01402380903230637.

Wall, M., M. L. Sudulich, R. Costello, and E. Leon (2009). "Picking your party online-An investigation of Irelands first online voting advice application". In: *Information Polity* 14.3, pp. 203–218. DOI: 10.3233/IP-2009-0179.

Weech-Maldonado, R., M. N. Elliott, A. Oluwole, K. C. Schiller, and R. D. Hays (2008). "Survey Response Style and Differential Use of CAHPS Rating Scales by Hispanics". In: *Medical Care* 46 (9), pp. 963–968. DOI: 10.1097/MLR.0b013e3181791924.

Wheatley, J. (2015a). "Identifying Latent Policy Dimensions from Public Opinion Data: An Inductive Approach". In: *Journal of Elections, Public Opinion and Parties* 25 (2), pp. 215–233. DOI: 10.1080/17457289.2014.985222.

— (2015b). "Restructuring the policy space in England: The end of the Left–Right paradigm?" In: *British Politics* 10 (3), pp. 268–285. DOI: 10.1057/bp.2015.35.

— (2016). "Cleavage Structures and Dimensions of Ideology in English Politics: Evidence From Voting Advice Application Data". In: *Policy & Internet* 8 (4), pp. 457–477. DOI: 10.1002/poi3.129.

Wheatley, J., C. Carman, F. Mendez, and J. Mitchell (2014). "The dimensionality of the Scottish political space: Results from an experiment on the 2011 Holyrood elections". In: *Party Politics* 20.6, pp. 864–878. DOI: 10.1177/1354068812458614.

Wheatley, J. and F. Mendez (2018). "Reconceptualising dimensions of political competition in Europe: A demand side approach". In: *British Journal of Political Science*. Forthcoming.

Widaman, K. F. (2007). "Common Factors Versus Components: Principals and Principles, Errors and Misconceptions. Historical Developments and Future Directions". In: *Factor Analysis at 100*. Ed. by R. Cudeck and R. C. MacCallum. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 177–204.

Wilson, T. D. and S. D. Hodges (1992). "Attitudes as temporary constructions". In: *The construction of social judgments*. Ed. by L. L. Martin and A. Tesser. Hillsdale, NJ: Lawrence Erlbaum, pp. 37–65.

Windschitl, P. D. and G. L. Wells (1996). "Measuring psychological uncertainty: Verbal versus numeric methods". In: *Journal of Experimental Psychology: Applied* 2 (4), pp. 343–364. DOI: 10.1037/1076-898X.2.4.343.

Wismeijer, A. A. J., K. Sijtsma, M. A. L. M. van Assen, and A. J. J. M. Vingerhoets (2008). "A Comparative Study of the Dimensionality of the Self-Concealment Scale Using Principal Components Analysis and Mokken Scale Analysis". In: *Journal of Personality Assessment* 90.4, pp. 323–334. DOI: 10.1080/00223890802107875.

Yan, T. and F. Keusch (2015). "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey". In: *Public Opinion Quarterly* 79.1, pp. 145–165. DOI: 10.1093/poq/nfu062.

Yong, A. G. and S. Pearce (2013). "A Beginners Guide to Factor Analysis: Focusing on Exploratory Factor Analysis". In: *Tutorials in Quantitative Methods for Psychology* 9.2, pp. 79–94. DOI: 10.20982/tqmp.09.2.p079.

Young, I. (1990). *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press.

Yu, S.-C. and W.-H. Hsu (2013). "Applying structural equation modeling methodology to test validation: an example of cyberspace positive psychology scale". In: *Quality & Quantity* 47 (6), pp. 3423–3434. DOI: 10.1007/s11135-012-9730-3.

Yu, S.-C., Y.-H. Lin, and W.-H. Hsu (2013). "Applying structural equation modeling to report psychometric properties of Chinese version 10-item CES-D depression scale". In: *Quality & Quantity* 47 (3), pp. 1511–1518. DOI: 10.1007/s11135-011-9604-0.

Zacks, J., E. Levy, B. Tversky, and D. J. Schiano (1998). "Reading bar graphs: Effects of extraneous depth cues and graphical context". In: *Journal of Experimental Psychology: Applied* 4 (2), pp. 119–138. DOI: 10.1037/1076-898X.4.2.119.

Ziemkiewicz, C., A. Ottley, R. J. Crouser, K. Chauncey, S. L. Su, and R. Chang (Nov. 2012). "Understanding Visualization by Understanding Individual Users". In: *IEEE Computer Graphics and Applications* 32.6, pp. 88–94. DOI: 10.1109/MCG.2012.120.

Zmerli, S. and T. W. G. van der Meer (2017). *Handbook on Political Trust*. Cheltenham, UK: Edward Elgar.

A | Inventory Codebook

Variable	Obs	Miss	Mean	Min	Max	Label	Categories
gen_country_first	1102	0	23	1	50	Country (1st Level)	-
gen_country_second	1102	399	-	-	-	Country (2nd Level)	-
gen_name	1102	0	-	-	-	Name	-
gen_year	1102	0	2011	1989	2017	Year	-
gen_weights	1102	-	.29	$3.2 \cdot 10^{-}3$	1	Weights	-
gen_election	1102	0	3.6	1.0	7	Election Type	1 (EP) 2 (Lower House) 3 (Upper House) 4 (Municipality) 5 (Region) 6 (President) 7 (Other)
ques_respositions	1102	2	3-9	2	50	# Response Options	-
ques_numitems	1102	84	29	9	88	# Items	-
ques_dnk	1102	0	0.6	0	1	"Do Not Know" Included	0 (No) 1 (Yes)
ques_weights	1102	0	0.5	0	1	Weights Included	0 (No) 1 (Yes)

(continues on the next page)

Table A.1 continued from previous page

Variable	Obs	Miss	Mean	Min	Max	Label	Categories
type	1102	0	1.5	1	2	VAA Format	1 (Candidate) 2 (Party)
format	1102	0	1.6	1	2	Response Format	1 (Options) 2 (Scale)
pos_method	1102	0	3.1	1	6	Positioning Method	1 (Manifesto) 2 (Candidate) 3 (Elite) 4 (Expert) 5 (Kieskompas) 6 (Delphi)
match_cityblock	1102	0	0.1	0	1	City-Block Matching	0 (No) 1 (Yes)
match_euclidean	1102	0	0.1	0	1	Euclidean Matching	0 (No) 1 (Yes)
match_agreement	1102	0	0.7	0	1	Agreement Matching	0 (No) 1 (Yes)
match_2d	1102	0	0.2	0	1	2D Matching	0 (No) 1 (Yes)
match_3d	1102	0	0	0	1	3D Matching	0 (No)

(continues on the next page)

Table A.1 continued from previous page

Variable	Obs	Miss	Mean	Min	Max	Label	Categories
							1 (Yes)
match_6d	1102	0	0	0	1	6D Matching	0 (No)
							1 (Yes)
vis_single	1102	0	0	0	1	Single Party Visual	0 (No)
							1 (Yes)
vis_rank_quant	1102	0	0.9	0	1	Quant Rank Visual	0 (No)
							1 (Yes)
vis_rank_notquant	1102	0	0	0	1	Non-Quant Rank Visual	0 (No)
							1 (Yes)
vis_twodim	1102	0	0.2	0	1	2D Visual	0 (No)
							1 (Yes)
vis_threedim	1102	0	0	0	1	3D Visual	0 (No)
							1 (Yes)
vis_spider	1102	0	0	0	1	Spider Graph	0 (No)
							1 (Yes)
designer_uni	1102	0	0.3	0	1	Designer - University	0 (No)
							1 (Yes)
designer_comp	1102	0	0.2	0	1	Designer - Company	0 (No)

(continues on the next page)

Table A.1 continued from previous page

Variable	Obs	Miss	Mean	Min	Max	Label	Categories
							1 (Yes)
designer_media	1102	0	0.4	0	1	Designer - Media	0 (No)
							1 (Yes)
designer_ngo	1102	0	0.4	0	1	Designer - NGO	0 (No)
							1 (Yes)
elec_vaa	1102	213	52.5	2	2901	# Party/Candidate (VAA)	-
elec_ballot	1102	889	129.3	2	3473	# Party/Candidate (Ballot)	-
elec_parliament	1102	1051	27.4	3	85	# Party/Candidate (Parl.)	-
elec_electors	1102	917	$5.2 \cdot 10^6$	1572	$1.3 \cdot 10^8$	# Electors	-
elec_users	1102	1014	$9.7 \cdot 10^5$	658	$1.3 \cdot 10^7$	# VAA Users	-
website	1102	94	-	-	-	Website	-
Note	1102	1067	-	-	-	Notes	-

B | Stemconsult Data

Stemconsult Questionnaire

	Version A		Version B			
#	Question	Type	Question	Type	Exp/Imp	Scale
1	Op kunst en cultuur mag niet worden bezuinigd (*Subsidies for arts and culture may not be cut*)	-	Op kunst en cultuur mag worden bezuinigd (*Subsidies for arts and culture may be cut*)	+	EXP	EC
2	De publieke omroep in de huidige vorm moet worden behouden (*The public broadcaster in its current form has to be maintained*)	+	De publieke omroep in de huidige vorm moet worden afgeschaft (*The public broadcaster in its current form has to be abolished*)	-	IMP	EC
3	Nederland moet meer uitgeven aan ontwikkelings samenwerking (*The Netherlands has to spend more on developmental cooperation*)	+	Nederland moet meer uitgeven aan ontwikkelings samenwerking (*The Netherlands has to spend more on developmental cooperation*)	+		SO
4	De AOW leeftijd moet terug naar 65 jaar (*The pension age has to revert to 65 years*)	+	De AOW leeftijd moet terug naar 65 jaar (*The pension age has to revert to 65 years*)	+		EC
5	De hypotheekrenteaftrek moet behouden blijven (*Mortgage relief has to be maintained*)	+	De hypotheekrenteaftrek moet worden afgeschaft (*Mortgage relief has to be abolished*)	-	IMP	EC
6	Anoniem solliciteren moet bij de overheid de norm worden (*Anonymous application must become the norm for government jobs*)	+	Anoniem solliciteren moet bij de overheid de norm worden (*Anonymous application must become the norm for government jobs*)	+		SO

Table B.1 continued from previous page

	Version A		Version B			
#	Question	Type	Question	Type	Exp/Imp	Scale
7	Zelfstandigen zonder personeel (ZZP'ers) moeten verplicht worden zich te verzekeren tegen ziekte en arbeidsongeschiktheid (*Entrepreneurs without personnel should be obliged to insure themselves against illness and disability*)	+	Zelfstandigen zonder personeel (ZZP'ers) moeten verplicht worden zich te verzekeren tegen ziekte en arbeidsongeschiktheid (*Entrepreneurs without personnel should be obliged to insure themselves against illness and disability*)	+		EC
8	Ouderen hebben recht een einde aan hun leven te maken wanneer ze vinden dat het voltooid is (*Elderly people have the right to end their lives when they consider it finished*)	+	Ouderen hebben recht een einde aan hun leven te maken wanneer ze vinden dat het voltooid is (*Elderly people have the right to end their lives when they consider it finished*)	+		SO
9	De Islam vormt een bedreiging voor de Nederlandse normen en waarden (*Islam forms a threat for Dutch norms and values*)	+	De Islam vormt een bedreiging voor de Nederlandse normen en waarden (*Islam forms a threat for Dutch norms and values*)	+		SO
10	Softdrugs moeten verboden worden (*Soft drugs should be forbidden*)	-	Softdrugs moeten gelegaliseerd worden (*Soft drugs should be legalised*)	+	IMP	SO
11	Nederland moet in de Europese Unie blijven (*The Netherlands has to remain in the European Union*)	+	Nederland moet uit de Europese Unie (*The Netherlands has to leave the European Union*)	-	IMP	SO
12	Het is goed als meer landen toetreden tot de Europese Unie (*It is good if more countries join the European Union*)	+	Het is goed als meer landen toetreden tot de Europese Unie (*It is good if more countries join the European Union*)	+		SO

Table B.1 continued from previous page

	Version A		Version B			
#	Question	Type	Question	Type	Exp/Imp	Scale
13	Immigranten moeten zich aanpassen aan de Nederlandse normen en waarden (*Immigrants have to adapt themselves to Dutch norms and values*)	+	Immigranten moeten zich aanpassen aan de Nederlandse normen en waarden (*Immigrants have to adapt themselves to Dutch norms and values*)	+		SO
14	Nederland moet meer vluchtelingen toelaten dan nu het geval is (*The Netherlands has to accept more refugees than is now the case*)	+	Nederland moet meer vluchtelingen toelaten dan nu het geval is (*The Netherlands has to accept more refugees than is now the case*)	+		SO
15	Voor milieumaatregelen mag geen belastingverhoging plaatsvinden(*For environmental measures, taxes may not be raised*)	-	Voor milieumaatregelen mag een belastingverhoging plaatsvinden(*For environmental measures, taxes may be raised*)	+	EXP	EC
16	De overheid moet bedrijven zelf laten bepalen of ze energiebesparende maatregelen nemen (*The government has to leave it to the companies to take energy saving measures*)	+	De overheid mag bedrijven dwingen tot het nemen van energiebesparende maatregelen (*The government may force companies to take environmental measures*)	-	IMP	EC
17	Alle kolencentrales in Nederland moeten dicht (*All coal plants in the Netherlands should be closed*)	-	De kolencentrales in Nederland mogen openblijven (*All coal plants in the Netherlands may remain open*)	+	IMP	SO
18	Het leenstelsel voor studenten moet worden af- geschaft (*The loan system for students should be abolished*)	-	Het leenstelsel voor studenten moet behouden blijven (*The loan system for students should be maintained*)	+	IMP	EC
19	Er moet een bindend referendum worden ingevoerd (*A binding referendum has to be introduced*)	+	Er moet een bindend referendum worden ingevoerd (*A binding referendum has to be introduced*)	+		SO

Table B.1 continued from previous page

#	Version A		Version B		Exp/Imp	Scale
	Question	Type	Question	Type		
20	Er moet extra geld worden geïnvesteerd in Defensie (*Additional funds must be invested in Defense*)	+	Er moet extra geld worden geïnvesteerd in Defensie (*Additional funds must be invested in Defense*)	+		SO
21	Op welzijns werk mag niet worden bezuinigd (*There can be no cuts in spending on social work*)	-	Op welzijns werk mag worden bezuinigd (*There can be cuts in spending on social work*)	+	EXP	EC
22	Het eigen risico in de zorg moet worden afgeschaft (*The own risk in healthcare should be abolished*)	-	Het eigen risico in de zorg moet behouden blijven (*The own risk in healthcare should be maintained*)	+	IMP	EC
23	Door vrije marktwerking functioneert de gezondheidszorg beter (*Through free market operation, healthcare functions better*)	+	Door vrije marktwerking functioneert de gezondheidszorg beter (*Through free market operation, healthcare functions better*)	+		EC
24	Er moet een kilometerheffing komen ter vervanging van de huidige belasting van personenauto's en motorrijwielen (*There must be a mileage charge to replace the current tax on passenger cars and motorcycles*)	+	Er moet een kilometerheffing komen ter vervanging van de huidige belasting van personenauto's en motorrijwielen (*There must be a mileage charge to replace the current tax on passenger cars and motorcycles*)	+		EC
25	De multiculturele samenleving is een goede zaak (*The multicultural society is a good thing*)	+	De multiculturele samenleving is geen goede zaak (*The multicultural society is not a good thing*)	-	EXP	SO

Stem Consult items with English translation italicised in brackets. Type indicates whether or not the question is positive or negative, Exp/Imp whether the negation is implicit or explicit, and Scale to which of the two *ex ante* designated scales the question belongs.

Party Positions for Stem-Consult (Version A)

#	Completely Disagree	Disagree	Neutral	Agree	Completely Agree	NA
1	D66 GL PvdA PvdD SP	50Plus CDA CU SGP VVD FvD	DENK VNL		PVV	
2	VNL	50Plus CDA CU PvdA PvdD VVD DENK		D66 GL SGP SP	PVV FvD	
3	CDA CU D66 GL PvdA PvdD DENK	SP	SGP	50Plus VVD	PVV VNL FvD	
4	50Plus PVV SP	PvdD	PvdA	CU GL SGP DENK VNL FvD	CDA D66 VVD	
5	50Plus	CDA CU D66 PvdA PVV SGP SP VVD DENK	VNL FvD	GL PvdD		
6	GL PvdA	50Plus PvdD	D66	CDA CU PVV SGP SP VVD VNL	DENK	FvD
7	CU GL DENK	50Plus CDA PvdA		D66 PvdD SGP SP VNL	VVD FvD	PVV
8	50Plus D66 GL VVD VNL FvD	PvdA PvdD	PVV	CDA SP DENK	CU SGP	

Table B.2 continued from previous page

#	Completely Disagree	Disagree	Neutral	Agree	Completely Agree	NA
9	PVV VNL	SGP	50Plus CDA VVD FvD	CU D66 GL PvdA PvdD SP	DENK	
10	CU SGP	CDA PVV	VVD	50Plus PvdA DENK VNL	D66 GL PvdD SP FvD	
11	50Plus CDA CU D66 GL PvdA VVD	PvdD SGP SP DENK		VNL	PVV FvD	
12		D66 GL DENK	VVD	CDA CU PvdA SGP SP VNL	50Plus PvdD PVV FvD	
13	PVV SGP VNL	50Plus CDA CU D66 GL PvdD SP VVD FvD	PvdA	DENK		
14		GL DENK	CDA CU D66 PvdA PvdD SP	50Plus SGP VVD	PVV VNL FvD	
15		PVV VVD VNL	CDA	50Plus D66 PvdA SGP DENK	CU GL PvdD SP	FvD
16		CDA D66 PVV SGP VVD FvD	VNL	50Plus CU PvdA SP DENK	GL PvdD	

Table B.2 continued from previous page

#	Completely Disagree	Disagree	Neutral	Agree	Completely Agree	NA
17	CU D66 GL PvdD DENK	50Plus PvdA SP	FvD	CDA PVV SGP VVD VNL		
18	50Plus CDA CU PvdD SP DENK FvD		PVV SGP	D66 GL PvdA VNL	VVD	
19	50Plus PVV VNL FvD	D66 PvdD SP		CU GL PvdA	CDA SGP VVD DENK	
20	CDA CU D66 PVV SGP VVD VNL FvD	50Plus PvdA		GL PvdD SP DENK		
21	CU PVV	50Plus CDA D66 GL PvdA PvdD SGP SP VVD DENK VNL	FvD			
22	50Plus GL PvdA PvdD PVV SP DENK			CDA CU SGP VNL FvD	D66 VVD	
23	D66 VVD	SGP	CDA CU GL PvdD PVV DENK VNL FvD		50Plus PvdA SP	
24	GL PvdA PvdD DENK	D66 SGP SP	50Plus FvD	CDA CU PVV	VVD VNL	

Table B.2 continued from previous page

#	Completely Disagree	Disagree	Neutral	Agree	Completely Agree	NA
25	D66 GL PvdA SP DENK	CDA CU	VVD	SGP VNL FvD	PVV	50Plus PvdD

Model

The linear model used to estimate the marginal and conditional effects is:

$$\begin{aligned} match \sim\ & pos\_neg + age + sex + education + interestInPolitics + gutmann + \\ & pos\_neg * age + pos\_neg * sex + pos\_neg * education + \\ & pos\_neg * interestInPolitics + pos\_neg * gutmann \end{aligned} \tag{B.1}$$

Here, the *match* refers to the percentage match between user and party and can range from −100 to +100. The meaning of this depends on the matching algorithm used. In the case of the Hybrid algorithm, −100 indicates that in all cases where the user completely disagreed/agreed, the party completely agreed/disagreed, while +100 means that in all cases where the user completely agreed/disagreed, the party also completely agreed/disagreed, and 0 means that either the party or the user was neutral while the other completely agreed or disagreed. In the case of the scalar algorithm, −100 indicates that in all cases where the user completely disagreed/agreed, the party completely agreed/disagreed, while +100 means that in all cases where the user completely agreed/disagreed, the party also completely agreed/disagreed, and 0 means that either the party or the user was neutral. For the euclidean algorithm, −100 indicates that in all cases where the user completely disagreed/agreed, the party completely agreed/disagreed, while +100 means that in all cases the user and the party had the same opinion, and 0 means that there is an equal number of agreements and disagreements. For the city block algorithm, −100 indicates that in all cases where the user completely disagreed/agreed, the party completely agreed/disagreed, while +100 means that in all cases the user and the party had the same opinion, and 0 means that in all cases, the user had an opinion "on the other side" - that is, if the party completely agreed, the user was either neutral, disagreed, or completely disagreed. See for a further explanation of the algorithms Mendez (2012, 2014b).

In addition, *pos_neg* is a binary variable, which is 0 when the question is positive and 1 when the question is negative, *age* is a numeric variable giving the self-reported age of the user ranging between 16 and 79, *sex* is a binary variable with 0 for women and 1 for men, *education* is a 7-point ordinal variable running from 1 "Not completed primary education" to 7 "Postgraduate education", *interestInPolitics* is 5-point ordinal variable running from 1 "Very Interested in Politics" to 5 "Not At All Interested in Politics", and *gutmann* are the number of Gutmann errors ranging between 0 and 130. In addition, interactions between *pos_neg* and *age*, *sex*, *education*, *interestInPolitics* and *gutmann* have been added.

All linear models were calculated using the **lm** command in R (R Core Team 2018).

	Version A								Version B							
	Ex Ante				Quasi-Inductive				Ex Ante				Quasi-Inductive			
	EC		CU		EC		CU‡		EC		CU		EC		CU	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.14*	49							0.11	30						
2	0.09*	29							0.10	32					0.30	18
3			0.31	0			0.51	0			0.04	66			0.46	0
4	0.02*	181					0.38*	10	0.02*	121						
5	0.05	115							0.05*	112						
6			0.21	27					0.06	59						
7	0.09*	60							−0.02*	220						
8			0.04	134					−0.03	198						
9			0.34*	0			0.53*	0			0.02*	130			0.49*	0
10			0.14*	74					−0.09	477						
11			0.23	0			0.54	0			−0.08*	466			0.48	0
12			0.22	19			0.34	7			0.04	58			0.31	10
13			0.32*	0			0.47*	3			0.08*	0			0.46*	0
14			0.34	0			0.54	0			0.05	32			0.52	0
15	0.02	104							0.10*	26						
16	0.13	22							0.07*	90						
17			0.21	27							−0.08*	383				
18	0.08*	51							−0.01	206						
19			−0.27	644			0.45*	0			−0.05	356			0.38*	9
20			0.18*	65							0.06*	65				
21	0.16*	23							0.06	72						
22	0.11*	42							−0.10	449						
23	0.12	31							−0.06	278						
24	0.11*	46							−0.08*	377						
25			0.33	0			0.52	0			−0.09*	445			0.36	72

Table B.3 continued from previous page

	Version A								Version B							
	Ex Ante				Quasi-Inductive				Ex Ante				Quasi-Inductive			
	EC		CU		EC		CU‡		EC		CU		EC		CU	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
Overall H	0.09		0.19				0.47		0.02		−0.01				0.42	
α	0.52		0.72				0.86		0.15		−0.12				0.84	
α (ordinal)	0.57		0.77				0.89		0.22		−0.03				0.87	
ω (ordinal)	0.67		0.89				0.92		0.67		0.88				0.90	
LCRC †	0.57 (4)		0.77 (5)				0.88 (5)		0.29 (4)		0.22 (5)				0.84 (3)	
N	1328		1328				1328		1346		1346				1346	

∗ Items with an asterisk (∗) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.
‡ Question 22 was dropped as the content of the topic did not belong to a cultural scale.

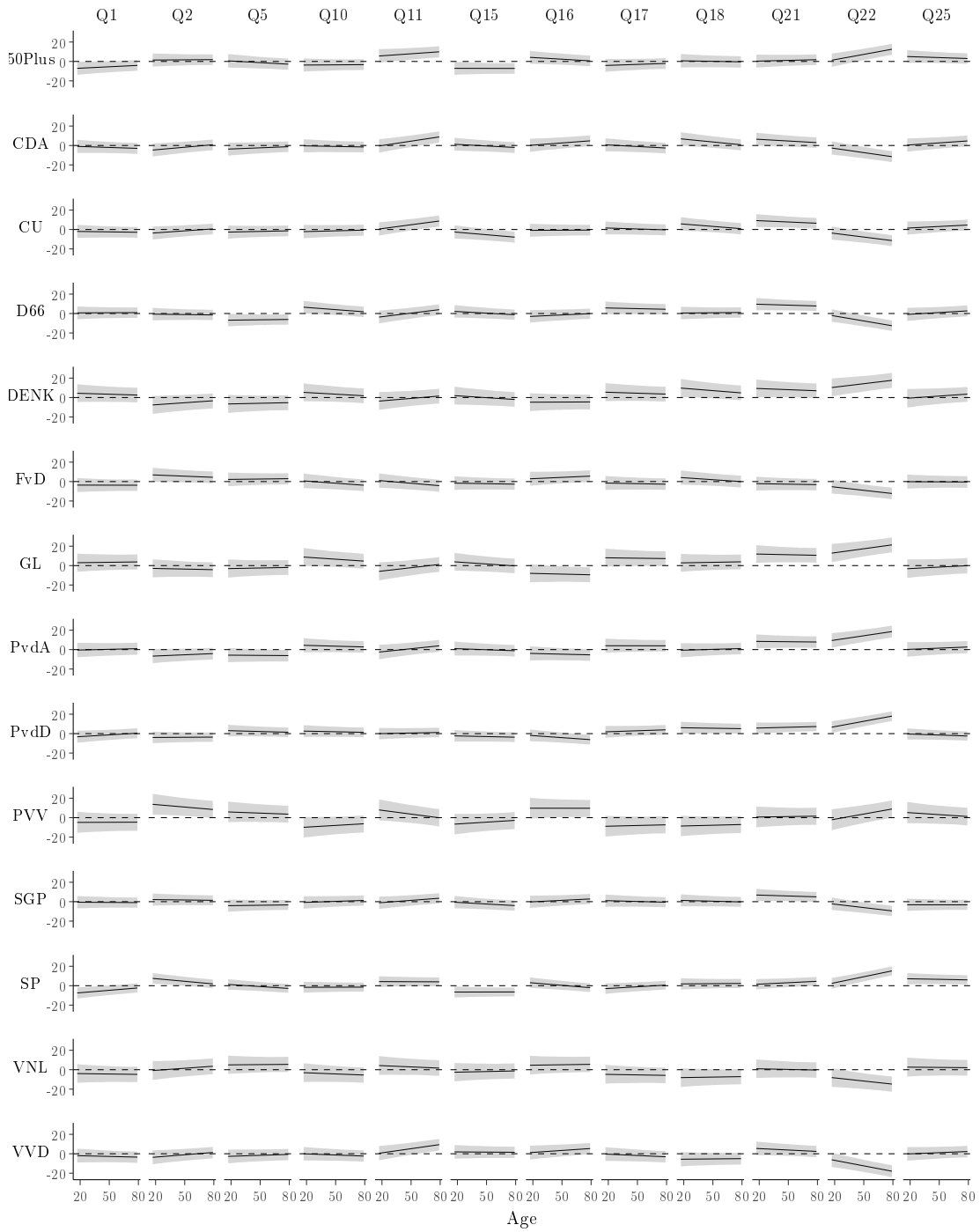
Average Marginal Effects with 95% CI

Party	Question 1		Question 2		Question 5		Question 10		Question 11		Question 15	
50Plus	−0.51	[−1.45;0.43]	−1.54	[−2.48;−0.60]	−4.52	[−5.50;−3.55]	1.64	[0.70;2.58]	−4.22	[−5.19;−3.25]	−0.97	[−1.91;−0.03]
CDA	−1.02	[−1.96;−0.07]	−1.03	[−1.98;−0.09]	−2.19	[−3.13;−1.24]	−2.61	[−3.55;−1.66]	−3.71	[−4.68;−2.74]	−0.75	[−1.68;0.18]
CU	−0.94	[−1.88;0.01]	−1.12	[−2.06;−0.17]	−2.27	[−3.21;−1.33]	−3.45	[−4.42;−2.48]	−3.79	[−4.76;−2.83]	−1.83	[−2.79;−0.87]
D66	−0.23	[−1.15;0.69]	1.13	[0.23;2.02]	−3.36	[−4.26;−2.46]	3.38	[12.47;4.29]	−4.89	[−5.81;−3.96]	−0.30	[−1.20;0.59]
DENK	0.16	[−1.12;1.45]	−1.73	[−3.02;−0.44]	−2.88	[−4.18;−1.59]	1.83	[0.54;3.12]	−2.77	[−4.06;−1.49]	−0.78	[−2.07;0.50]
FvD	−1.20	[−2.20;−0.20]	3.22	[2.19;4.24]	0.85	[−0.13;1.83]	2.02	[0.98;3.07]	5.22	[4.18;6.26]	−0.85	[−1.83;0.14]
GL	0.39	[−0.93;1.70]	0.50	[−0.79;1.79]	1.71	[0.40;3.01]	4.01	[2.71;5.30]	−5.51	[−6.82;−4.20]	−0.11	[−1.40;1.17]
PvdA	−0.38	[−1.42;0.67]	−2.07	[−3.09;−1.05]	−3.22	[−4.24;−2.20]	2.17	[1.16;3.18]	−4.75	[−5.79;−3.70]	−0.44	[−1.45;0.57]
PvdD	−0.42	[−1.28;0.44]	−2.02	[−2.84;−1.20]	2.52	[1.68;3.35]	3.19	[2.34;4.04]	−3.06	[−3.89;−2.24]	−0.92	[−1.76;−0.09]
PVV	−0.23	[−1.74;1.29]	3.34	[1.84;4.84]	−1.88	[−3.36;−0.39]	−2.91	[−4.40;−1.43]	5.35	[3.84;6.85]	−0.24	[−1.71;1.24]
SGP	−0.61	[−1.50;0.27]	1.90	[1.02;2.77]	−2.59	[−3.46;−1.72]	−3.13	[−4.03;−2.23]	−2.48	[−3.37;−1.60]	−1.07	[−1.97;−0.18]
SP	−1.37	[−2.17;−0.57]	2.26	[1.50;3.02]	−2.22	[−2.98;−1.46]	2.25	[1.45;3.04]	−2.12	[−2.89;−1.35]	−1.87	[−2.66;−1.09]
VNL	−0.96	[−2.28;0.37]	−1.28	[−2.65;0.08]	1.09	[−0.22;2.41]	0.71	[−0.64;2.06]	3.52	[2.17;4.87]	−0.35	[−1.68;0.98]
VVD	−1.05	[−2.04;−0.06]	−1.00	[−2.00;−0.01]	−2.16	[−3.14;−1.17]	−0.79	[−1.77;0.20]	−3.68	[−4.70;−2.66]	0.04	[−0.95;1.03]

(continues on the next page)

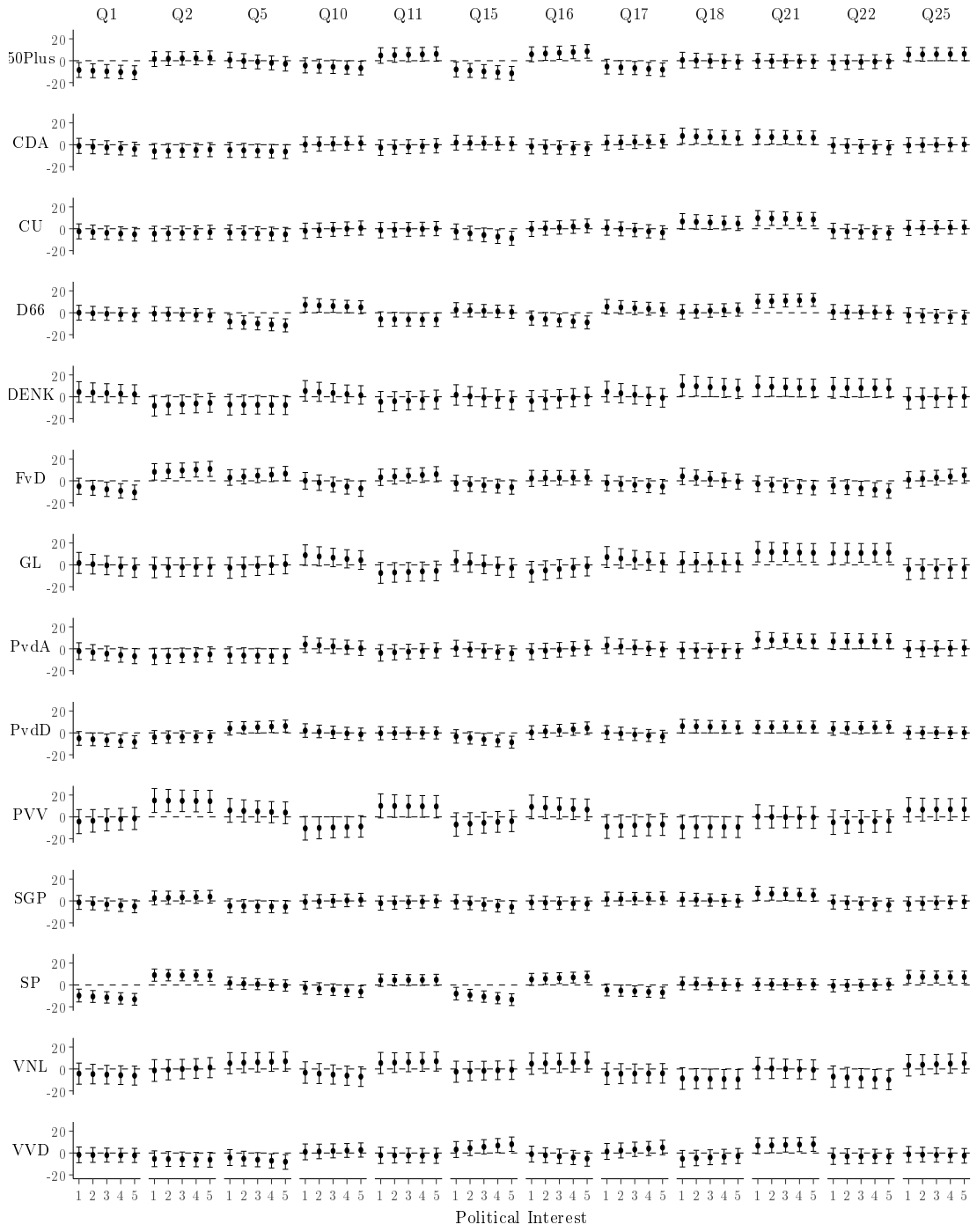
Party	Question 16		Question 17		Question 18		Question 21		Question 22		Question 25	
50Plus	−2.39	[−3.32;−1.46]	1.38	[0.44;2.31]	1.69	[0.71;2.66]	3.75	[2.82;4.68]	3.54	[2.57;4.51]	0.16	[−0.77;1.08]
CDA	3.20	[2.26;4.14]	−2.22	[−3.16;−1.27]	1.18	[0.21;2.15]	3.24	[2.31;4.18]	−2.83	[−3.79;−1.86]	−1.20	[−2.15;−0.24]
CU	−1.96	[−2.90;−1.02]	1.88	[0.93;2.83]	1.26	[0.29;2.23]	5.50	[4.57;6.44]	−2.75	[−3.70;−1.79]	−1.28	[−2.23;−0.33]
D66	2.02	[1.13;2.92]	2.98	[2.06;3.89]	−0.60	[−1.51;0.31]	4.42	[3.53;5.31]	−3.18	[−4.10;−2.26]	−3.08	[−4.02;−2.13]
DENK	−2.58	[−3.87;−1.28]	2.50	[1.19;3.80]	1.87	[0.56;3.19]	3.94	[2.66;5.22]	3.72	[2.42;5.03]	−2.60	[−3.93;−1.26]
FvD	3.38	[2.38;4.39]	−0.88	[−1.88;0.12]	0.99	[−0.04;2.02]	−0.91	[−1.91;0.09]	−3.01	[−4.01;−2.01]	2.47	[1.45;3.49]
GL	−4.91	[−6.22;−3.61]	3.60	[2.29;4.90]	0.02	[−1.27;1.32]	5.04	[3.77;6.32]	4.83	[3.52;6.13]	−3.70	[−5.03;−2.37]
PvdA	−2.91	[−3.93;−1.90]	1.90	[0.89;2.91]	−0.74	[−1.76;0.28]	4.28	[3.27;5.28]	4.06	[3.03;5.10]	−2.94	[−4.00;−1.87]
PvdD	−4.10	[−4.94;−3.26]	2.79	[1.94;3.63]	2.16	[1.30;3.03]	4.23	[3.42;5.04]	4.01	[3.16;4.87]	−0.32	[−1.12;0.48]
PVV	3.51	[2.02;5.00]	−2.52	[−4.02;−1.03]	−1.08	[−2.55;0.39]	5.11	[3.62;6.61]	2.72	[1.21;4.23]	3.54	[2.01;5.07]
SGP	2.79	[1.91;3.68]	−1.81	[−2.69;−0.93]	−0.37	[−1.24;0.50]	3.65	[2.78;4.52]	−2.42	[−3.31;−1.53]	1.88	[0.99;2.77]
SP	−1.92	[−2.68;−1.16]	0.91	[0.15;1.66]	1.22	[0.41;2.02]	3.28	[2.54;4.03]	3.07	[2.29;3.85]	−1.94	[−2.75;−1.13]
VNL	1.14	[−0.20;2.48]	−2.64	[−3.99;−1.29]	−2.20	[−3.54;−0.85]	2.82	[1.48;4.16]	−3.25	[−4.59;−1.91]	2.71	[1.35;4.07]
VVD	3.23	[2.23;4.22]	−2.24	[−3.24;−1.25]	−2.53	[−3.57;−1.50]	3.22	[2.23;4.20]	−4.38	[−5.41;−3.36]	0.49	[−0.50;1.48]

Conditional Effect of Age



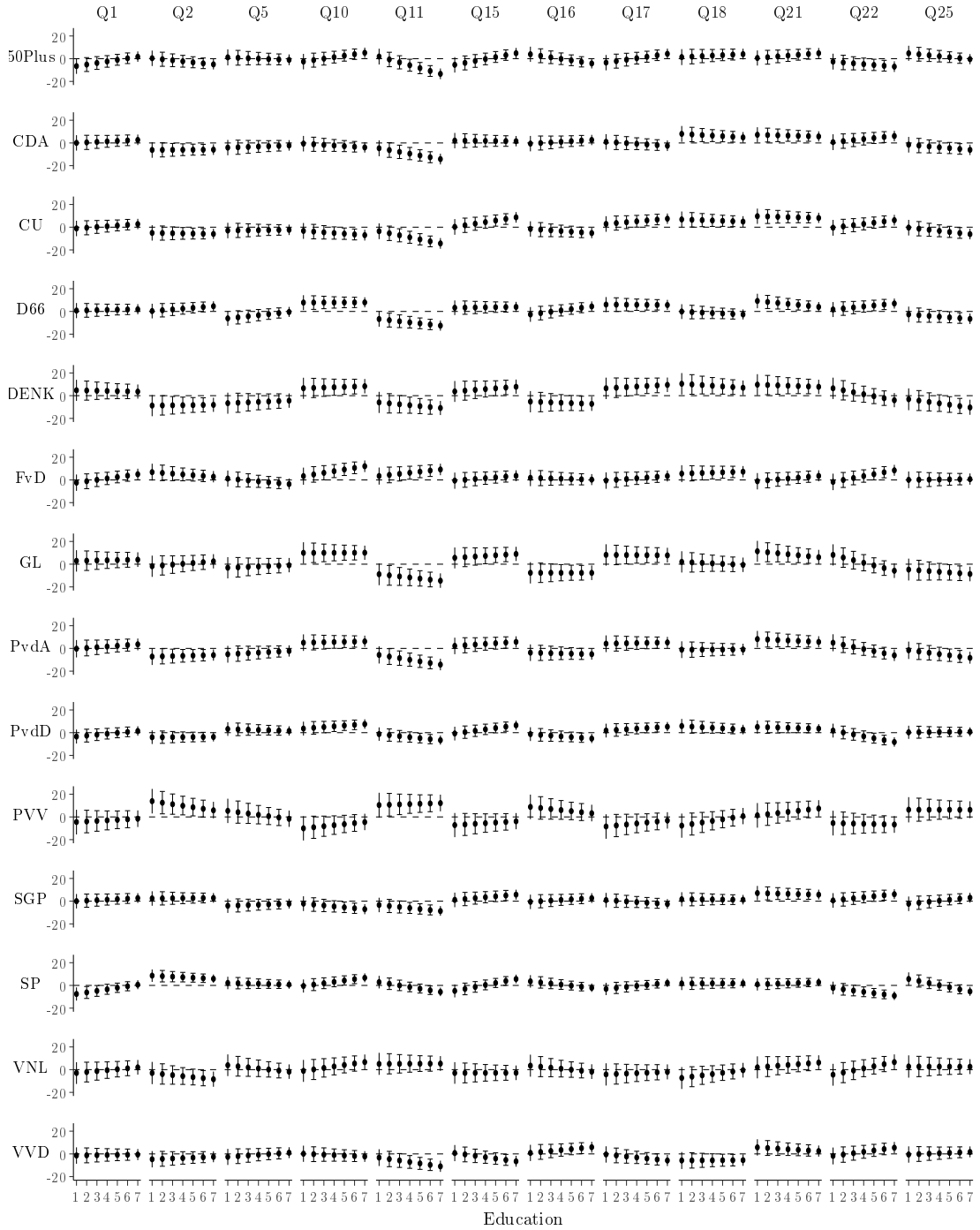
Conditional Effect of age on the marginal effect of the positive-negative condition on the match for the Hybrid algorithm. The y-axis shows the magnitude of the coefficient of the positive-negative condition on the match. 95% confidence intervals are also shown.

Conditional Effect of Political Interest



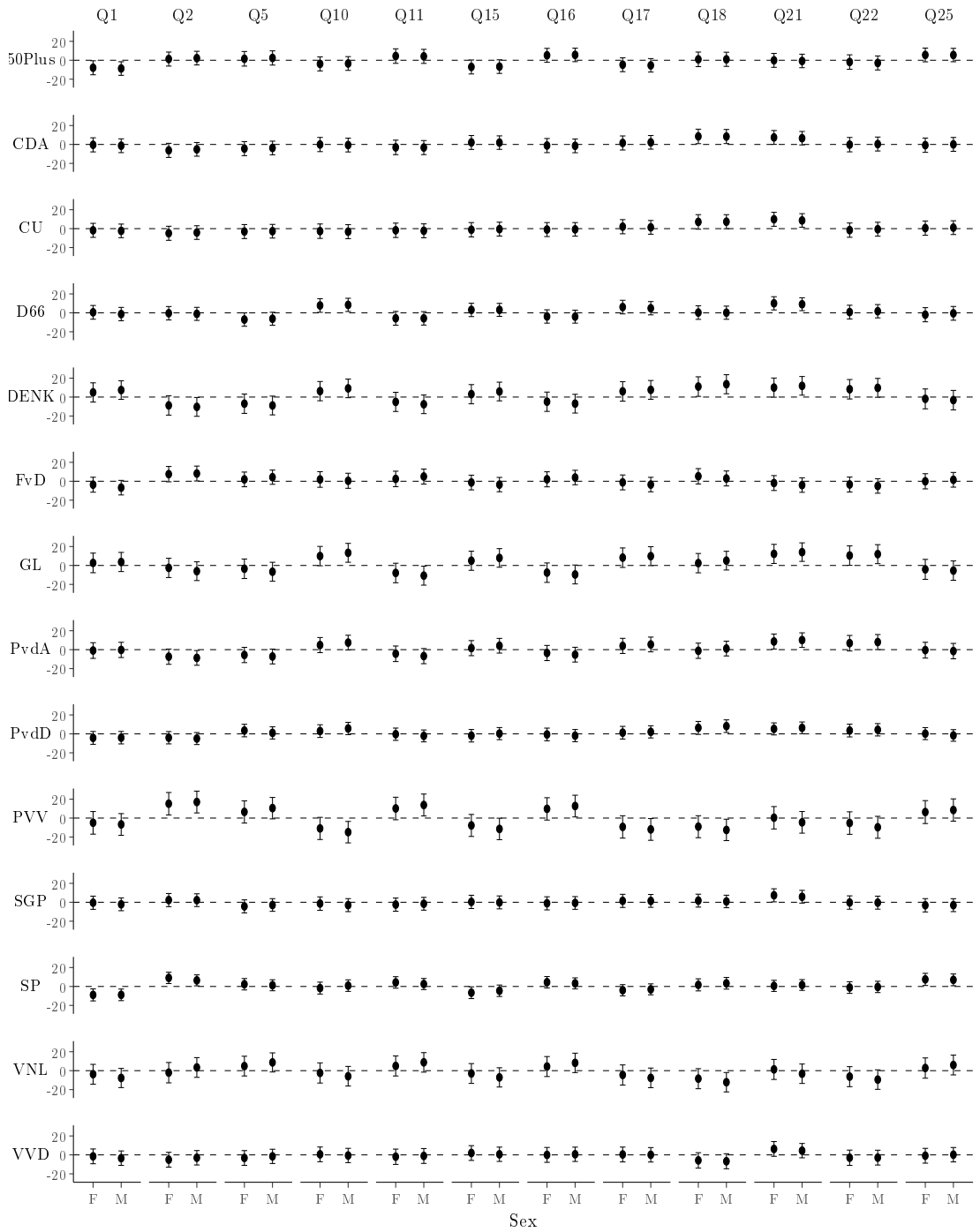
Conditional Effect of the 5-point political interest scale on the marginal effect of the positive-negative condition on the match for the Hybrid algorithm. The y-axis shows the magnitude of the coefficient of the positive-negative condition on the match. 95% confidence intervals are also shown.

Conditional Effect of Education



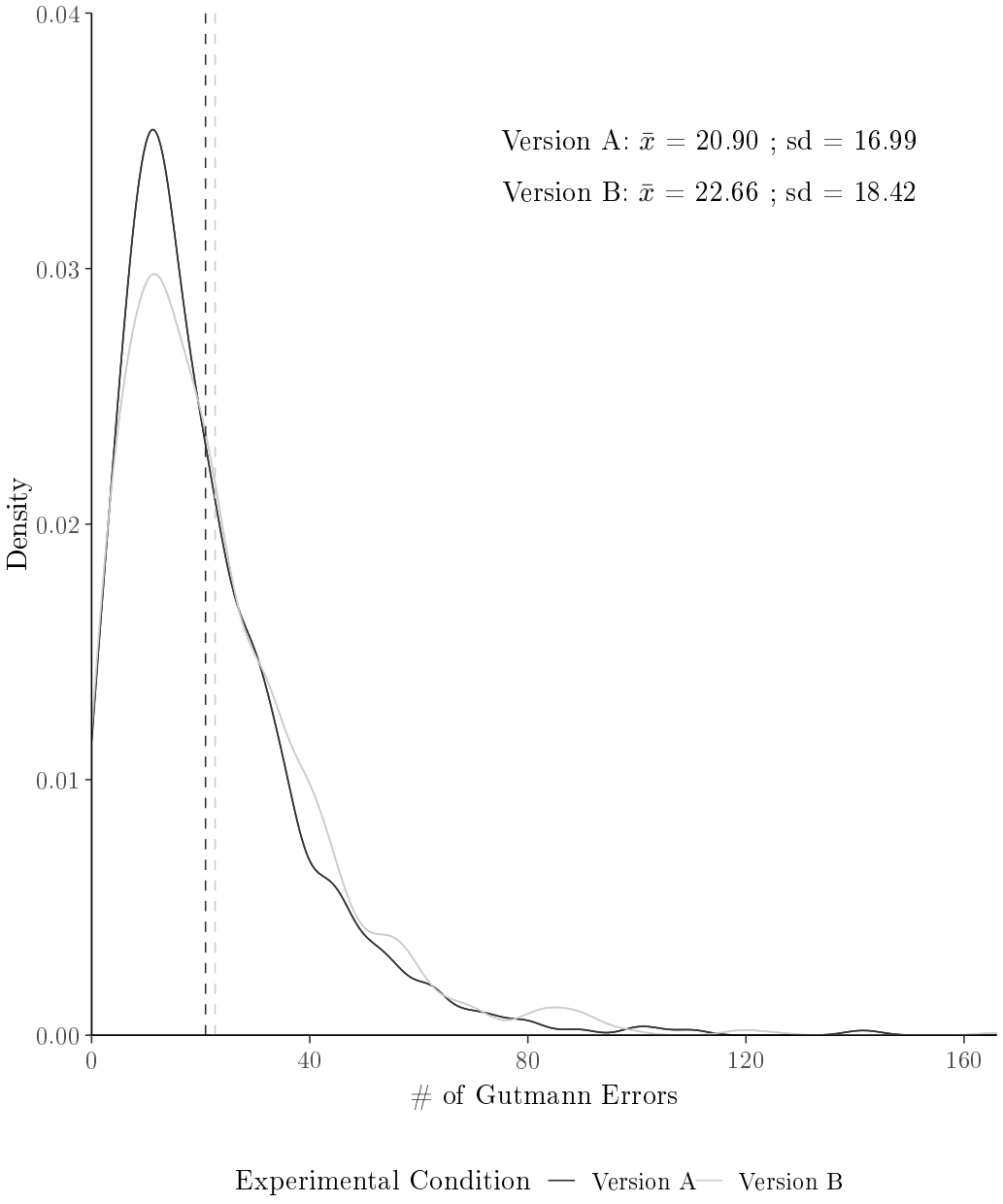
Conditional Effect of education on the marginal effect of the positive-negative condition on the match for the Hybrid algorithm. The y-axis shows the magnitude of the coefficient of the positive-negative condition on the match. 95% confidence intervals are also shown.

Conditional Effect of Sex



Conditional Effect of sex on the marginal effect of the positive-negative condition on the match for the Hybrid algorithm. The y-axis shows the magnitude of the coefficient of the positive-negative condition on the match. 95% confidence intervals are also shown.

Distribution of the # of Gutmann Errors



Distribution of the # of Gutmann Errors for the Stemconsult VAA for both the experimental groups

C | Dynamic Scale Validation Scales

Austria

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.53*	4					0.57*	0				
2	0.36*	14					0.39*	7				
3	0.42*	11					0.46*	0				
4	0.42	13					0.46	0				
5	0.36	36					0.40	33				
6	0.53*	0					0.56*	0				
7												
8	0.30	19					0.33	40				
9	0.28	54										
10												
11	0.33*	16										
12			0.32	16					0.36	38		
13			0.24	28					0.28	10		
14			0.35	8					0.41	0		
15			0.28*	33					0.34*	0		
16			0.27	14					0.34	11		
17			0.30	14					0.33	26		
18			0.11	156								
19			0.28	20					0.34	9		
20			0.17	33								
21					0.37	24					0.39	31
22					0.24	27					0.26	44
23					0.33	37					0.35	54
24					0.17*	66						
25					0.36*	7					0.38*	0
26					0.28*	15					0.30*	14

Table C.1 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.29*	17					0.30*	30
29					0.33	9					0.35	28
30												
Total H	0.39		0.26		0.30		0.45		0.34		0.33	
α	0.82		0.73		0.74		0.83		0.76		0.74	
α (Ord)	0.86		0.77		0.78		0.86		0.79		0.79	
ω (Ord)	0.85		0.68		0.77		0.86		0.73		0.78	
LCRC†	0.83 (6)		0.74 (6)		0.77 (8)		0.83 (6)		0.76 (6)		0.77 (7)	
N	5421		5421		5421		5421		5421		5421	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Croatia

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.36*	0					0.44*	0				
2	0.29*	58					0.35*	14				
3	0.29*	38					0.36*	33				
4	0.3	61					0.34	41				
5	0.02	238										
6	0.39*	29					0.47*	0				
7												
8			0.2	47								
9			0.21	0					0.43	0		
10			0.2	60					0.3	0		
11			0.15*	95								
12			0.16	49								
13			0.22	39					0.33	29		
14			0.13	72								
15			0.09*	97								
16			0.16*	60								
17			0.11	66								
18			0.21	47					0.43	0		
19			-0.01*	349								
20												
21					0.29	39					0.36	85
22					0.16	105						
23					0.12	150						
24					0.14*	89						
25					0.46*	0					0.59*	-1
26					0.39*	41					0.52*	25

Table C.2 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.35*	36					0.41*	52
28					0.43*	0					0.56*	36
29					0.36	10					0.49	42
30												
Total H	0.28		0.15		0.32		0.39		0.37		0.5	
α	0.68		0.65		0.77		0.74		0.66		0.83	
α (Ord)	0.7		0.69		0.79		0.78		0.71		0.86	
ω (Ord)	0.68		0.63		0.71		0.76		0.68		0.85	
LCRC†	0.7 (4)		0.68 (8)		0.81 (7)		0.74 (4)		0.67 (4)		0.85 (5)	
N	4126		4126		4126		4126		4126		4126	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Czech Republic

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.50*	58					0.50*	58				
2	0.43*	6					0.43*	6				
3	0.32*	170					0.32*	170				
4	0.46	0					0.46	0				
5	0.32	80					0.32	80				
6	0.57*	0					0.57*	0				
7												
8	0.34	12					0.34*	12				
9	0.33	91					0.33*	91				
10	0.54*	0					0.54*	0				
11			0.3	35					0.33	84		
12			0.29	39					0.33	40		
13			0.37	39					0.44	20		
14			0.28*	56					0.34*	41		
15			0.2	99								
16			0.37	9					0.44	0		
17			0	526								
18			0.29*	112					0.34*	128		
19			0.33	12					0.4	3		
20					0.25	29					0.53	0
21					0.16	51						
22					0.19	74						
23					0.10*	147						
24					0.25*	30					0.35*	36
25					0.16*	118						
26					0.22*	38						

Table C.3 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.25	29					0.53	0
28											0.49	0
29												
30												
Total H	0.43		0.27		0.2		0.43		0.37		0.47	
α	0.85		0.75		0.62		0.85		0.79		0.73	
α (Ord)	0.87		0.77		0.65		0.87		0.82		0.8	
ω (Ord)	0.87		0.61		0.26		0.87		0.77		0.79	
LCRC†	0.87 (9)		0.78 (9)		0.68 (8)		0.87 (9)		0.79 (7)		0.74 (4)	
N	17833		17833		17833		17833		17833		17833	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Denmark

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.26*	31					0.28*	0				
2	0.24*	80					0.26*	65				
3	0.31*	46					0.32*	30				
4	0.25	18					0.26	28				
5	0.23	75					0.23	0				
6	0.34*	47					0.36*	31				
7												
8	-0.46	1887					-0.48	1822				
9	0.28	33					0.28	12				
10	0.29*	55					0.31*	32				
11	0.25	33					0.27	30				
12	0.06	234										
13	0.29	42					0.29	39				
14			0.35	0					0.39	0		
15			0.4	0					0.44	0		
16			0.32	82					0.37	60		
17			0.28*	0					0.33*	0		
18			0.39	25					0.43	0		
19			0.24	32					0.24	60		
20			0.07	298								
21									0.32	12		
22												
23					0.38	0					0.46	0
24					0.23	76					0.28	77
25					0.26	48					0.32	26
26					0.20*	165						

Table C.4 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.27*	0					0.27*	43
28					0.19*	81						
29					0.23*	55						
30					0.35	42					0.42	0
Total H	0.2		0.3		0.27		0.21		0.36		0.35	
α	0.72		0.73		0.68		0.73		0.78		0.68	
α (Ord)	0.75		0.76		0.74		0.77		0.82		0.74	
ω (Ord)	0.8		0.69		0.68		0.82		0.76		0.71	
LCRC†	0.81 (12)		0.76 (7)		0.73 (8)		0.81 (11)		0.8 (7)		0.72 (5)	
N	55832		55832		55832		55832		55832		55832	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Estonia

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.33*	9					0.40*	25				
2	0.21*	61					0.33*	0				
3	0.22*	22					0.33*	34				
4	0.23	67					0.28	67				
5	0.05	241										
6	0.34*	0										
7												
8	0.1	116										
9	0.17	71										
10			0.16	75								
11			0.12	73					0.31	0		
12			0.2	0					0.25	38		
13			0.20*	16								
14			0.1	63					0.27	36		
15			0.2	24					0.22	71		
16			0.06	139								
17			0.05*	140								
18												
19			0.20*	23								
20					0.16	78						
21					0.13	89						
22					0.17	58						
23					0.10*	111						
24					0.15*	59					0.26	35
25					0.08*	143					0.24	91
26					0.15*	70					0.19	108

Table C.5 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.10*	56						
28												
29												
30												
Total H	0.2		0.15		0.13		0.33		0.26		0.23	
α	0.63		0.57		0.49		0.61		0.53		0.4	
α (Ord)	0.68		0.6		0.54		0.68		0.57		0.47	
ω (Ord)	0.58		0.42		0.27		0.62		0.56		0.5	
LCRC†	0.65 (7)		0.63 (9)		0.54 (7)		0.6 (4)		0.56 (4)		0.45 (3)	
N	7953		7953		7953		7953		7953		7953	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Finland

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.50*	-2					0.54*	-1				
2	0.40*	9					0.42*	28				
3	0.43*	0					0.44*	15				
4	0.4	17					0.42	30				
5	0.25	67										
6	0.53*	-2					0.57*	0				
7							0.39	28				
8												
9												
10												
11												
12			0.36	38					0.39	10		
13			0.37	14					0.42	3		
14			0.34	12					0.39	28		
15			0.33*	37					0.37*	0		
16			0.39	30					0.43	0		
17			0.35	17					0.39	0		
18			0.12	152								
19												
20												
21												
22					0.28	53					0.39*	7
23					0.25	31						
24					0.31	35						
25					0.19*	35						
26					0.34*	11					0.41	4

Table C.6 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.23*	27					0.35	14
28					0.29*	0						
29												
30											0.28	28
Total H	0.42		0.33		0.28		0.46		0.4		0.36	
α	0.79		0.75		0.66		0.81		0.78		0.59	
α (Ord)	0.82		0.78		0.72		0.85		0.82		0.69	
ω (Ord)	0.83		0.76		0.69		0.83		0.76		0.66	
LCRC†	0.8 (4)		0.77 (7)		0.72 (6)		0.82 (4)		0.78 (6)		0.65 (4)	
N	4664		4664		4664		4664		4664		4664	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

France

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.49*	0					0.53*	0				
2	0.32*	38					0.35*	65				
3	0.38*	40					0.41*	13				
4	0.39	47					0.41	36				
5	0.36	43					0.37	31				
6	0.50*	6					0.53*	0				
7												
8	0.25	95										
9	0.29	74					0.3	76				
10												
11			0.36	34					0.38	42		
12			0.46	5					0.5	3		
13			0.41	23					0.44	2		
14			0.40*	17					0.43*	11		
15			0.44	29					0.46	19		
16			0.43	31					0.46	9		
17			0.17	187								
18			0.39*	19					0.43*	12		
19			0.44	24					0.47	7		
20												
21					0.49	4					0.51	-1
22					0.31	97					0.33	68
23					0.34	18					0.36	27
24					0.27*	52						
25					0.43*	32					0.45*	26
26					0.34*	46					0.35*	42

Table C.7 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.34*	55					0.34*	33
28					0.46	6					0.48	0
29					0.52	8					0.54	2
30					0.43	24					0.45	32
Total H	0.37		0.39		0.4		0.42		0.45		0.42	
α	0.81		0.83		0.83		0.82		0.84		0.83	
α (Ord)	0.83		0.86		0.87		0.84		0.87		0.87	
ω (Ord)	0.86		0.8		0.87		0.87		0.84		0.87	
LCRC†	0.82 (6)		0.84 (6)		0.85 (7)		0.83 (6)		0.85 (6)		0.85 (7)	
N	5268		5268		5268		5268		5268		5268	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Germany

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.53*	0					0.55*	0				
2	0.35*	51					0.37*	37				
3	0.45*	7					0.49*	8				
4	0.45	14					0.47	41				
5	0.37	47					0.41	27				
6	0.54*	-1					0.56*	0				
7												
8	0.32	55					0.34	77				
9	0.29	64										
10												
11												
12												
13			0.35	9					0.38	13		
14			0.27	12					0.32	28		
15			0.38	12					0.43	0		
16			0.28*	29					0.32*	15		
17			0.3	36					0.34	33		
18			0.32	12					0.35	34		
19			0.08	191								
20												
21												
22					0.34	19					0.37	38
23					0.27	22					0.3	57
24					0.26	61					0.3	20
25					0.16*	65						
26					0.33*	14					0.35*	10

Table C.8 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.26*	50					0.29*	40
28					0.30*	17					0.31*	15
29												
30												
Total H	0.41		0.28		0.28		0.46		0.36		0.32	
α	0.83		0.71		0.69		0.84		0.74		0.69	
α (Ord)	0.85		0.74		0.74		0.86		0.78		0.75	
ω (Ord)	0.86		0.58		0.7		0.88		0.72		0.72	
LCRC†	0.85 (8)		0.73 (6)		0.71 (6)		0.85 (7)		0.75 (6)		0.73 (6)	
N	5447		5447		5447		5447		5447		5447	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Greece

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.39*	5					0.46*	4				
2	0.27*	89					0.32*	0				
3	0.24*	102					0.27*	106				
4	0.33	29					0.36	31				
5	-0.06	453										
6	0.39*	32					0.46*	0				
7												
8	0.22	74					0.26	50				
9												
10			0.43	0					0.46	17		
11			0.44	3					0.47	19		
12			0.41	0					0.45	0		
13			0.37	12					0.41	25		
14			0.28*	31								
15			0.33	13					0.36	0		
16			0.4	0					0.43	0		
17			0.28	53								
18			0.36	0					0.38	0		
19			0.38	10					0.41	11		
20			0.28	50								
21					0.34	301					0.37	272
22					0.28	139						
23					0.24	127						
24					0.17*	219						
25					0.42*	0					0.52*	0
26					0.31*	65					0.38*	50

Table C.9 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.33*	58					0.38*	32
28					0.41*	22					0.49*	2
29					0.43	0					0.5	0
30					0.41*	27					0.50*	17
Total H	0.27		0.36		0.34		0.36		0.42		0.45	
α	0.69		0.84		0.81		0.74		0.83		0.82	
α (Ord)	0.69		0.87		0.84		0.77		0.87		0.86	
ω (Ord)	0.52		0.86		0.87		0.78		0.85		0.85	
LCRC†	0.74 (7)		0.85 (11)		0.84 (10)		0.76 (6)		0.84 (8)		0.84 (7)	
N	36967		36967		36967		36967		36967		36967	

∗ Items with an asterisk (∗) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Hungary

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.53*	21					0.58*	0				
2	0.43*	41					0.48*	60				
3	0.46*	9					0.51*	9				
4	0.5	24					0.57	2				
5	-0.1	390										
6	0.55*	2					0.62*	0				
7												
8	0.31	20					0.34	78				
9	0.52*	0					0.58*	-2				
10												
11							0.46	33				
12			0.3	0					0.44	0		
13			0.23	45					0.3	46		
14			0.31	11					0.48	3		
15			-0.08*	381								
16			0.14	73								
17			0.19	27								
18			0.25	64					0.43	8		
19												
20												
21												
22					0.35	31					0.40*	0
23					0.27	25					0.29*	26
24					0.21	91						
25					0.18*	104						
26					0.41*	5					0.51	0

Table C.10 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.35*	8					0.46	7
28					0.31*	35					0.39	6
29					0.19	63						
30												
Total H	0.42		0.2		0.29		0.52		0.41		0.41	
α	0.83		0.6		0.73		0.88		0.72		0.73	
α (Ord)	0.84		0.62		0.76		0.9		0.75		0.78	
ω (Ord)	0.85		0.51		0.72		0.9		0.67		0.76	
LCRC†	0.86 (6)		0.67 (5)		0.76 (7)		0.89 (6)		0.72 (4)		0.76 (4)	
N	3616		3616		3616		3616		3616		3616	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Ireland

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.34*	7					0.45*	0				
2	0.23*	24					0.34*	0				
3	0.26*	39					0.32*	13				
4	0.29	42					0.33	31				
5	0.14	138										
6	0.35*	0					0.46*	0				
7												
8	0.17	47										
9	0.17	56										
10			0.32	17					0.39	0		
11			0.29	19					0.37	11		
12			0.27	59					0.35	0		
13			0.30*	36					0.37*	38		
14			0.28	36					0.35	9		
15			0.29	40					0.35	24		
16			0.22	121					0.21	186		
17			0.17*	70								
18			0.17	70								
19												
20			0.25*	26					0.29*	43		
21			0.05	243								
22			0.31	20					0.36	48		
23					0.24	17						
24					0.23	60					0.26	63
25					0.26	29					0.3	40
26					0.18*	52						

Table C.11 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.34*	6					0.36*	0
28					0.27*	20					0.32*	0
29					0.28*	20					0.33*	0
30					0.26*	62						
Total H	0.24		0.24		0.26		0.38		0.34		0.31	
α	0.69		0.77		0.69		0.71		0.79		0.66	
α (Ord)	0.73		0.8		0.76		0.77		0.83		0.73	
ω (Ord)	0.69		0.8		0.78		0.72		0.81		0.77	
LCRC†	0.7 (6)		0.79 (8)		0.73 (7)		0.71 (4)		0.8 (6)		0.71 (5)	
N	5325		5325		5325		5325		5325		5325	

$*$ Items with an asterisk $(*)$ were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Italy

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.45*	0					0.52*	0				
2	0.30*	75					0.35*	59				
3	0.37*	15					0.38*	13				
4	0.36	59					0.39	77				
5	0.07	203										
6	0.42*	0					0.49*	0				
7							0.35	0				
8			0.26	17					0.32	0		
9			0.22	13					0.36	0		
10			0.25	33					0.34	27		
11			0.25*	38								
12			0.23	15					0.34	0		
13			0.26	17					0.32	10		
14			0.14	137								
15			0.17*	76								
16			0.13*	123								
17												
18			0.18*	44								
19					0.37	51					0.43	0
20					0.18	60						
21					0.32	48					0.34	73
22					0.16*	103						
23					0.41*	0					0.47*	0
24					0.33*	21					0.39*	16
25					0.33*	30					0.36*	24
26					0.39	0					0.44	0

Table C.12 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.36*	32					0.41*	24
28												
29												
30												
Total H	0.33		0.21		0.32		0.42		0.33		0.41	
α	0.72		0.7		0.78		0.78		0.69		0.8	
α (Ord)	0.75		0.74		0.82		0.81		0.73		0.84	
ω (Ord)	0.81		0.68		0.82		0.81		0.69		0.85	
LCRC†	0.76 (6)		0.73 (10)		0.81 (9)		0.8 (6)		0.69 (5)		0.83 (7)	
N	20068		20068		20068		20068		20068		20068	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Lithuania

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.38*	0					0.44*	0				
2	0.25*	56					0.31*	48				
3	0.22*	25					0.25*	43				
4	0.27	61					0.3	71				
5	0.02	181										
6	0.40*	4					0.47*	0				
7												
8	0.23	28					0.26	38				
9	0.37*	0					0.44*	3				
10			0.23	34					0.31	40		
11			0.14	78								
12			0.24	14					0.31	28		
13			0.19*	25					0.26*	0		
14			0.1	73								
15			0.22	24					0.3	0		
16			0.07	140								
17			0.11*	53								
18												
19			0.21*	53					0.30*	69		
20												
21					0.2	46					0.27	20
22					0.18	66						
23					0.19	28					0.22	65
24					0.08*	134						
25					0.27*	0					0.38*	0
26					0.22*	35					0.32*	8

Table C.13 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.25*	12					0.30*	36
28					0.07*	95						
29												
30												
Total H	0.27		0.17		0.19		0.35		0.3		0.3	
α	0.71		0.61		0.6		0.75		0.64		0.62	
α (Ord)	0.76		0.65		0.64		0.81		0.69		0.68	
ω (Ord)	0.69		0.54		0.53		0.79		0.63		0.59	
LCRC†	0.73 (6)		0.63 (5)		0.63 (5)		0.77 (6)		0.63 (4)		0.66 (4)	
N	5317		5317		5317		5317		5317		5317	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Poland

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.45*	7					0.45*	7				
2	0.40*	84					0.40*	84				
3	0.28*	314					0.28*	314				
4	0.48	53					0.48	53				
5	0.34	115					0.34	115				
6	0.52*	26					0.52*	26				
7												
8	0.28	119					0.28	119				
9			0.29	28					0.42	0		
10			0.26	16					0.36	40		
11			0.3	0					0.45	6		
12			0.24*	148					0.38*	9		
13			0.19	116								
14			0.29	42					0.43	20		
15			-0.21	885								
16			0.14*	158								
17					0.13	70					0.29	27
18					0.01	340						
19					-0.01	493						
20					0.07*	144						
21					0.32*	25					0.53*	0
22					0.31*	31					0.53*	0
23					0.21*	48						
24					0.04*	169						
25											0.34	46
26												

Table C.14 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.32*	28					0.53*	0
28												
29					0.30*	36					0.42*	46
30					0.27*	65					0.52*	0
Total H	0.4		0.19		0.19		0.4		0.41		0.46	
α	0.79		0.62		0.69		0.79		0.74		0.83	
α (Ord)	0.82		0.68		0.7		0.82		0.8		0.86	
ω (Ord)	0.79		0.46		0.39		0.79		0.77		0.88	
LCRC†	0.81 (7)		0.69 (8)		0.77 (11)		0.81 (7)		0.75 (5)		0.85 (7)	
N	41554		41554		41554		41554		41554		41554	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Portugal

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.34*	0					0.47*	0				
2	0.22*	92					0.28*	76				
3	0.19*	146										
4	0.28	108					0.33	86				
5	0.01	281										
6	0.35*	0					0.45*	0				
7												
8	0.38*	0					0.46*	0				
9			0.36	0					0.38	0		
10			0.38	11					0.42	0		
11			0.35	22					0.37	22		
12			0.26*	71					0.29*	22		
13			0.3	43					0.33	0		
14			0.36	6					0.41	26		
15			0.31	114					0.34	72		
16			0.12*	152								
17			0.35	7					0.39	0		
18			0.30*	23					0.33*	42		
19			0.24	48								
20			0.25*	59								
21			0.37*	42					0.40*	9		
22					0.2	66					0.26	0
23					0.23	40					0.28	40
24					0.29	11					0.35	0
25					0.08*	174						
26					0.34*	25					0.41*	0

Table C.15 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.31*	0					0.36*	30
28					0.30*	31					0.33*	0
29					0.16*	94						
30			0.27*	37					0.29*	0		
Total H	0.26		0.3		0.25		0.4		0.36		0.33	
α	0.68		0.84		0.68		0.74		0.84		0.71	
α (Ord)	0.7		0.87		0.73		0.78		0.87		0.77	
ω (Ord)	0.3		0.86		0.7		0.73		0.86		0.76	
LCRC†	0.72 (7)		0.85 (14)		0.71 (8)		0.75 (5)		0.85 (11)		0.73 (6)	
N	32647		32647		32647		32647		32647		32647	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Slovakia

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.37*	45					0.40*	15				
2	0.24*	35					0.26*	54				
3	0.25*	38					0.28*	14				
4	0.36	11					0.38	32				
5	0.1	170										
6	0.39*	10					0.43*	0				
7												
8	0.23	68										
9	0.26	28					0.3	43				
10												
11							0.23	60				
12	0.25	28					0.26	25				
13			0.28	10					0.35	0		
14			0.22	18					0.31	0		
15			0.31	45					0.42	26		
16			0.19*	64								
17			0.15	33								
18			0.3	28					0.41	0		
19			0.09	123								
20			0.22	29					0.32	17		
21			0.12*	73								
22					0.24	24					0.33	40
23					0.12	94						
24					0.14	77						
25					0.09*	61						
26					0.33*	0					0.46*	0

Table C.16 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.26*	15					0.38*	10
28					0.29*	12					0.35*	0
29					0.23	33					0.33	47
30												
Total H	0.27		0.21		0.22		0.32		0.37		0.37	
α	0.74		0.67		0.65		0.75		0.71		0.7	
α (Ord)	0.78		0.72		0.67		0.81		0.76		0.75	
ω (Ord)	0.71		0.69		0.57		0.78		0.72		0.77	
LCRC†	0.75 (6)		0.69 (5)		0.72 (7)		0.77 (6)		0.72 (5)		0.76 (5)	
N	4689		4689		4689		4689		4689		4689	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

United Kingdom (England)

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.57*	0					0.61*	0				
2	0.46*	21					0.50*	0				
3	0.58*	0					0.63*	0				
4	0.49	0					0.53	48				
5	0.48	24					0.52	0				
6	0.63*	0					0.68*	0				
7												
8	0.22	272										
9	0.66	0					0.71	0				
10	0.57*	74					0.62*	72				
11			0.37	23								
12			0.4	0					0.5	0		
13			0.36	0					0.45	0		
14			0.37*	30					0.45*	0		
15			0.39	0					0.48	0		
16			0.36	10					0.43	0		
17			0.01	597								
18			0.11*	340								
19			0.39	0					0.47	0		
20			0.34	13								
21					0.49	0					0.61	0
22					0.27	145						
23					0.34	91						
24					0.34*	87					0.41*	54
25					0.42*	8					0.49*	0
26					0.27*	199						

Table C.17 continued from previous page

	Ex Ante						Final					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.31*	44						
28					0.45	0					0.58	0
29					0.46*	0					0.58*	0
30					0.47	0					0.57	0
Total H	0.52		0.32		0.38		0.6		0.46		0.54	
α	0.89		0.81		0.83		0.91		0.82		0.83	
α (Ord)	0.91		0.82		0.86		0.93		0.85		0.88	
ω (Ord)	0.91		0.73		0.85		0.93		0.81		0.87	
LCRC†	0.91 (9)		0.83 (10)		0.85 (10)		0.92 (8)		0.82 (6)		0.84 (6)	
N	61435		61435		61435		61435		61435		61435	

∗ Items with an asterisk (∗) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

D | Quasi-Inductive Scales

Austria

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.53*	4					0.51*	10				
2	0.36*	14					0.32*	36				
3	0.42*	11					0.49*	14				
4	0.42	13					0.37	42				
5	0.36	36					0.42	40				
6	0.53*	0					0.52*	3				
7												
8	0.30	19										
9	0.28	54										
10							0.43*	37				
11	0.33*	16					0.31*	72				
12			0.32	16					0.38*	40		
13			0.24	28								
14			0.35	8					0.42*	0		
15			0.28*	33					0.36	8		
16			0.27	14					0.34*	10		
17			0.30	14					0.33*	36		
18			0.11	156								
19			0.28	20					0.35*	0		
20			0.17	33								
21					0.37	24	0.43*	33				
22					0.24	27					0.38*	34
23					0.33	37					0.42*	7
24					0.17*	66						
25					0.36*	7	0.34	40				
26					0.28*	15						

Table D.1 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.29*	17					0.32	0
28							0.4	32				
29					0.33	9	0.46*	15				
30												
Total H	0.39		0.26		0.30		0.41		0.37		0.37	
α	0.82		0.73		0.74		0.87		0.75		0.59	
α (Ord)	0.86		0.77		0.78		0.9		0.79		0.67	
ω (Ord)	0.85		0.68		0.77		0.92		0.72		0.64	
LCRC†	0.83 (6)		0.74 (6)		0.77 (8)		0.89 (9)		0.75 (6)		0.58 (3)	
N	5421		5421		5421		5421		5421		5421	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Croatia

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.36*	0							0.46	0		
2	0.29*	58							0.36	23		
3	0.29*	38					0.32*	24				
4	0.3	61							0.36*	45		
5	0.02	238										
6	0.39*	29							0.48	0		
7												
8			0.20	47								
9			0.21	0							0.49	17
10			0.20	60								
11			0.15*	95								
12			0.16	49								
13			0.22	39							0.34	42
14			0.13	72								
15			0.09*	97								
16			0.16*	60								
17			0.11	66								
18			0.21	47							0.51	1
19			-0.01*	349								
20												
21					0.29	39	0.36*	34				
22					0.16	105						
23					0.12	150						
24					0.14*	89						
25					0.46*	0	0.56	0				
26					0.39*	41	0.48	2				

Table D.2 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.35*	36	0.38	20				
28					0.43*	0	0.52	26				
29					0.36	10	0.48	13				
30												
Total H	0.28		0.15		0.32		0.45		0.41		0.45	
α	0.68		0.65		0.77		0.83		0.71		0.65	
α (Ord)	0.70		0.69		0.79		0.86		0.75		0.70	
ω (Ord)	0.68		0.63		0.71		0.85		0.73		0.72	
LCRC†	0.70 (4)		0.67 (8)		0.82 (7)		0.85 (5)		0.72 (4)		0.67 (3)	
N	4126		4126		4126		4126		4126		4126	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Czech Republic

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU‡		Economic‡	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.5*	58					0.47*	77		
2	0.43*	6					0.4*	24		
3	0.32*	170								
4	0.46	0					0.43	24		
5	0.32	80					0.32	44		
6	0.57*	0					0.55*	0		
7										
8	0.34	12					0.34	26		
9	0.33	91					0.32	42		
10	0.54*	0					0.52*	0		
11			0.3	35					0.33*	59
12			0.29	39					0.36*	37
13			0.37	39					0.47*	0
14			0.28*	56					0.35	10
15			0.2	99						
16			0.37	9					0.44*	0
17			0	526			0.32	63		
18			0.29*	112						
19			0.33	12					0.39*	0
20					0.25	29	0.3*	17		
21					0.16	51				
22					0.19	74				
23					0.1*	147				
24					0.25*	30				
25					0.16*	118				
26					0.22*	38				

Table D.3 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU‡		Economic‡	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.25	29				
28							0.32*	23		
29										
30										
Total H	0.43		0.27		0.2		0.4		0.39	
α	0.85		0.75		0.62		0.86		0.77	
α (Ord)	0.87		0.77		0.65		0.88		0.81	
ω (Ord)	0.87		0.61		0.26		0.9		0.75	
LCRC†	0.87 (9)		0.78 (9)		0.68 (8)		0.88 (11)		0.77 (6)	
N	17833		17833		17833		17833		17833	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.
‡ Item 18 was dropped from the EU scale, and Item 3$*$ was dropped from the Economic scale because of crit values > 80.

Denmark

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.26*	31					0.35*	9		
2	0.24*	80					0.32*	27		
3	0.31*	46					0.45*	19		
4	0.25	18					0.31	12		
5	0.23	75					0.31	38		
6	0.34*	47					0.46*	27		
7										
8	-0.46	1887					0.48*	0		
9	0.28	33					0.4	0		
10	0.29*	55					0.44*	0		
11	0.25	33					0.32	0		
12	0.06	234								
13	0.29	42					0.39	38		
14			0.35	0					0.42	0
15			0.4	0					0.46	0
16			0.32	82					0.41	34
17			0.28*	0					0.36*	0
18			0.39	25					0.45	0
19			0.24	32						
20			0.07	298						
21									0.35	0
22										
23					0.38	0	0.37*	14		
24					0.23	76				
25					0.26	48				
26					0.2*	165				

Table D.4 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.27*	0				
28					0.19*	81				
29					0.23*	55				
30					0.35	42	0.4*	23		
Total H	0.2		0.3		0.27		0.39		0.41	
α	0.72		0.73		0.68		0.88		0.79	
α (Ord)	0.75		0.76		0.74		0.9		0.82	
ω (Ord)	0.8		0.69		0.68		0.92		0.8	
LCRC†	0.81 (12)		0.76 (7)		0.73 (8)		0.89 (13)		0.8 (6)	
N	55832		55832		55832		55832		55832	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Estonia

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.33*	9					0.46	35				
2	0.21*	61					0.34	0				
3	0.22*	22										
4	0.23	67					0.31*	31				
5	0.05	241							0.31	42		
6	0.34*	0					0.46	30				
7							0.32*	57				
8	0.10	116										
9	0.17	71									0.39	23
10			0.16	75								
11			0.12	73								
12			0.20	0								
13			0.20*	16					0.55	0		
14			0.10	63								
15			0.20	24								
16			0.06	139								
17			0.05*	140								
18												
19			0.20*	23					0.44	20		
20					0.16	78					0.34	31
21					0.13	89						
22					0.17	58						
23					0.10*	111						
24					0.15*	59						
25					0.08*	143						
26					0.15*	70						

Table D.5 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.10*	56					0.36*	42
28												
29												
30												
Total H	0.20		0.15		0.13		0.38		0.44		0.36	
α	0.63		0.57		0.49		0.69		0.65		0.54	
α (Ord)	0.68		0.60		0.54		0.76		0.70		0.64	
ω (Ord)	0.58		0.42		0.27		0.73		0.65		0.52	
LCRC†	0.65 (7)		0.63 (9)		0.54 (7)		0.69 (5)		0.67 (3)		0.55 (3)	
N	7953		7953		7953		7953		7953		7953	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Finland

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Economic		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.50*	-2							0.48	3		
2	0.40*	9							0.37	49		
3	0.43*	0							0.44	5		
4	0.4	17							0.38*	19		
5	0.25	67					0.31*	21				
6	0.53*	-2							0.52	7		
7									0.33*	42		
8												
9							0.35*	52				
10												
11												
12			0.36	38			0.30	40				
13			0.37	14			0.36	13				
14			0.34	12			0.34	40				
15			0.33*	37			0.37*	47				
16			0.39	30			0.36	16				
17			0.35	17			0.30	53				
18			0.12	152								
19							0.35*	24				
20												
21												
22					0.28	53			0.35	43		
23					0.25	31					0.40	33
24					0.31	35					0.39	0
25					0.19*	35						
26					0.34*	11			0.32*	25		

Table D.6 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Economic		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.23*	27						
28					0.29*	0					0.36*	17
29												
30							0.30*	44				
Total H	0.42		0.33		0.28		0.33		0.4		0.39	
α	0.79		0.75		0.66		0.82		0.81		0.61	
α (Ord)	0.82		0.78		0.72		0.85		0.85		0.69	
ω (Ord)	0.83		0.76		0.69		0.84		0.87		0.65	
LCRC†	0.8 (4)		0.77 (7)		0.72 (6)		0.83 (8)		0.83 (6)		0.6 (3)	
N	4664		4664		4664		4664		4664		4664	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

France

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		Economic		EU	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.49*	0							0.49	5
2	0.32*	38							0.33	32
3	0.38*	40					0.33	65		
4	0.39	47							0.38*	9
5	0.36	43					0.35*	52		
6	0.50*	6							0.49	0
7										
8	0.25	95								
9	0.29	74							0.32*	50
10									0.36*	17
11			0.36	34			0.30	59		
12			0.46	5			0.41	51		
13			0.41	23			0.36	47		
14			0.40*	17			0.40*	26		
15			0.44	29			0.39	75		
16			0.43	31			0.36	69		
17			0.17	187						
18			0.39*	19			0.40*	60		
19			0.44	24			0.41	45		
20										
21					0.49	4	0.43	43		
22					0.31	97				
23					0.34	18	0.35	57		
24					0.27*	52				
25					0.43*	32			0.34*	21
26					0.34*	46			0.30*	23

Table D.7 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		Economic		EU	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.34*	55				
28					0.46	6	0.40	46		
29					0.52	8	0.47	15		
30					0.43	24	0.42	28		
Total H	0.37		0.39		0.40		0.39		0.38	
α	0.81		0.83		0.83		0.89		0.81	
α (Ord)	0.83		0.86		0.87		0.91		0.84	
ω (Ord)	0.86		0.80		0.87		0.92		0.89	
LCRC†	0.82 (6)		0.84 (6)		0.85 (7)		0.90 (10)		0.83 (7)	
N	5268		5268		5268		5268		5268	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Germany

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU‡		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.53*	0					0.52	20				
2	0.35*	51					0.32	68				
3	0.45*	7					0.49	34				
4	0.45	14					0.43*	51				
5	0.37	47					0.43*	57				
6	0.54*	-1					0.52	14				
7											0.41	0
8	0.32	55					0.32*	51				
9	0.29	64									0.32	29
10												
11							0.41*	27				
12												
13			0.35	9					0.38	13		
14			0.27	12					0.32	28		
15			0.38	12					0.43	0		
16			0.28*	29					0.32*	15		
17			0.3	36					0.34	33		
18			0.32	12					0.35	34		
19			0.08	191							0.33	25
20												
21							0.39*	45				
22					0.34	19	0.44	44				
23					0.27	22						
24					0.26	61						
25					0.16*	65						
26					0.33*	14	0.40*	39				

Table D.8 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU‡		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.26*	50						
28					0.30*	17						
29											0.37*	0
30											0.40*	0
Total H	0.41		0.28		0.28		0.43		0.36		0.37	
α	0.83		0.71		0.69		0.87		0.74		0.71	
α (Ord)	0.85		0.74		0.74		0.89		0.78		0.76	
ω (Ord)	0.86		0.58		0.7		0.91		0.74		0.77	
LCRC†	0.85 (8)		0.73 (6)		0.71 (6)		0.88 (9)		0.75 (6)		0.74 (5)	
N	5447		5447		5447		5447		5447		5447	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.
‡ Item 12 was dropped from the EU scale because of a crit value > 80.

Greece

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.39*	5					0.42*	7		
2	0.27*	89								
3	0.24*	102							0.38*	34
4	0.33	29								
5	-0.06	453								
6	0.39*	32					0.39*	24		
7							0.32	44		
8	0.22	74								
9										
10			0.43	0			0.44	0		
11			0.44	3			0.44	0		
12			0.41	0			0.39	22		
13			0.37	12			0.35	27		
14			0.28*	31						
15			0.33	13			0.31	41		
16			0.40	0			0.41	4		
17			0.28	53			0.32	53		
18			0.36	0			0.35	43		
19			0.38	10			0.40	0		
20			0.28	50						
21					0.34	301				
22					0.28	139	0.33	43		
23					0.24	127	0.45	3		
24					0.17*	219				
25					0.42*	0			0.52	0
26					0.31*	65			0.38	36

Table D.9 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.33*	58			0.39	28
28					0.41*	22			0.49	19
29					0.43	0			0.48*	1
30					0.41*	27			0.51	7
Total H	0.27		0.36		0.34		0.38		0.45	
α	0.69		0.84		0.81		0.88		0.83	
α (Ord)	0.69		0.87		0.84		0.90		0.86	
ω (Ord)	0.52		0.86		0.87		0.92		0.83	
LCRC†	0.74 (7)		0.85 (11)		0.84 (10)		0.89 (14)		0.84 (7)	
N	36967		36967		36967		36967		36967	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Hungary

	Ex Ante						Quasi Inductive	
	EU		Economic		Cultural		EU‡	
#	H	crit	H	crit	H	crit	H	crit
1	0.53*	21					0.54*	4
2	0.43*	41					0.45*	39
3	0.46*	9					0.49*	33
4	0.50	24					0.54	18
5	-0.10	390						
6	0.55*	2					0.59*	-1
7								
8	0.31	20					0.33	72
9	0.52*	0					0.54*	0
10								
11							0.47	51
12			0.30	0			0.42	42
13			0.23	45				
14			0.31	11			0.42	65
15			-0.08*	381				
16			0.14	73				
17			0.19	27				
18			0.25	64			0.55	9
19								
20								
21								
22					0.35	31	0.43*	37
23					0.27	25		
24					0.21	91		
25					0.18*	104		
26					0.41*	5	0.54	13

Table D.10 continued from previous page

	Ex Ante						Quasi Inductive	
	EU		Economic		Cultural		EU‡	
#	H	crit	H	crit	H	crit	H	crit
27					0.35*	8	0.43	54
28					0.31*	35		
29					0.19	63		
30								
Total H	0.42		0.20		0.29		0.48	
α	0.83		0.60		0.73		0.92	
α (Ord)	0.84		0.62		0.76		0.93	
ω (Ord)	0.85		0.51		0.72		0.93	
LCRC†	0.86 (6)		0.67 (5)		0.76 (7)		0.92 (7)	
N	3616		3616		3616		3616	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.
‡ Item 28 was dropped from the EU scale because of a crit value > 80.

Ireland

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Economic		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.34*	7							0.40	11		
2	0.23*	24							0.30	31		
3	0.26*	39									0.32*	0
4	0.29	42							0.32*	35		
5	0.14	138					0.32*	45				
6	0.35*	0							0.43	6		
7									0.34*	31		
8	0.17	47										
9	0.17	56										
10			0.32	17			0.41	10				
11			0.29	19			0.40	25				
12			0.27	59			0.37	16				
13			0.30*	36			0.41*	0				
14			0.28	36			0.37	12				
15			0.29	40			0.36	17				
16			0.22	121								
17			0.17*	70								
18			0.17	70					0.36*	38		
19									0.36	0		
20			0.25*	26								
21			0.05	243								
22			0.31	20			0.37	33				
23					0.24	17					0.34*	9
24					0.23	60						
25					0.26	29						
26					0.18*	52						

Table D.11 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		Economic		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.34*	6					0.41	0
28					0.27*	20					0.30	12
29					0.28*	20						
30					0.26*	62					0.33	28
Total H	0.24		0.24		0.26		0.38		0.36		0.34	
α	0.69		0.77		0.69		0.80		0.75		0.68	
ω (Ord)	0.73		0.80		0.76		0.84		0.80		0.75	
ω (Ord)	0.69		0.80		0.78		0.83		0.82		0.75	
LCRC†	0.7 (6)		0.78 (8)		0.73 (7)		0.81 (7)		0.78 (7)		0.71 (5)	
N	5325		5325		5325		5325		5325		5325	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

Italy

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.45*	0					0.41	7				
2	0.30*	75										
3	0.37*	15					0.40	24				
4	0.36	59										
5	0.07	203										
6	0.42*	0					0.40	5				
7												
8			0.26	17					0.32	0		
9			0.22	13					0.36	0		
10			0.25	33					0.34	27		
11			0.25*	38								
12			0.23	15					0.34	0		
13			0.26	17					0.32	10		
14			0.14	137								
15			0.17*	76								
16			0.13*	123								
17							0.31	76				
18			0.18*	44								
19					0.37	51	0.41	0				
20					0.18	60						
21					0.32	48					0.31*	29
22					0.16*	103						
23					0.41*	0	0.40*	0				
24					0.33*	21	0.31*	16			0.38	26
25					0.33*	30					0.40	5
26					0.39	0	0.43	29				

Table D.12 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.36*	32	0.41*	12				
28												
29												
30												
Total H	0.33		0.21		0.32		0.39		0.33		0.36	
α	0.72		0.70		0.78		0.82		0.69		0.59	
α (Ord)	0.75		0.74		0.82		0.86		0.73		0.66	
ω (Ord)	0.81		0.68		0.82		0.89		0.69		0.66	
LCRC†	0.76 (6)		0.73 (10)		0.81 (9)		0.85 (9)		0.68 (5)		0.61 (3)	
N	20068		20068		20068		20068		20068		20068	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Lithuania

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.38*	0					0.46	0		
2	0.25*	56								
3	0.22*	25								
4	0.27	61								
5	0.02	181								
6	0.40*	4					0.48	0		
7							0.35*	27		
8	0.23	28					0.32*	47		
9	0.37*	0					0.42	27		
10			0.23	34					0.31*	37
11			0.14	78						
12			0.24	14					0.36*	0
13			0.19*	25						
14			0.10	73						
15			0.22	24					0.32*	0
16			0.07	140						
17			0.11*	53						
18							0.32*	34		
19			0.21*	53			0.44	9		
20							0.37*	8		
21					0.20	46				
22					0.18	66				
23					0.19	28				
24					0.08*	134				
25					0.27*	0	0.31*	38		
26					0.22*	35				

Table D.13 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.25*	12				
28					0.07*	95				
29										
30										
Total H	0.27		0.17		0.19		0.38		0.33	
α	0.71		0.61		0.60		0.81		0.55	
α (Ord)	0.76		0.65		0.64		0.87		0.60	
ω (Ord)	0.69		0.54		0.53		0.87		0.57	
LCRC†	0.73 (6)		0.63 (5)		0.63 (5)		0.82 (7)		0.54 (3)	
N	5317		5317		5317		5317		5317	

∗ Items with an asterisk (∗) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Poland

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU‡		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.45*	7					0.44*	0		
2	0.40*	84					0.42*	31		
3	0.28*	314								
4	0.48	53					0.47	0		
5	0.34	115							0.45	0
6	0.52*	26					0.52*	0		
7										
8	0.28	119								
9			0.29	28					0.39*	0
10			0.26	16					0.34*	35
11			0.3	0					0.42*	19
12			0.24*	148					0.44	0
13			0.19	116						
14			0.29	42					0.42*	0
15			-0.21	885						
16			0.14*	158						
17					0.13	70	0.32*	67		
18					0.01	340				
19					-0.01	493				
20					0.07*	144				
21					0.32*	25	0.52	0		
22					0.31*	31	0.46	28		
23					0.21*	48				
24					0.04*	169				
25									0.34	49
26										

Table D.14 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU‡		Economic	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.32*	28	0.46	22		
28										
29					0.30*	36				
30					0.27*	65	0.53	0		
Total H	0.4		0.19		0.19		0.46		0.4	
α	0.79		0.62		0.69		0.87		0.8	
α (Ord)	0.82		0.68		0.7		0.89		0.84	
ω (Ord)	0.79		0.46		0.39		0.91		0.84	
LCRC†	0.81 (7)		0.69 (8)		0.77 (11)		0.89 (9)		0.81 (7)	
N	41554		41554		41554		41554		41554	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.
‡ Items 3$*$,15, and 29 were dropped from the EU scale because of crit values > 80.

Portugal

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.34*	0					0.34	16		
2	0.22*	92								
3	0.19*	146								
4	0.28	108								
5	0.01	281								
6	0.35*	0					0.31	40		
7										
8	0.38*	0								
9			0.36	0			0.37*	0		
10			0.38	11			0.39*	0		
11			0.35	22			0.33*	65		
12			0.26*	71						
13			0.30	43						
14			0.36	6			0.42*	0		
15			0.31	114			0.37*	38		
16			0.12*	152						
17			0.35	7			0.40*	22		
18			0.30*	23			0.35	51		
19			0.24	48						
20			0.25*	59						
21			0.37*	42			0.42	8		
22					0.20	66				
23					0.23	40				
24					0.29	11	0.34*	26		
25					0.08*	174				
26					0.34*	25			0.51*	0

Table D.15 continued from previous page

	Ex Ante						Quasi Inductive			
	EU		Economic		Cultural		EU		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.31*	0			0.49*	0
28					0.30*	31			0.43*	24
29					0.16*	94				
30			0.27*	37						
Total H	0.26		0.30		0.25		0.37		0.48	
α	0.68		0.84		0.68		0.84		0.7	
α (Ord)	0.7		0.87		0.73		0.87		0.77	
ω (Ord)	0.3		0.86		0.7		0.87		0.74	
LCRC†	0.72 (7)		0.85 (14)		0.72 (8)		0.85 (11)		0.69 (3)	
N	32647		32647		32647		32647		32647	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

Slovakia

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
1	0.37*	45					0.38	11				
2	0.24*	35										
3	0.25*	38					0.35	18				
4	0.36	11					0.33*	34				
5	0.1	170										
6	0.39*	10					0.42	22				
7												
8	0.23	68										
9	0.26	28					0.31*	37				
10												
11												
12	0.25	28										
13			0.28	10					0.35	0		
14			0.22	18					0.31	0		
15			0.31	45					0.42	26		
16			0.19*	64								
17			0.15	33								
18			0.3	28					0.41	0		
19			0.09	123								
20			0.22	29					0.32	17		
21			0.12*	73							0.35	0
22					0.24	24	0.34	19				
23					0.12	94						
24					0.14	77						
25					0.09*	61						
26					0.33*	0	0.33*	6				

Table D.16 continued from previous page

	Ex Ante						Quasi Inductive					
	EU		Economic		Cultural		EU		Economic		Cultural	
#	H	crit	H	crit	H	crit	H	crit	H	crit	H	crit
27					0.26*	15					0.40*	0
28					0.29*	12					0.34*	0
29					0.23	33	0.31	67				
30							0.37	64				
Total H	0.27		0.21		0.22		0.35		0.37		0.36	
α	0.74		0.67		0.65		0.78		0.71		0.6	
α (Ord)	0.78		0.72		0.67		0.83		0.76		0.67	
ω (Ord)	0.71		0.69		0.57		0.85		0.68		0.6	
LCRC†	0.75 (6)		0.69 (5)		0.72 (7)		0.81 (7)		0.71 (5)		0.6 (3)	
N	4689		4689		4689		4689		4689		4689	

$*$ Items with an asterisk ($*$) were reversed for analysis.

† The Bayesian Information Criterion was used to determine the number of latent classes.

United Kingdom (England)

	Ex Ante						Quasi Inductive	
	EU		Economic		Cultural		EU	
#	H	crit	H	crit	H	crit	H	crit
1	0.57*	0					0.50	46
2	0.46*	21					0.41	69
3	0.58*	0					0.55	14
4	0.49	0					0.44*	4
5	0.48	24					0.48*	0
6	0.63*	0					0.55	39
7								
8	0.22	272						
9	0.66	0					0.56*	55
10	0.57*	74					0.50	22
11			0.37	23				
12			0.40	0			0.43	7
13			0.36	0			0.38	47
14			0.37*	30			0.45*	39
15			0.39	0			0.41	50
16			0.36	10			0.33	59
17			0.01	597				
18			0.11*	340				
19			0.39	0			0.40	69
20			0.34	13			0.34	65
21					0.49	0	0.52	18
22					0.27	145		
23					0.34	91		
24					0.34*	87		
25					0.42*	8	0.42*	48
26					0.27*	199		

Table D.17 continued from previous page

	Ex Ante						Quasi Inductive	
	EU		Economic		Cultural		EU	
#	H	crit	H	crit	H	crit	H	crit
27					0.31*	44		
28					0.45	0	0.49	55
29					0.46*	0	0.51*	25
30					0.47	0	0.54	23
Total H	0.52		0.32		0.38		0.46	
α	0.89		0.81		0.83		0.94	
α (Ord)	0.91		0.82		0.86		0.95	
ω (Ord)	0.91		0.73		0.85		0.97	
LCRC†	0.91 (9)		0.83 (10)		0.85 (10)		0.94 (20)	
N	61435		61435		61435		61435	

$*$ Items with an asterisk ($*$) were reversed for analysis.
† The Bayesian Information Criterion was used to determine the number of latent classes.

E | Visualization Data

Responses for the open-ended question "You will now see a plot with three points. One point represents you, the three other points represent three political parties. How would you describe your position in the graph?" Responses are categorised by the four categories set out in the paper.

E.0.1 Whole Graph - Relative Qualifiers

- on the right
- Upper right
- right of center
- near center of top right
- In the upper right quadrant in between the origin and point B
- I am in the upper right quadrant, one unit to the left and two units down from B.
- more central
- Somewhere in the middle
- My position is in the middle of the other 6.
- I'm in the middle on the Y axis
- close to the middle
- Fairly centrist in both ways, not in an extreme position.
- nearly equal distance from each political party
- my position fairly neutral.
- my position is the nearest near plot A and B between y-axis and x-axis line graph
- Positive on both axis
- On the positive side of both the x and y axes, not quite to point B, and well away from points A, C and D.
- Near to half of the peoples views
- I am on to the positive side on both x and y axis
- it is between x and y
- I am moderately for X and Y.

- somewhat close
- +

E.0.2 Whole Graph - Absolute Qualifiers

- two points away in the positive side of X and Y axis from the origin
- line 8 in y line 4 in x
- 2,2
- y2,x2
- I am located at (2,2).
- My Position is at X2,Y2
- 2,2
- I am at (2, 2) on the coordinate plane
- (2, 2)
- 2 points above x and 2 points right of y
- 2,2

E.0.3 Point B - Relative Qualifiers

- Closest to B
- Below the B party
- My position is closest to B.
- nearest to B
- B
- close to B
- Closest to B.
- closest to point B
- I am closest to party B and farthest to party D
- Closest to B
- B
- closest to D middle of the road
- Very similar to B, pretty dissimilar to A, D, and C.
- CLOSET TO B
- B
- b is close
- Closest to B

- more in line with B but not the same
- I am somewhat close to B, and I'm somewhat far away from A and D and quite far away from C
- Some what closer
- near b
- b
- closer to A
- C
- similar d

E.0.4 Point B - Absolute Qualifiers

- 2 points from b vertically, -1 point from y horizontally
- I am 2 Y and 2 X and closest to the B.
- Two blocks below B
- I am two points away from B
- I am 2 bars down and one over from B