Quality & Quantity (2020) 54:1299–1316
https://doi.org/10.1007/s11135-020-00986-8

A comparison of measures to validate scales in voting advice applications

Bastiaan Bruinsma[1]

Published online: 29 April 2020

Abstract
Voting advice applications (VAAs) are online tools providing voting advice to their users. This voting advice is based on the match between the answers of the user and the answers of several political parties to a common questionnaire on political attitudes. To visualize this match, VAAs use a wide array of visualisations, most popular of which are the two-dimensional political maps. These maps show the position of both the political parties and the user in the political landscape, allowing the user to understand both their own position and their relation to the political parties. To construct these maps, VAAs require scales that represent the main underlying dimensions of the political space. This makes the correct construction of these scales important if the VAA aims to provide accurate and helpful voting advice. This paper presents three criteria that assess if a VAA achieves this aim. To illustrate their usefulness, these three criteria—unidimensionality, reliability and quality—are used to assess the scales in the cross-national EUVox VAA, a VAA designed for the European Parliament elections of 2014. Using techniques from Mokken scaling analysis and categorical principal component analysis to capture the metrics, I find that most scales show low unidimensionality and reliability. Moreover, even while designers can—and sometimes do—use certain techniques to improve their scales, these improvements are rarely enough to overcome all of the problems regarding unidimensionality, reliability and quality. This leaves certain problems for the designers of VAAs and designers of similar type online surveys.

Keywords Voting advice applications · Scales · Data quality

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11135-020-00986-8) contains supplementary material, which is available to authorized users.

✉ Bastiaan Bruinsma
bruinsma@soz.uni-frankfurt.de

[1] Institute of Political Science, Goethe University Frankfurt am Main, Frankfurt am Main, Germany

1 Introduction

Over the last years, an increasing number of voters have turned to voting advice applications (VAAs). These online platforms match the responses of their users and those of various political parties to a common set of questions about various political issues and attitudes. Of the various visualisations VAAs use to visualise this match, one of the most popular is the political map. Here, some of the questions are collapsed into scales, which are then used to construct a two-dimensional political space in which. As both users and parties are positioned in this space, this allows the user to not only see their position but those of the political parties as well.

For this visualisation to be accurate and useful, the scales underlying the dimensions need to be correct. Yet, VAA scales more often than not score low on basic criteria for scale construction (Louwerse and Otjes 2012; Gemenis 2013). This is worrying, as defective scales might lead to political maps that are not only incorrect and misleading but also hard to understand. In the worst case, this might confuse the user instead of helping them. To avoid this, VAA designers have made several suggestions on how to improve these scales. The most prominent of these is the idea of *dynamic scale validation* (DSV), first introduced by Germann et al. (2015). Here, VAA designers use the responses of an early group of users to confirm—or if necessary alter—the original scales.

In this paper, I expand on the idea of dynamic scale validation. I do so by looking at three criteria useful for scale validation: unidimensionality, reliability, and quality. For unidimensionality, I will use Loevinger's coefficient of homogeneity H (Loevinger 1947, 1948), for reliability the Latent Class Reliability Coefficient (LCRC) (van der Ark 2012), and for quality the *dirty data index* (DDI) as proposed by Blasius and Thiessen (2012). For each of these, I will provide a rationale for why they are required, and what they can tell about the quality of the scales.

From here, the paper will proceed as follows. First, I will discuss how VAAs generate scales and how they use them. Then, I will briefly discuss the concepts of unidimensionality, reliability, and quality and how we can measure them. I will use these concepts to assess a series of scales derived from the EUVox VAA, that was launched for the European Parliament elections of 2014. I conclude with a brief reflection on the use of the criteria and how VAA designers can use them to build more accurate scales.

2 Scales in voting advice applications

In 2006, *Kieskompas* was the first VAA to use scales to position users and parties on a political map. Before this, VAAs showed their matches as mere percentages, most often in the form of ordered bar plots. Yet, while intuitive, such plots tend to include a piece of rather explicit voting advice: that of the party on top (Germann and Mendez 2016). Also, they do not distinguish between different issues but only add them together. Thus, the content of the issues does not affect the visualisation of the result. To see this, and on which issue they agreed or disagreed with the parties, a user has to compare their responses to those of each party for each of the issues. The political map addresses all these problems by combining the issues into two or more meaningful dimensions. Also, by using them to construct a two-dimensional space, they can show both the distance between the user and a party as well as the direction of that distance (Fossen et al. 2012).

This idea to combine issues into dimensions is part of a long tradition of spatial modelling ranging back to Downs (1957). At the core of this tradition is the idea that voters and parties do not choose their positions on issues at random. Instead, their position on one or more broader, underlying dimensions, determines their positions. Thus, to represent these dimensions, VAA designers combine several of the issues into scales. Which issue belongs to which scale depends either on theory or on data. Thus, an issue on taxes can belong to an economic left-right scale *in theory*, while an issue on abortion can belong to a conservative-progressive scale, again *in theory*. Yet, while using theory is simple and convenient, scales constructed this way are not without problems. Most often, they lack unidimensionality, which means that they represent more than a single dimension at once (Louwerse and Otjes 2012; Gemenis 2013; Otjes and Louwerse 2014). This not only makes the scale difficult to interpret but also gives a false idea of what the political space looks like.

Thus VAA designers have lately turned to data to construct their scales. In the beginning, they often did so by using a combination of Exploratory (EFA) and Confirmatory Factor Analysis (CFA) (Wheatley 2015b, a), as these methods were well-understood and produced helpful goodness-of-fit indexes. Yet, as both methods assume metric data while VAA data are often ordinal, this often led to an overestimation of the number of dimensions. To address this, they turned to IRT (Item Response Theory) models, such as Mokken Scaling Analysis (MSA) as these do allow for ordinal data (Katsanidou and Otjes 2016; Mendez and Wheatley 2014; Wheatley 2016; Wheatley et al. 2014; Wheatley and Mendez n.d.). Moreover, apart from constructing new scales, these models allowed VAA designers to validate any theoretically constructed scales as well. It was therefore that Germann et al. (2015); Germann and Mendez (2016) suggested MSA a way to validate theoretically constructed scales in VAAs, and construct new scales based on early user data if needed.

Yet, basing the scales on user data has its problems. To begin with, it assumes that scales derived from the responses of users are better than scales constructed by VAA designers. Also, as this process requires user data, this means that VAA designers have to wait until after they have taken the VAA offline. And while they might choose to collect data earlier and only use data of a set of early users, this leads to later users receiving advice based on different scales. It also implies that early users are more likely to have inaccurate results, and receive advice different from a later user, even though they might have the same responses to each of the items.

Yet, we must remember that while it might be uncertain to base scales solely on the responses of voters, it is as uncertain to base them solely on theory. In other words: users are no less correct or incorrect than VAA designers. Thus, the adequacy of VAA scales should depend both on theoretical and empirical criteria (Carmines and Zeller 1979). Here, the theoretical criteria are those VAA designers use to decide which items to include and in which scales to place them initially, while the empirical criteria are those decisions on whether to include or remove items from the scales based on the user responses. Also, if this changes the scales, this is conceptually reasonable as the goal of VAAs is to match voters with parties from the voter perspective. In other words, VAAs match voters with parties, not the other way around.

Moreover, while some voters will indeed receive different advice, this is a good price to pay for more accurate scales (Germann et al. 2015; Germann and Mendez 2016). This as alternatives, such as pre-administrating the VAA questionnaire, often come with high costs. Also, with the often high number of VAA users, reaching the required number of users for validation is most likely a matter of days (or hours in some cases). And finally, as scales are only needed to construct the political maps, VAA designers could choose to

de-activate those until the scales are validated. In the meanwhile then, users could use the other visualisations—such as bar graphs or spider graphs—that most VAAs offer.

3 Criteria for scales

When validating their scales with empirical criteria, there are three that VAA designers should keep in mind. These are the unidimensionality, reliability, and quality of the scale. While the first and second of these criteria are well known, the third is less so. This is as it focuses more on the actual responses of the users to the items than the items themselves. To understand what this means, I will briefly discuss this concept, and the metric it bases itself on, together with a review of the other two criteria.

3.1 Unidimensionality

The underlying idea of any scale is that by adding the responses to all its items together, one arrives at a position on an underlying dimension. Yet, it is often so that items measure more than a single dimension. In other words: the scale is multidimensional. This is problematic, as the sum score of the scale is then difficult to interpret. So, what we want is a uni-dimensional scale: a scale that measures a single dimension and only that dimension (Gerbing and Anderson 1988). There are various methods to assess this, with the most popular methods being the earlier mentioned EFA and CFA. Here, one can use CFA to validate scales, and EFA to build new scales. Yet, apart from both requiring metric data, in EFA the number of dimensions extracted from a set of items depends on the eigenvalues. As there is some discussion at which eigenvalue a scale forms a dimension, extracting scales this way can be somewhat arbitrary. Also, EFA does not allow for correlated errors and cannot provide estimations of model fit (Bollen 1989). CFA does provide such estimations but needs to assume that the specified number of dimensions is correct.

MSA addresses this using two concepts to assess unidimensionality: homogeneity and monotonicity. Homogeneity measures how well the items belong together while monotonicity assesses if a higher score on a single item relates to a higher score on the scale as a whole. 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 respondents on the underlying dimension (Mokken et al. 1986), (b) a value H_{ij}, which indicates how well items i and j co-vary (Loevinger 1947), and (c) a value H_i that tells us how well an item co-varies with the other items in the scale (see also Appendix A). For scale construction, H and H_i are the most important. Values of both lie between 0 and 1 (Hemker et al. 1995). Here, 0 indicates there is no relation between the items at all and 1 that there is a perfect relation between them. As a rule of thumb, scales with $H < 0.30$ lack unidimensionality, while scales between $H = 0.30$ and $H = 0.40$ show weak, scales between $0.40 < H < 0.50$ show medium and scales with $H > 0.50$ show strong unidimensionality. For a Mokken scale, the H of the scale needs to be > 0.30, as do the individual H_i values for each of the items. To measure monotonicity, we can use the *restscore* of the item. This is the score that remains when we subtract the score for the item from the sum score of all items in the scale. 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 an item. From the different ways to measure this, the most useful is the *crit*-value Sijtsma and Molenaar (2002).

This value not only looks at the rest score but combines it the with characteristics of the scale. In general, any item with a *crit*-value over 80 can be said to violate monotonicity.

As well as confirming the unidimensionality of the scales like CFA, we can also use MSA to generate new scales, as EFA would. To do so, we can use two algorithms: the standard *automated item search procedure* (aisp) or the newer *genetic algorithm* (ga) (Straat et al. 2013). Here, the *aisp* starts with those two items that have the highest H_{ij} value. Then, it adds the item with the highest H_{ij} value relative to the original two items until the H value of the scale drops below a set criterion. This procedure is then repeated for the remaining items until the algorithm can form no more scales (van der Ark 2012). Yet, while quick and straightforward, *aisp* does not always lead to an optimal partition of the items into scales. This is because as soon as an item is only a little under the lower bound for the criterion, *aisp* does not consider it. Also, because the method works hierarchically, much depends on which items *aisp* chooses as its starting items. The *ga* procedure addresses this by considering all possible combinations of scales and selecting those that are the longest. Yet, this comes at the cost of the algorithm becoming time-consuming when the number of items is large (Straat et al. 2013).

3.2 Reliability

The common understanding of the reliability of a scale is that under similar circumstances, it should produce similar results. In terms of classical test theory, this means that we want the true score variation to account for as large a degree of the total variation as possible (Carmines and Zeller 1979). Yet, as we cannot observe this true score, we have to estimate it. To do so, the most established technique is that of internal consistency,. This rests on the idea that the higher the correlation between items in a scale, the more reliable the scale is. The most popular measure to establish such internal consistency is Cronbach's α. This measure compares the total variance between the items to the total variance of the scale. The higher this ratio, the more the items have in common (as they share a higher degree of variance) and thus the more reliable the scale.

While popular, using α to estimate the reliability of scales generated for VAAs is problematic. Not only does α need metric data, but it also assumes tau-equivalence. This means that each item in the scale should be equal and carry equal importance (Sijtsma 2009). Yet, both assumptions seem inappropriate for the hierarchical and ordinal data that VAAs generate. Some items might be more difficult, or less popular than others, sometimes even so by design. Thus, Germann and Mendez (2016) advise to use the Latent Class Reliability Coefficient (LCRC) (van der Ark et al. 2011) instead (see also Appendix B). What the LCRC does is correlate a set of observed categorical variables to a set of latent unobserved categorical variables using latent classes. This addresses both the problems of requiring metric data and tau-equivalence. The LCRC itself is then a value between $0-1$ that has the same interpretation as Cronbach's α, with 0.90 being a useful lower bound Germann and Mendez (2016).

3.3 Quality

Both reliability and unidimensionality focus on the relation *between* the items in the scale. Quality, on the other hand, focuses on the items themselves. The basic assumption of quality is that users respond to each item as we expect them to respond to it (Blasius and Thiessen 2012). That is, we expect them to take careful consideration of both item and response

options, and then find a proper connection between them. Yet, users often *simplify* this process by skimming over the items, using only a few of the response options or fabricate their responses altogether (Blasius and Thiessen 2012, 2015).

To measure to which degree this behaviour affects the quality of the items, Blasius and Thiessen (2012) propose a *dirty data index*. This index tests the quality of the responses to the item by looking at how metric this response is. For Likert scales, such as occur in VAAs, metric means that there are uniform distances between each of the categories. So, the distance between *completely disagree* and *disagree* is the same as the distance between *completely agree* and *agree*. Also, both are half of the distance between *completely disagree* and *either agree or disagree*. But if users do not understand the item or the response categories, violations might occur. So, they may perceive the distance between *completely disagree* and *disagree* as smaller than the distance between *disagree* and *either agree or disagree*. In the worst scenario, this might even lead to ties. Here, either the distance between the two options is too small to be different, or the order of the response categories is wrong. This happens when *either agree or disagree* is more negative than *disagree*. If any of these things happen, the data are more ordinal than they are metric, and thus of lower quality.

To calculate the DDI, we compare the PCA solution of the scale with its categorical PCA (catPCA) solution. This is possible as catPCA does not assume that the distances between the categories are equal as PCA does. Instead, catPCA calculates these distances itself. So, while in PCA the values of the response categories (1, 2, 3 …, etc.) are actual values, catPCA uses *optimal scores* to replace the original categories. The process that does so, optimal quantification, ensures that each optimal score accounts for the highest amount of variation (Linting 2007, p. 338). After the calculation of these scores, catPCA than proceeds in the same way as regular PCA. If the responses for all the items are metric, the optimal scores are the same as the PCA scores (Blasius and Thiessen 2012, pp. 133–138). In such a case, the DDI will be 0. The farther away from 0 the values are, the less equal the distances between the categories are and the less metric the responses to the items. Blasius and Thiessen (2012) consider DDI values smaller than 0.30 and 0.15 to indicate data of *good* and *exceptional* quality while values exceeding 0.5 indicate data of *bad* quality (see Appendix C for an overview of how to calculate the DDI).

4 Empirical analysis

To show the usefulness of these criteria, I will now turn to an empirical application using data from the EUVox VAA (Mendez and Manavopoulos 2018). This VAA, launched for the 2014 elections for the European Parliament, had versions for 28 countries and contained 30 items. The questionnaire itself used a 5-point Likert-type scale with an additional *no opinion* option (coded in the data-set as missing or NA). Each version of the VAA had an English version as well as a version in the main language of the country that version focused on. Of the 30 items, 21 core items occurred in all countries (20 in the case of France). Seven (six in the case of France) of these core items concerned powers of the EU. Another seven handled economic issues, and a final seven handled cultural issues. The issues and their content were such as to cover the main issues as found by the Chapel Hill Expert Survey (Polk et al. 2017). Country experts decided on the remaining issues (9 for most countries and 10 in the case of France) to capture salient topics for their countries.

Appendix E shows the English version of the 21 core questions. Of the 28 available countries, I excluded 11 from this analysis. The Netherlands and Sweden because these VAAs had different questionnaires; Belgium, Latvia, Luxembourg, and Malta because of the small number of users; and Bulgaria, Cyprus, Romania, Slovenia and Spain as not all information on them was available. Also, as the data-set for England was larger than those for Northern Ireland, Scotland and Wales, I excluded the latter three.

Table 1 shows the number of respondents for the remaining 22 countries. Here, the third column shows the entries remaining after cleaning. Cleaning data generated by VAAs is necessary as VAAs are online unsupervised questionnaires. Thus, users can fill out the survey many times, click through the items, or use the same response category for each of the items. As this can lead to a large number of nonsense answers., we should always clean VAA data (Andreadis 2014; Mendez et al. 2014). Also, we should keep in mind that although the data might be cleaner, it is still far away from a standard survey in which an interviewer asks the questions and records the responses. As such, it should be no surprise if VAA scales will perform less well than scales derived from a supervised survey.

Following the advice by Andreadis (2014) and Mendez et al. (2014), I remove all those entries that: 1) users filled out using a mobile phone (as in the mobile version the *No Opinion* option was not clearly shown), 2) returning users filled out (based on an identifier based on their IP-address), 3) where the time taken to complete all the 30 items was less than 120 seconds, 4) where the time taken to complete all the 30 items was more than 5400 seconds (90 min), 5) where the time taken to respond to any of the items 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 items using the *No Opinion* response. The number of entries remaining varies between countries and ranges between 65% (for Ireland) and 84% (for the Czech

Table 1 Number of users in the EUVox data-set

Country	Raw dataset	Clean dataset	% Clean
Austria	10,669	7527	71
Croatia	7666	52,81	69
Czech Republic	28,630	24,084	84
Denmark	126,261	92,633	73
Estonia	18,172	12,267	68
Finland	8274	6729	81
France	8704	6656	76
Germany	9658	7208	75
Greece	63,687	46,098	72
Hungary	6711	5536	82
Ireland	9523	6198	65
Italy	36,614	26,235	72
Lithuania	9072	7050	78
Poland	73,521	58,429	79
Portugal	54,165	42,199	78
Slovakia	7238	5905	82
United Kingdom*	100,897	77,403	77

*Includes only England

Appendix D shows a full overview of the cleaning procedure

Republic). Most users are removed either because they were returning users or because they used their mobile phone (see also Appendix D).

During its time in operation, EUVox used two versions of each of the economic (EC), European Union (EU) and cultural (CU) scales. The first version—the original scales—were the scales designers based on theory and used before they launched the VAA. The second version—the DSV scales—were those constructed after DSV and which were implemented after the first group of users completed the VAA. Besides these two, I will also assess a series of scales based on a quasi-inductive method (Wheatley 2015b). These are constructed similarly as the scales created by the DSV, though wherein DSV restrains the MSA to the items assigned to it in the original scale, here these restraints are removed. While this often leads to longer scales, it also means that sometimes only a single emerges. This makes the method problematic if one needs two scales to construct a political map.

4.1 Original scales

Appendix F shows the *original* scales. These scales include all the items designed for the VAA, except for the EU scale, where the *EU*7 item relating to referenda on EU treaties is left out. This item was most likely not included because the designers decided in a late-stage that the item did not fit in well with the EU dimension. Besides these items, each country adds some additional items to some of the scales, with the type and number of items differing per country. For example, where Finland uses none of the additional items, other countries sometimes use almost all, most often for the economic scale. Portugal, for example, uses 7 of the 9 additional items for the economic scale, while both Croatia and Italy use 4 extra items. Moreover, Denmark takes six additional items on the EU scale while Poland takes four items on the cultural scale. The content of these additional items is most often related to specific issue relevant to the country. Thus, in the case of the United Kingdom, the three additional items that deal with Islam, asylum seekers and immigration are on the cultural scale. The underlying idea here is that voters think about these issues like other cultural issues such as abortion.

4.2 DSV scales

Appendix G shows a second version of the scales. These scales are the result of a DSV for each of the scales. While almost all scales lost items, the precise implications were different for each of the scales. The EU dimension lost the least number of items, with items *EU*1, *EU*2, and *EU*4 remaining for all countries. This is a first sign that these items (relating to the Euro, treaty change and common foreign policy) seem to capture the underlying EU dimension well. This was not the case for item *EU*5—related to redistribution—which DSV dropped in most cases. Something similar happened to the item *CU*4 that asked about community service. DSV dropped this item in all countries except the United Kingdom. On the other hand, DSV has retained other items, such as item *CU*5 in all cases, as well as item *CU*1 (except for Estonia and Ireland).

In all cases, the resulting scales are shorter than the original ones. Here, Estonia is the most noticeable, with 4 items for the EU and economic scales, and 3 for the cultural scale. To get an idea of why this is the case, we can look at the individual *H* values. For Estonia, the H scores 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 items in all scales passed the 0.30 mark and most of the *crit* values were high. Yet, even when removing several items, the

H values for the scales were only 0.33, 0.26 and 0.23. Besides, most of the H_i values for the individual items are below 0.30, while the crit values remain high in several cases as well. Most problematic is the CU scale, with only three items, with none of them reaching the 0.3 mark and two of the three items have crit values > 80. Even in this version, the cultural scale does not represent a true scale that measures a single underlying latent dimension. An altogether different case is Hungary. Here, the EU scale scored high *H* values for both the original scale and even higher ones for the DSV scale. In this case, DSV improved the original scale by removing the worst-performing item *EU*5 and adding another.

4.3 Quasi-inductive scales

Restraining the analysis to the set of items related to either of the three scales has as the advantage that it will result in a similar number of scales. Yet, we also saw that this restriction leads to some unsatisfactory or very short scales. So, as another way to improve upon the scales, we can remove any notion of to which scale an item *should* belong. This is the quasi-inductive approach. Inductive as we build the scales based on the data, and quasi because the scales are still constrained by the number and type of items included in the VAA (Wheatley 2016). To generate these scales, I run an MSA procedure using the genetic algorithm and $H = 0.30$ and $H_i = 0.30$ as unidimensionality criteria for the scale and items respectively. Also, I used a rest score group of 100 when calculating the **crit** values and dropped an item if the *crit* value was > 80. After discarding an item, I repeated the procedure from the start, as the H_i of any item depends upon the other items included in the scale. For the full procedure I used the **mokken** package (van der Ark 2007, 2012) as implemented in **R**. For a full overview of this procedure, see Appendix I.

Appendix H shows the results of this procedure. At first glance, there are some interesting differences. To begin with, MSA did not maintain all the scales for each of the countries. In the case of the United Kingdom and Hungary, there was only a single EU scale. In most other countries there were either two or three scales, with one of them also being an EU scale. The latter is not strange, as the VAA focused on the European Parliament elections. Estonia, which had troublesome scales in its original and DSV form, has three scales, though they are both short and low in *H* values. The same goes for Lithuania, which lost a cultural scale, but whose economic scale is short (containing 3 items). The DSV already dropped *CU*4 in all countries but the United Kingdom, but now it disappears completely from any scale. Yet, items *EU*1 and *EU*6—relating to the Euro and EU membership, are in each EU-scale, again underlining their importance.

5 Results

Figures 1, 2, and 3 show the results for all countries for all three three versions of the scales, for the unidimensionality, reliability, and quality respectively. In each figure, squares represent the values for the economic scales, circles for the EU scales and diamonds for the cultural scales. Also, open symbols represent the values for the original scales, closed ones for the DSV scales, and crossed ones for the quasi-inductive scales if these scales were present.

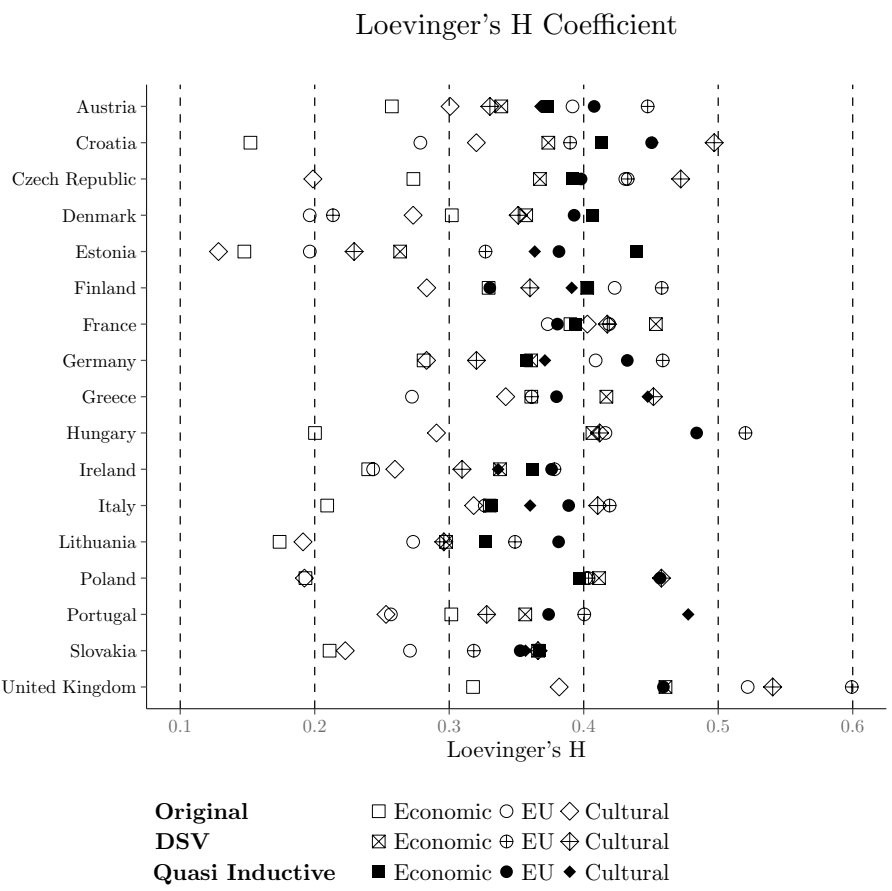

Fig. 1 Overview of the values for Loevinger's H coefficient for the original, DSV and quasi-inductive scales

5.1 Unidimensionality

With regards to unidimensionality, Fig. 1 reveals that the Loevinger's H values (see also Appendix J) lie between 0.1 and 0.6. This means some countries have scales with high unidimensionality while others lack unidimensionality. This is especially so for the DSV and quasi-inductive scales as these are restrained to only include items in the scale when $H_i > 0.30$. The exception to this is the DSV economic scale in Denmark and the DSV EU and cultural scale in Estonia. Here, the VAA included them despite their $H_i < 0.30$ as no other possibilities where available.

We find the highest values 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 (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. Estonia and Denmark have the lowest scores, as mentioned before. Despite changing the scales, even the DSV scales for the EU and cultural dimensions in Estonia were well

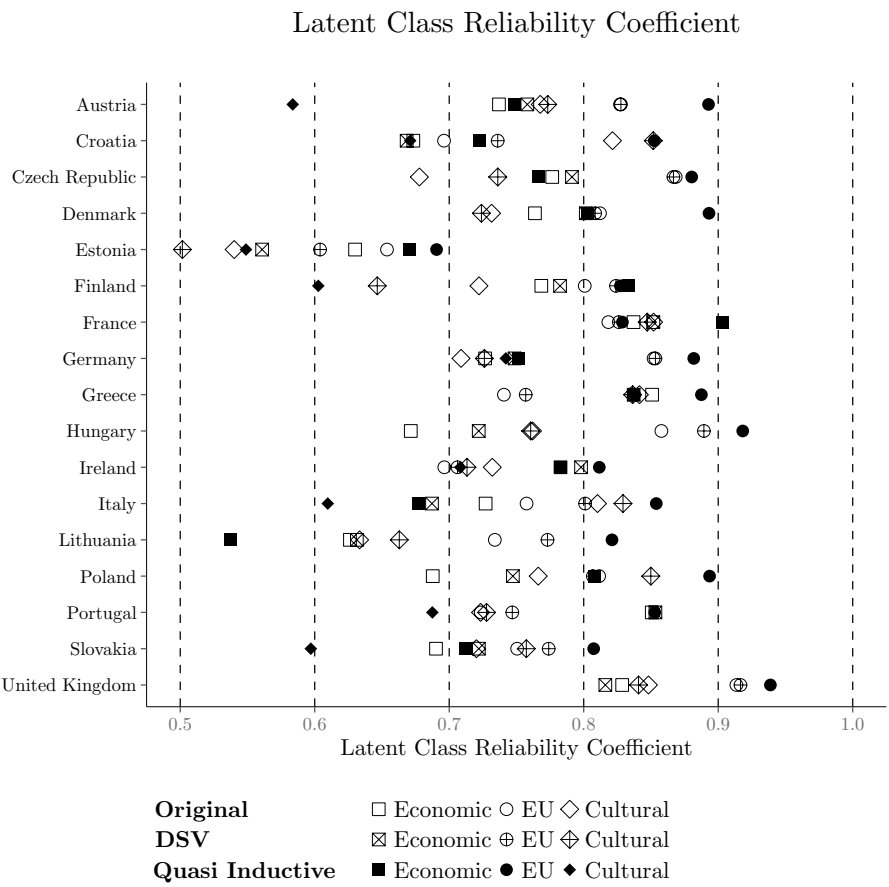

Fig. 2 Overview of the latent class reliability coefficient for the original, DSV and quasi-inductive scales

below the 0.30 mark. Also, the value for the economic scale is low, at 0.33. The quasi-inductive scales fare better but are small. Both the economic and cultural scales consist of 3 items, while the EU scale has 5 items.

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 with several items less. In the case of the EC scale, these comprise the items relating to cutting government spending (EC5) and loans from external institutions (EC7) as well as an item about the protection of the environment versus economic growth (AD5). Each of the items correlated low with the other items. The loans from external institutions item even show a negative correlation with the others ($H_i = -0.20$). This means that for that item we observe more errors than expected under statistical independence. This is most likely because the item is not in the correct direction. So, users who agreed with the other items of the scale disagreed with this item.

Other countries in which the value of H improved between the original and DSV version of the scale was Croatia. Here, 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,

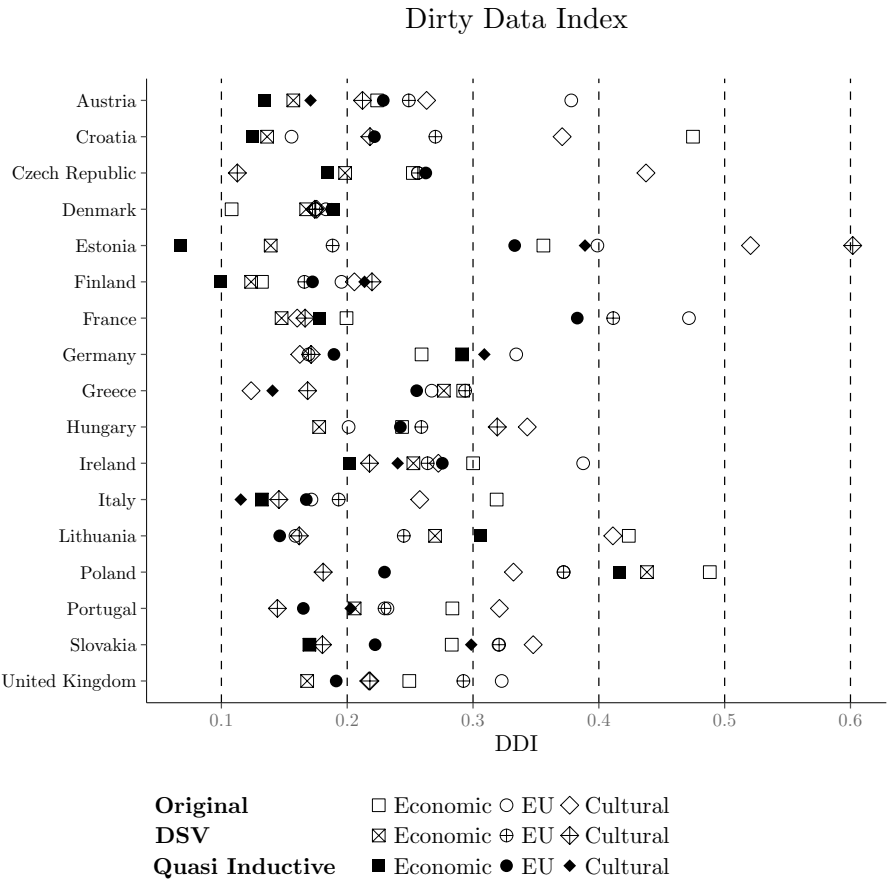

Fig. 3 Overview of the values for the DDI for the original, DSV and quasi-inductive scales

Portugal and Denmark, the improvement (in this case for the CU scale) was low and did little to help reach the 0.30 mark. Thus, the scales for EC and CU improved the most between the original and DSV scales, while the EU scale showed less improvement.

5.2 Reliability

Figure 2 shows the values for reliability in the form of the LCRC (see also Appendix K). As with other measures of reliability, the possible values lie between 0 and 1, with values closer to 1 indicating higher reliability. Most of the values of the LCRC lie between 0.50 and 0.90. Only two countries (Hungary and the United Kingdom) show values above 0.90, in both cases for the EU scale. This means that for most cases 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, I conclude it is not possible to

generate valid scales for Estonia. We can find another case of a low LCRC 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. It had 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 as well as the DSV scales in case of the EC scale. Yet, they perform worse in the case of the cultural scale. This is most likely because the cultural scales in their quasi-inductive form were small. Often, they contained only 3 items (as was the case in Austria, Estonia, Finland, France, Italy, Portugal, and Slovakia). Keeping the analogy with α, it is likely that the LCRC decreased because a higher number of items leads to higher reliability of the scale (Nunnally 1967).

5.3 Quality

For quality, Fig. 3 shows the results for the DDI (see also Appendix L). Opposite to the values of H and the LCRC, we want the DDI to be as low as possible. Here, we observe only two cases in which the DDI comes above the 0.5 mark, which Blasius and Thiessen (2012) argue indicates data of *bad* quality. Both cases are for the original and DSV version of the CU scale in Estonia. Interesting about this is that Estonia has both the highest and lowest DDI values. Its DSV cultural scale has a value of 0.52, while its quasi-inductive economic scale has a value of 0.07. The different values show that while the first scale is ordinal, the latter is near metric. In any case, the original values for Estonia are all above the 0.30 mark, as are two of the quasi-inductive scales. This indicates that even after running MSA in two different ways, these two scales remain problematic. This illustrates that MSA cannot guarantee a high quality of the scales.

Most of the other values are between 0.10 and 0.30, indicating metric data. For all scales and countries, the values belonging to the DSV and quasi-inductive scales are lower than the scores for the original scales, though the differences are often small. Besides, there is little difference between the quasi-inductive and the DSV scales. For the countries, Finland, Denmark and Italy perform the best, while Estonia and Poland perform the worst. This is because of the problematic EC scale in Poland, which performed badly in all its iterations and the high value for the CU scale in Estonia. Also, while the economic scale in Poland changed little in its DDI values (from 0.49 to 0.44 and 0.42), it improved in both its values of H (from 0.19 to 0.41 and 0.40) and LCRC (from 0.69 to 0.81 and 0.81). Thus, improving the unidimensionality and reliability does not ensure the quality of the scale. Another instance where this is the case is Ireland. Here, the *H* values for the economic scale increased from 0.24 to 0.34 between the original and quasi-inductive scales, while the DDI was 0.25 and 0.20

5.4 Relationships between the measures

While unidimensionality, reliability and quality measure different aspects of a scale, there is overlap between them. This is especially so for unidimensionality and reliability, as the first is a necessary condition for the second. Indeed, most reliability metrics assume unidimensionality at the risk of underestimating the reliability (Tavakol and Dennick 2011, p. 54). Thus, a low degree of unidimensionality should lead to a low degree of reliability, while a high unidimensionality should relate to a higher degree of reliability. Besides, both the estimates for the concepts base themselves on the degree of inter-item co-variances. They do so either in a direct way (Loevinger's *H*) or through an estimator (LCRC).

The opposite is true for the correlation between either the LCRC and the Loevinger's *H* values, and the DDI. Not only is there no clear link—as the focus of the DDI is on the categories of the item and not on the item itself—but the DDI is also formulated such that higher values indicate a negative performance of the scale, instead of a positive one.

As we already saw some evidence that the quality if not related to either unidimensionality or reliability, we will now look a bit closer at their precise relationship. Thus, Fig. 4 shows the correlations between each of the three possible pairs of criteria for each of the three versions of scales (over all countries and types), together with their 95% confidence intervals (see also Appendix M).

Starting with the original scales, I find high and significant correlations between the LCRC and *H* values as we would expect. Besides, the confidence intervals are small. For the correlations between the DDI and H values, I find conflicting results. While the CU 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. Yet, here the confidence intervals of the CU and EC scales are more extended. This means that for the economic and cultural scales an increase in the LCRC or *H* value goes together with 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, there is no relation between the two.

For the DSV scales, the relations between the LCRC and *H* values are all positive. Yet, 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. Only the CU scale indicates a significant negative correlation, like the one seen at the original scales.

For the quasi-inductive scales, none of the correlations is significant. The exception is the EU scales for the relation between H-LCRC. Here, the other two scales, while positive, are not 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 ($r = -0.01$). The other two scales are negative but also not different from 0. This is also the case for the relation between the LCRC and the DDI. Here, both the CU and EC scales are 0 ($r = -0.03$ and $r = -0.05$), and the EU scale is negative but different from 0. The reason for this, especially for the absence of a positive relationship between the H-value and the LCRC, is that the quasi-inductive algorithm produced only scales with a $H > 0.30$. As this reduced the number of possible values H could take, the spread of the H-value is limited, resulting in lower values of *r*.

So, while we can observe the expected relationship between reliability and unidimensionality, the relationship between these two and the measure of quality is not so straightforward. Thus, it is possible to construct a scale that shows high reliability and

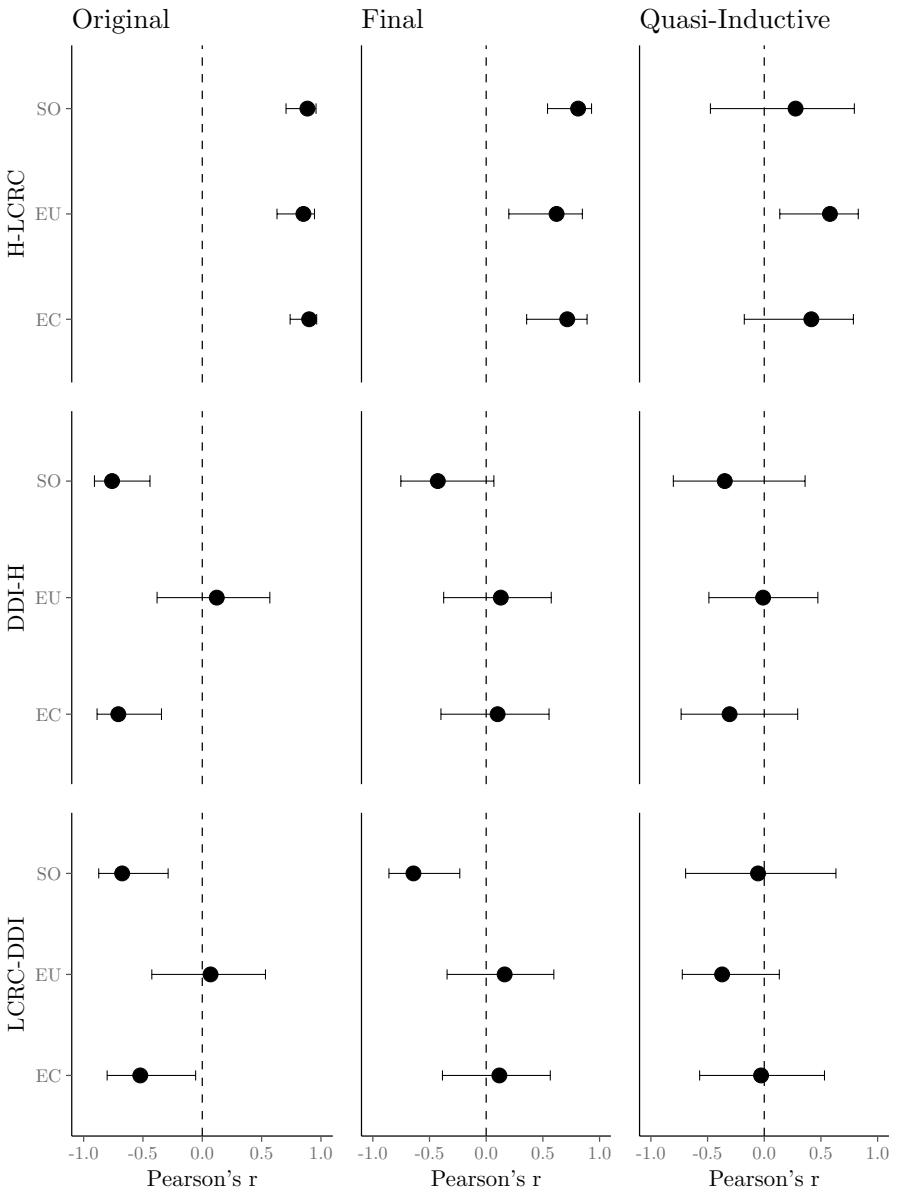

Fig. 4 Pearson's r correlations between the DDI, Loevinger's *H*, and the latent class reliability coefficient with 95% confidence intervals

unidimensionality, but from which the items themselves are not well understood by the users. Similarly, the degree to which the users understood the items says nothing about their combined behaviour on a scale. So, when confronted with a scale of items that show high unidimensionality and reliability, but low quality, it depends on the designer whether to decide to include or drop a certain item.

6 Conclusion

For the dimensions in VAA political maps to be useful and informative, VAA designers must validate the scales that underlie them. Here, I discussed three criteria they can use to do so: unidimensionality, reliability, and quality. As an example, I applied these to available data from the EUVox VAA. I did so for scales on three topics—economy, culture and European integration —for three different versions of the scales. I found that in only a few cases any of these scales meet the requirements demanded by each of the three criteria. Moreover, I reached similar conclusions when performing a robustness check using the data from the *euandi* VAA for the same elections (see Appendix N).

The scales scored especially low on reliability, with only five of the scales surpassing the lower bound of 0.90. For quality, the results were mixed, with most scales scoring a DDI between 0.10 and 0.30, indicating *relatively good* data (Blasius and Thiessen 2012, p. 136), while others performed worse and two even showed *poor* data. For unidimensionality, the results depended on the version of the scale. While most of the original versions of the scales scored below $H = 0.30$, the other two versions scored around 0.40. This result is further proof that VAA designers should not only base scales on theory but should also validate them using data. One method of validation—known as dynamic scale validation (Germann et al. 2015; Germann and Mendez 2016) - was tested here in one version of the scales, and in almost all cases, these scales performed better than their original versions. A third version of the scales, using a quasi-inductive method, did not perform better than either the original or the DSV scales.

From this, we can draw three main conclusions. First, that VAA designers should never base their scales on theoretical assumptions alone. Second, that even though using DSV can improve the scales, these improvements are only guaranteed for the unidimensionality criterion. This as the MSA algorithm takes only that criterion into account. Other criteria, such as the reliability and the quality, might improve, but not necessarily. Third, that while there is both a theoretical and empirical relationship between the unidimensionality and the reliability of the scale, such a relationship does not exist between either of these two measures and the quality of the scale.

While establishing if scales conform to standards is important in all aspects of political science, it is even more so in the case of VAAs. Not only because voters might be unaware of any deficiencies, but also because incorrect scales might lead to incorrect advice. This is especially problematic as there is increasing evidence that VAAs do not only influence voter turnout but also voter choice. Do so would go against the idea of most scholars in the VAA community that VAAs should benefit the voter. Thus, scale validation should become standard. Each of the three criteria shown here covers another element of validation and should be performed as soon as it is workable to do so. Moreover, carrying out this validation is relatively easy and requires only a few steps (see for an example Appendix O). Doing so would allow empirical reality to correct theoretical assumptions, and improve the quality of the voting advice.

Acknowledgements Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The author declares that there are no conflict of interest.

References

Andreadis, I.: Data quality and data cleaning. In: Garzia, D., Marschall, S. (eds.) Matching Voters with Parties and Candidates: Voting Advice Applications in a Comparative Perspective, pp. 79–92. ECPR Press, Colchester (2014)

Blasius, J., Thiessen, V.: Assessing the Quality of Survey Data. SAGE, London (2012)

Blasius, J., Thiessen, V.: Should we trust survey data? Assessing response simplification and data fabrication. Soc. Sci. Res. **52**, 479–493 (2015). https://doi.org/10.1016/j.ssresearch.2015.03.006

Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)

Carmines, E.G., Zeller, R.A.: Reliability and validity assessment. Sage, Newbury Park (1979)

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

Fossen, T., Anderson, J., Tiemeijer, W.: Wijzer stemmen? StemWijzer, Kieskompas en het voorgeprogrammeerde electoraat. In: van 't Hof, C.C.G., Timmer, J., van Est, R. (eds.) Voorgeprogrammeerd. Hoe Internet Ons Leven Leidt, pp. 163–188. Boom Lemma, The Hague (2012)

Gemenis, K.: Estimating parties' policy positions through voting advice applications: Some methodological considerations. Acta Polit. **48**(3), 268–295 (2013). https://doi.org/10.1057/ap.2012.36

Gerbing, D.W., Anderson, J.C.: An updated paradigm for scale development incorporating unidimensionality and its assessment. J. Mark. Res. **25**(2), 186–192 (1988). https://doi.org/10.2307/3172650

Germann, M., Mendez, F.: Dynamic scale validation reloaded—assessing the psychometric properties of latent measures of ideology in VAA spatial maps. Qual Quant **50**(3), 981–1007 (2016). https://doi.org/10.1007/s11135-015-0186-0

Germann, M., Mendez, F., Wheatley, J., Serdült, U.: Spatial maps in voting advice applications: the case for dynamic scale validation. Acta Polit **50**(2), 214–238 (2015). https://doi.org/10.1057/ap.2014.3

Hemker, B.T., Sijtsma, K., Molenaar, I.W.: Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. Appl. Psychol. Meas. **19**(4), 337–352 (1995). https://doi.org/10.1177/014662169501900404

Katsanidou, A., Otjes, S.: How the European debt crisis reshaped national political space: the case of Greece. Eur. Uni. Polit. **17**(2), 262–284 (2016). https://doi.org/10.1177/1465116515616196

Linting, M.: Nonparametric Inference in Nonlinear Principal Components Analysis: Exploration and Beyond. PhD thesis, Universiteit Leiden (2007)

Loevinger, J.: A systematic approach to the construction and evaluation of tests of ability. Psychol. Monogr. **61**(4), 1–49 (1947). https://doi.org/10.1037/h0093565

Loevinger, J.: The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychol. Bull. **45**(6), 507–529 (1948). https://doi.org/10.1037/h0055827

Louwerse, T., Otjes, S.: Design challenges in cross-national VAAs: the case of the EU Profiler. Int. J. Electr. Gov. **5**(3/4), 279–297 (2012). https://doi.org/10.1504/IJEG.2012.051305

Mendez, F., Manavopoulos, V.: EUvox2014: Voting Advice Application data for the 2014 European Parliament Elections (2018). https://doi.org/10.7802/1750. Licences: CC BY-NC 4.0

Mendez, F., Wheatley, J.: Using VAA-generated data for mapping partisan supporters in the ideological space. In: Garzia, D., Marschall, S. (eds.) Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective, pp. 161–173. ECPR Press, Colchester (2014)

Mendez, F., Gemenis, K., Djouvas, C.: Methodological challenges in the analysis of voting advice application generated data. In: Ninth International Workshop on Semantic and Social Media Adaptation and Personalization. IEEE Computer Society, Los Alamitos, CA, pp. 142–148 (2014). https://doi.org/10.1109/SMAP.2014.32

Mokken, R.J.: A theory and procedure of scale analysis—with applications in political research. Mouton, The Hague (1971)

Mokken, R.J., Lewis, C., Sijtsma, K.: Rejoinder to "the Mokken scale: a critical discussion". Appl. Psychol. Meas. **10**(3), 279–285 (1986). https://doi.org/10.1177/014662168601000306

Nunnally, J.C.: Psychometric Theory. McGraw Hill, New York (1967)
Otjes, S., Louwerse, T.: Spatial models in voting advice applications. Electoral. Stud. **36**, 263–271 (2014). https://doi.org/10.1016/j.electstud.2014.04.004
Polk, J., Rovny, J., Bakker, R., Edwards, E., Hooghe, L., Jolly, S., Koedam, J., Kostelka, F., Marks, G., Schumacher, G., Steenbergen, M., Vachudova, M., Zilovic, M.: Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data. Res. Polit. **4**(1), 1–9 (2017). https://doi.org/10.1177/2053168016686915
Sijtsma, K.: On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika **74**(1), 107–120 (2009). https://doi.org/10.1007/s11336-008-9101-0
Sijtsma, K., Molenaar, I.W.: Introduction to Nonparametric Item Response Modeling. Sage, Thousand Oaks, CA (2002)
Straat, J.H., van der Ark, L.A., Sijtsma, K.: Comparing optimization algorithms for item selection in Mokken scale analysis. J. Classif. **30**(1), 75–99 (2013). https://doi.org/10.1007/s00357-013-9122-y
Tavakol, M., Dennick, R.: Making sense of Cronbach's alpha. Int. J. Med. Educ. **2**, 53–55 (2011). https://doi.org/10.5116/ijme.4dfb.8dfd
van der Ark, L.A.: Mokken scale analysis in R. J. Stat. Softw. **20**(11), 1–19 (2007). https://doi.org/10.18637/jss.v020.i11
van der Ark, L.A.: New developments in Mokken scale analysis in R. J. Stat. Softw. **48**(5), 1–27 (2012). https://doi.org/10.18637/jss.v048.i05
van der Ark, L.A., van der Palm, D.W., Sijtsma, K.: A latent class approach to estimating test-score reliability. Appl. Psychol. Meas. **35**(5), 380–392 (2011). https://doi.org/10.1177/0146621610392911
Wheatley, J.: Restructuring the policy space in England: the end of the left-right paradigm? Br Polit **10**(3), 268–285 (2015a). https://doi.org/10.1057/bp.2015.35
Wheatley, J.: The use VAA-generated data to identify ideological dimensions: the case of ecuador. In: Second International Conference on eDemocracy & eGovernment (ICEDEG), pp. 55–60 (2015b). https://doi.org/10.1109/ICEDEG.2015.7114470
Wheatley, J.: Cleavage structures and dimensions of ideology in English politics: evidence from voting advice application data. Pol. Internet **8**(4), 457–477 (2016). https://doi.org/10.1002/poi3.129
Wheatley, J., Mendez, F.: (n.d.) Reconceptualising dimensions of political competition in Europe: a demand side approach. Br. J. Polit. Sci. 1–20 (forthcoming). https://doi.org/10.1017/S0007123418000571
Wheatley, J., Carman, C., Mendez, F., Mitchell, J.: The dimensionality of the Scottish political space: results from an experiment on the 2011 Holyrood elections. Party Polit. **20**(6), 864–878 (2014). https://doi.org/10.1177/1354068812458614