Quality & Quantity
https://doi.org/10.1007/s11135-024-01845-6

IRT for voting advice applications: a multi-dimensional test that is adaptive and interpretable

Karl Sigfrid[1]

Accepted: 16 January 2024

Abstract
Voting advice applications rely on user input to match user preferences to political parties or candidates. Providing the input can be time-consuming, which may have a negative effect on participation. For individuals who are under time constraints or who are affected by survey fatigue, the participation threshold may be lowered if there is an option to conclude the test without answering all question items. The test result should ideally be close to the result that the participant would have gotten after answering the full battery of questions. We propose a method that allows respondents to conclude a VAA early and still get results with sufficient accuracy.The method proposed here extends the Graded Response Model and the Maximum Information Criterion, used in Item Response Theory. The aim of the method is to allow the user to control the length of the test. Furthermore, we want a simpler interpretation of multidimensional parameter estimates than we get from traditional MIRT. To achieve this, we propose an algorithm for adaptive IRT capable of selecting from a pool of items that belong to separate unidimensional scales. Using both simulated data and response data from a voting advice application project, we evaluate the accuracy of shorter tests implemented with our adaptive method. When only a few test items are answered, our proposed method outperforms a static-order IRT test of equal length in identifying the best match. We expect that implementation of the method can increase participation and engagement in voting advice applications.

Keywords Item response theory · Computerized adaptive tests · Content-based recommender system · Voting advice application

1 Introduction

Voting advice applications (VAA) are popular in many European countries. In Belgium, Finland and the Netherlands more than one in three eligible voters claim to have used a VAA. In Norway, about one in two eligible voters make the same claim (Garzia and Marschall 2019). Ahead of the Swedish national election in 2018, 56 percent of the voters stated that they had used a VAA (Svenberg et al. 2022).

✉ Karl Sigfrid
karl.sigfrid@stat.su.se

[1] Department of Statistics, Stockholm University, Albanovägen 12, 114 19 Stockholm, Sweden

Published online: 05 March 2024

A VAA asks the respondent to take a position on a number of political issues deemed by experts to be relevant. When a test taker finishes the test, parties or candidates are ranked by how well they match the test taker's preferences.

A voting advice application typically requires the user to answer a substantial number of items. Ahead of the 2018 Swedish national election, the test offered by the Public Service TV company contained 33 items, the newspaper Svenska Dagbladet included 25 items and the newspaper Aftonbladet included 30. The VAA EUANDI had 22 items ahead of the 2019 elections to the European Parliament. The VAA smartvote had one version of the questionnaire with 59 items and a shorter version with 30 items for the 2023 Swiss national elections.

Long questionnaires can be time-consuming. Also, some participants find it difficult to complete the full battery of questions due to survey fatigue, a well explored phenomenon found to be associated with factors such as lower income, lower educational level and lower proficiency in the language of the survey (Le et al. 2021).

Survey fatigue in the context of VAAs can result in a test that is prematurely abandoned and does not provide a result. There is also a risk that the response quality dives toward the end of the test due to behaviors such as straightlining, resulting in an inaccurate conclusion (Herzog and Bachman 1981).

For the purpose of this paper we use three different terms that describe how a test session can come to an end. Respondents can finish the test, which means that all the items have been answered. Respondents can also abandon the test, which means that they walk away from an unfinished test and do not get a result. These are the only two options for a test of fixed length. A test of variable length can also be concluded before it is finished. In this case, the respondent chooses an option to end the sequence of items and gets a result. In this paper we propose that a VAA includes a button which ends the test and presents the result based on the items answered so far. The button may appear after a minimum number of items have been answered.

The method proposed in this paper aims to provide a sufficiently accurate result based on only a subset of the full question battery. It should be emphasized that adding an option to conclude the test early does not limit the number of items available to highly engaged respondents. To the contrary, an adaptive test that can be concluded at any time can have a larger total number of items without deterring users.

The method in this paper builds on the graded response model (GRM) proposed by Samejima (1968) and on adaptive item response theory (IRT) using the Maximum Information Criterion (MIC). The method aims to offer the respondent an interpretation of the result that is simple. Multivariate latent model parameters, such as those described in Reckase (2009), are known to have a more complex interpretation (de la Torre and Patz 2005). To avoid this, we let each item belong to only one scale. Whereas MIC is used to select the next best item from a scale, it does not tell us from which scale the next question is to be taken when we have multiple scales, as we do when we present the result in a low-dimensional space.

The main contribution of this paper is an algorithm that combines MIC with dimensional weights. By using a measure that both accounts for these dimensional weights and for an item's expected Fisher information, we can rank the usefulness of our remaining items across all scales. The proposed method can be applied to low-dimensional VAAs, i.e. VAAs that present the results in a low-dimensional political space, such as a 2-dimensional map.

The next section discusses the purpose of Voting Advice Applications, and common designs. We then describe survey drop-off and strategies that have been employed to lower

drop-off rates. In the subsequent executive summary, we give an overview of our proposed method.

We then describe the established models and algorithms that the method builds on, such as the graded response model and the Maximum Information Criterion. After these preliminaries we describe the proposed method in more technical detail. Finally we compare our proposed method to a fixed-order IRT model to demonstrate that our method has advantages when we want to provide an accurate result after only a few questions.

2 The purposes and effects of voting advice applications

Voting Advice Applications often state their purpose as educational. By responding to the questions and receiving matching scores the voter becomes better informed about relevant issues and party positions on these issues.

In some instances a goal may be to encourage participation in the electoral process. The German Wahl-O-Mat came into being as a response to a negative trend in voting turnout on sub-national level, especially among young voters. The tool aims to show that there are real differences between the political parties and that voting is therefore meaningful. The Wahl-O-Mat also attempts to make complex issues easier to understand (Marschall 2008).

The VAA EUANDI, previously named the EU Profiler, aims to make elections to the European Parliament more relevant in the eyes of citizens and to encourage citizens to be more politically active at the European level (Spaceu2019 2024).

The stated purpose of the Swedish Public Service VAA for the 2022 elections was to help voters familiarize themselves with relevant political issues, and to provide information on party positions. The VAA described itself as serious but yet entertaining (SVTNyheter 2022). The Swedish Public Service VAA does not state an ambition to influence user behavior, for instance by encouraging voter participation. It recommends that the matching scores should be viewed only as one of many factors when voting.

Whereas the stated goals of different VAAs vary to some extent, they all provide matches between voters and parties or candidates based on policy positions. Fossen and Anderson (2014) argue that traditional VAAs are built on the implicit assumptions of social choice theory, according to which the aim of elections is to aggregate given policy preferences of voters. Alternative theories, e.g. those that place importance on challenging your own opinions or those that question the policy agenda offered by the dominant political alternatives, may view the traditional VAA as less relevant. The decision to treat both the policy agenda and the voter preferences as given can be seen as normative decisions (Fossen and Anderson 2014).

In this paper we do not explore further the normative foundations of of VAAs, but will assume that knowledge about party positions is valuable to the voter. We will here focus on methodological choices that help us find a good measure of similarity. We also make the assumption that experts have formulated questions that are relevant to most voters.

In a meta-analysis of VAA effects, Munzert and Ramirez-Ruiz (2021) note an emerging consensus that VAAs can increase voter turnout and influence voter decisions. They also point out that self-selection issues may in some cases inflate estimated effects. A number of rigorous experimental studies have failed to show substantive VAA impact on voter turnout and voter choices at the polls (Munzert and Ramirez-Ruiz 2021). However, a more recent experimental study by Germann et al. (2023) with real VAA data from several countries showed that VAAs affected at least short-term voter preferences on an aggregated level.

The study gave some evidence, although inconsistent, that VAAs have a stronger influence on voters that are undecided and who have less political interest.

On the matter of whether VAAs influence voter decisions, selective exposure, i.e. the tendency to select information that agrees with one's pre-existing attitudes, may play a role. Steppat et al. (2022) found than selective exposure is a real phenomenon across Europe and in the United States. Especially individuals who previously had less opportunity to tailor their information sources to their views tended to do so when the opportunity arose.

Wall et al. (2014) found that respondents to the VAA kiescompass.nl in the Netherlands generally did not follow the advice when the suggested party was not already considered an option. On the other hand, when the VAA suggested a party that the respondent already considered a good option, this preference was strengthened. The VAA advice had more influence on uncertain voters.

In a review of early VAA literature, Albertsen (2022) found that early adopters were typically young, well-educated men with an interest in politics. While the same study found that the gender gap has disappeared and that the role of education has been reduced, age and political interest remained important predictors for who engaged with VAAs.

The method proposed in this paper aims to offer an option to design VAAs that are adopted by broader groups of voters. A test that allows the user to determine the number of questions may appeal also to voter who are less interested than the current users but still willing to complete a shorter test.

3 The design of voting advice applications

Garzia and Marschall (2016) list three methodological decisions that need to be made when designing a VAA. First, the choice and phrasing of items, which typically come in the form of policy statements. Secondly, the process to establish the positions of parties or candidates, and thirdly the measure of proximity between a test taker and a party or a candidate. To this we can add a fourth decision about the ordering of the items, which is central to this paper.

An assumption that we make here is that the VAA is designed by someone who wants to provide neutral advice. We also assume that the respondents provide answers that reflect their true preferences. If a respondent gives insincere input, perhaps to explore the VAA or to create a recommendation that is congruent with pre-existing attitudes, our method will match parties to the respondent's expressed policy positions rather than to the true policy positions of the respondent.

The choice of items has been shown to have a substantial impact on the VAA result. Walgrave et al. (2009) concluded from a simulation study, where each simulated sample consisted of 36 items chosen from an item pool of 50, that the proportion of respondents who got a certain party on top of their ranking could vary greatly from sample to sample. This stresses the need for a careful selection of items, especially when few items are included in the test.

The positions of political parties or political candidates can be determined through either input from party representatives, expert assessments or a combination of the two (Garzia and Marschall 2016). As an example, in prepararing the EU-Profiler, representatives from political parties were given the opportunity to fill out the survey. Independently from the party responses, groups of political experts coded party positions based on multiple sources such as party information material. Discrepancies between the party responses and

the expert coding were resolved through dialogue. However, the experts made the final decision (Trechsel 2011).

The decision on how to measure proximity between voters and parties is intimately connected to the spatial model. The spatial model determines how policy positions are represented in a space where each dimension measures a latent trait. For instance, a one-dimensional space implies that every item measures the same trait, e.g. a position on an ideological scale. A common spatial model is the two-dimensional space, where each dimension is a different ideological scale. For instance, the Portugese VAA Bússola Eleitoral used for the 2009 elections let one dimension represent the left/right scale, concerned with taxes, regulations and the roles of the public vs private sectors. Another dimension represented the GAL/TAN scale, concerned with lifestyle choices, migration and European integration (Lobo et al. 2010).

The two-dimensional representation is also used in Kieskompas in the Netherlands, with the implicit assumption that a closer distance between two positions in the space implies a better match. A space of more than two dimensions is more difficult to illustrate, but has nevertheless been used in EUANDI and the Swiss VAA smartvote. In the case of smartvote, a radar chart, also known as spider chart, illustrates the proximity of policy positions in eight dimensions (Louwerse and Rosema 2014). These charts may be perceived as less easy to read than a two-dimensional map, and one chart typically compares the user's policy positions to the positions of only one party or candidate.

Many VAAs do not use a low-dimensional space, in which case we refer to the VAA as high-dimensional since each item can be viewed as measuring its own dimension. A VAA that simply counts the number of items where the user agrees with a party is thus a special case of a high-dimensional VAA as defined here. This is how StemWijzer in the Netherlands has calculated matching scores since 2006 (Louwerse and Rosema 2014).

VAAs sometimes combine low-dimensional representations of the results with high-dimensional matching scores, as was the case with the Portugese Bússola Eleitoral and EUANDI 2019 (Lobo et al. 2010; Michel et al. 2019).

Given a spatial model, a VAA also needs a distance measure if it is to present a ranking of the parties or candidates. This can be the Euclidean distance, which is the length of a straight line between two points in the model space. It could also be the Manhattan distance, sometimes called the city block distance. To find the Manhattan distance between two points, one measures the distance between the points in each dimension separately, and then sums these distances.

It is natural to measure the Manhattan distance for high-dimensional representations. For instance, calculating the percentage of items where a respondent agrees with a political party is a measure that uses Manhattan distances. In the low-dimensional case it is more natural to measure the Euclidean distance. With a low-dimensional representation of the result, such as a two-dimensional ideological map, it can also be left to the user to visually assess the relative distances to the different political parties or candidates.

4 Survey response rates and survey drop-off

The purpose of our proposed method is to make VAAs appeal to a broader group of users. The problem that we address is similar to the problem of increasing survey response rates and minimizing drop-off rates. We will here look at how these problems have been approached.

The survey response rate can be calculated in different ways. Here we define it as the number of people who responded to a survey divided by the number of people invited to participate. We define the dropout rate as the number of respondents who started the questionnaire but did not finish it divided by the number of people who started the questionnaire.

Vicente and Reis (2010) mention two main strategies to increase survey participation: The use of incentives and designing a questionnaire that makes the respondent more inclined to cooperate. We will here focus on the latter.

One factor explored by Manfreda et al. (2002) was the choice between one-page scrolling web questionnaires and multi-page web questionnaires where each page presented one item or one group of similar items. No significant difference in dropout rate between one-page questionnaires and multi-page questionnaires were found. We can note that the method proposed in this paper depends on the multi-page design as it repeatedly selects the next item based on previous responses.

A second factor is the presence of a progress indicator. Progress indicators have shown to be either statistically insignificant or detrimental for the purpose of lowering the dropout rates depending on how the speed of the progress is presented (Villar et al. 2013).

A third factor is questionnaire length. Cook et al. (2000) found no significant correlation between the number of survey items and the response rate. However, more recent studies have found questionnaire length to be an important factor. Shorter versions of surveys have been associated with both higher response rates (Deutskens et al. 2004) and lower dropout rates (Ganassali 2008). In addition, Deutskens et al. (2004) found that the rate of incomplete responses and the use of the "don't know" option was higher among respondents who answered the long version of a survey.

In addition to the actual survey length, potential respondents were affected by information given prior to taking the survey. Crawford et al. (2001) found that people who were informed that a web survey would take 8 to 10 min more often began the survey than people who were informed that the same survey would take about 20 min. The true amount time to finish the survey was estimated to somewhere between 10 and 20 min, although in fact the median time consumption turned out to be almost 20 min. The group who was told that the survey would be shorter were more willing to respond, but they had a lower completion rate. Respondents who expected a longer survey were more inclined to finish it, given that they had agreed to start the survey.

Based on these findings, VAA participation could possibly be increased if potential participants are promised an option to answer fewer items. This option should then be communicated prior to the decision on whether to participate.

Since we want the VAA to appeal not only to those who prefer a quick test but also to the enthusiasts, the prior information should make it clear that the user has the options to either conclude the test after a few items or to answer a large battery of items. After a minimum number of items have been answered, the respondent should continuously have the choice to either conclude the test and see the result or to answer the next item. In this respect our approach is different from the approach of simply limiting the number of items in a fixed-length test.

As the respondent does not have to decide beforehand on the number of answered items, it is possible that some participants who are drawn in by the promise of a short test still decide to answer a larger number of items. It is also possible that a respondent who initially intended to answer all items concludes the test early.

When a respondent who otherwise would have done the full test instead concludes the test early, it means that potential answers are lost and cannot be included in follow-up

analyses. This may be a negative, but it could to some extent also be a positive consequence. When a respondent who is tired of the test answers the remaining items, there is a risk of low-quality responses. Missing values may then be preferable.

Our method is designed to increase the accuracy of comparisons between respondents and political parties in a low-dimensional space. The dimensions typically represent ideological scales or broad topics. The accuracy of the results will be lower with only a subset of the items answered than with all items answered. However, with the dynamic item ordering proposed here, the accuracy from a subset of answers will be better than it would have been with a static item ordering. Considerations related to high-dimensional representations of the results are outside the scope of the method proposed in this paper, which is exclusively for low-dimensional matching.

5 Executive summary of the proposed method

Our method assumes that we have a pool of items and that each item belongs to one and only one unidimensional scale. One such scale measures a latent trait, which may be the position on an ideological scale. The number of dimensions in the political policy space can be anywhere from one to a large arbitrary number. We use $\theta_{j,k}$ to denote the true position of respondent j on scale k. We let $\hat{\theta}_{j,k}$ denote our estimate of this position.

Further, we assume that the scales have been validated, i.e. that all items on one scale actually measure the same latent trait. To do this, we suggest Mokken Scale Analysis, which has previously been proposed for VAA applications (Germann and Mendez 2016).

When a respondent goes through the VAA session, items are continuously selected from an item pool. After each answered item, our proposed method ranks the remaining items based on usefulness, and the highest-ranked item is selected. The usefulness of an item is here a measure that takes into account two different factors: First, how much do we expect that a response to this item will increase the accuracy of $\hat{\theta}_{j,k}$? Secondly, how important is it to increase the accuracy of the estimate in dimension k? As we define accuracy here, an item increases the accuracy more if it leads to a greater decrease in the standard deviation of $\hat{\theta}_{j,k}$.

The usefulness of an item i is the product of its expected contribution to the accuracy of $\hat{\theta}_{j,k}$ and the importance of dimension k. We use the notation $\xi_{i,k}$ for the usefulness of item i that measures dimension k. We use $\eta_{i,k}$ to denote the expected increase in accuracy from choosing item i, and δ_k denotes the importance of improving the accuracy in dimension k. The item usefulness is then calculated

$$\xi_{i,k} = \eta_{i,k} \cdot \delta_k. \tag{1}$$

Whereas it is clear why we want to measure $\hat{\theta}_{i,k}$ with high accuracy, the idea of dimension importance requires an explanation.

Every time that the respondent has answered another item, we calculate a new provisional estimate $\hat{\theta}_j$, which positions the repsondent in our low-dimensional space. In this same political space we have the positions of political parties or political candidates, and we are mostly interested in the ranking of political parties in the near neighborhood of the respondent. Given the positions of the respondent and the political parties, some dimensions are more important than others in determining the ranking among the top ranked parties. In the two-dimensional case, it may be that moving $\hat{\theta}_j$ in dimension 1 has little effect

on the ranking of the nearest parties, while moving $\hat{\theta}_j$ in dimension 2 has greater effect. In this scenario we assign dimension 2 a higher importance score than dimension 1. Since the ranking in this scenario is more sensitive to movements in dimension 2, we are more interested in improving the accuracy of $\hat{\theta}_j$ in dimension 2.

When a new respondent begins a test, we will have to make an initial guess about θ_j. One reasonable initial guess may be that the respondent is located in the center in every dimension. If we make the same guess for every new respondent, which is reasonable if we have no prior knowledge about the respondents, the first item will always be the same. After that, the sequence of items may be different for different respondents, as the sequence depends on their answers. If two respondents answer the items in a way that is similar but not identical, they may still get the same sequence of items.

When the respondent ends the test, either by clicking a button that concludes the test early or by finishing all items, a result is presented in the low-dimensional space. A ranking of the parties can also be presented. The algorithm for our proposed method can be summarized as follows.

1. Make a guess about the value of $\theta_{j,k}$ for each k. This guess serves as our initial estimate $\hat{\theta}_{j,k}$.
2. Given $\hat{\theta}_{j,k}$, calculate the usefulness of each remaining item in the item pool[1]:
   (a) Calculate $\eta_{i,k}$ as a measure of how much each item is expected to reduce the standard deviation of $\hat{\theta}_{j,k}$.
   (b) Calculate δ_k, which measures the importance of dimension k.
   (c) $\xi_{k,i} = \eta_{i,k} \cdot \delta_k$ is the usefulness for item i, which belongs to dimension k.
3. Select the item with the highest usefulness.
4. The respondent either answers the selected item or chooses an option to conclude the test.[2]
5. This step depends on the respondent's choice in step (4).
   - If the respondent answered the item, update $\hat{\theta}_{j,k}$ based on all the respondent's answers so far. Then return to step (2).
   - If the respondent chose to conclude the test, present the results.

6 Method preliminaries

Our proposed method leans heavily on item response theory. IRT is typically used to measure latent abilities. The term "ability" here refers loosely to an ability, a personal trait or an attitude. For instance, having an ideological position near the center of the scale can be viewed as analogous to having an average ability. An ideological position far from the center can be analogous to an ability far from the average.

We can reason in a similar manner regarding the IRT difficulty parameter. A response option that requires a strong preference or a large dose of an attitude may be viewed as

[1] If the pool of remaining items is empty, end the test and present the results.

[2] A minimum number of items may have to be answered before the option to conclude the test appears.

"difficult" to choose. A popular proposal can be considered "easier" to agree with than one that is unpopular.

6.1 The graded response model

The graded response model is an extension of the 2PL IRT model. It accommodates graded scores and is therefore a natural choice when graded scales such as the Likert scale are used for the response options. Each response option is associated with a level of achievement where 1 is the lowest level and M the highest. The probability that respondent j ends up at level m on test item i is

$$P(u_{i,j} = m|\theta_j) = P(u_{i,j} \geq m|\theta_j) - P(u_{i,j} \geq m+1|\theta_j), \tag{2}$$

where

$$P(u_{i,j} \geq m|\theta_j) = \frac{e^{a_i(\theta_j - b_{i,m})}}{1 + e^{a_i(\theta_j - b_{i,m})}},\ i \in \{1, 2, ..., n_I\},\ j \in \{1, 2, ..., n_J\}. \tag{3}$$

Equation (3) is used to calculate the probability that respondent j succeeds in reaching level m or higher. θ_j is the ability of respondent j, $b_{i,m}$ is the difficulty of reaching level m of item i, and a_i is the discrimination parameter of item i. The discrimination parameter is a measure of how well the item discriminates between respondents with different abilities in the region of $b_{i,m}$. Note that the ability parameter θ_j and the difficulty parameter $b_{i,m}$ are measured on the same scale. With only two response options, Eq. (3) becomes equivalent to the dichotomous 2PL model. A more detailed description of GRM can be found in Samejima (1997) or Baker (2004).

6.2 Adaptive tests with IRT and the maximum information criterion

Whereas there are different types of tests that can be called adaptive, we are here interested in adaptive tests in the context of questionnaires. Rather than a fixed sequence of items, adaptive tests have a pool of items, and previous answers determine the next item to be presented to the user. The computerized form of these tests has been around since the 1970s (Reckase 1974). They typically build on IRT and take advantage of a provisional estimate of the ability parameter.

The MIC approach states that we should select the item that provides the highest Fisher information conditioned on the current ability estimate of the respondent. For GRM, the Fisher information obtained from item i conditioned on the ability θ_j is calculated

$$I_i(\theta_j) = \sum_{m=1}^{M} \frac{a_i P_{i,m}(\theta_j)(1 - P_{i,m}(\theta_j)) - a_i P_{i,m+1}(\theta_j)(1 - P_{i,m+1}(\theta_j))}{P_{i,m}(\theta_j) - P_{i,m+1}(\theta_j)}, \tag{4}$$

where M is the number of achievement levels for item i. It will always be the case that $P_{i,1} = 1$, since it is trivially easy to reach the lowest level. It is also always the case that $P_{i,M+1} = 0$, since M is the highest level that the respondent can reach. When we calculate the Fisher information in a real scenario where we don't know the true ability of the respondent, we will use the estimate $\hat{\theta}_j$ in place of the true value.

For a set of items, the Fisher information is the sum of the information obtained from the individual items in the set. If U is a set of indices for the items in a set, then

$$I(\theta_j) = \sum_{i \in U} I_i(\theta_j). \tag{5}$$

The adaptive test may, but does not need to, have a stopping rule. Such rule can be that the test ends when the estimated variance has shrunk below a chosen threshold.

6.3 Scale validation

Previous work has shown that in low-dimensional representations of abilities, the scales are sometimes constructed without empirical justification. For instance, Germann et al. (2015) argue that the current practice is to let VAA scales be *a priori* constructions. Without empirical validation, the scales may not be unidimensional, i.e. the different items for the scale may not measure the same ability.

We suggest Mokken scale analysis (MSA) to validate our most important model assumptions (Mokken 1971). MSA has wide applications in the social sciencies as a tool to support the design of questionnaires (Watson et al. 2012). MSA uses non-parametric IRT models to validate latent-variable scales. As a non-parametric tool, it cannot address all the assumptions of parametric IRT models, such as the logistic shape of the item characteristic curve. However, it addresses the important assumptions of unidimensionality and monotonicity (Sijtsma and van der Ark 2017).

Rather than fitting item parameters and ability parameters, non-parametric IRT models rank test takers on an ordinal scale. The lack of a predetermined shape has advantages for scale validation. Stout argues that the lack of fit of a particular univariate model to the data does not prove that *no* unidimensional IRT model fits the data (Stout 2002). Because of their flexibility, non-parametric IRT models are less prone than parametric IRT models to mistake an inappropriate choice of univariate model family for data multidimensionality (Stout 2002).

As far as using non-parametric IRT models for adaptive testing, that comes with its own challenges. One challenge is that the often used Maximum Information Criterion relies on the derivative of the item response function with regards to the ability parameter. This derivative is less reliable under the non-parametric assumptions (Xu and Douglas 2006). While we rely on MSA for scale validation in this paper, we use parametric IRT for the adaptive test.

A central and widely used tool in the MSA toolbox is the scalability coefficient, also known as the Loevinger's coefficient (Loevinger 1948; Watson et al. 2012; Sijtsma and van der Ark 2017). The coefficient H measures how well a group of items form a unidimensional scale. Details on the calculation of the scalability coefficient can be found in Sijtsma and WMolenaar, I. (2002).

For a set of items to constitute a Mokken scale, H should be greater than some threshold value c. A rule of thumb is to use $c \geq 0.3$ as a lower threshold. It has been proposed that a scale should be considered weak if $0.3 \leq H \leq 0.4$, moderate if $0.4 \leq H \leq 0.5$ and strong if $H > 0.5$ (Ark and v. d. 2007).

In addition to assessing unidimensionality through the scalability coefficient, Sijtsma and van der Ark (2017) argue that a Mokken scale analysis should include tests for monotonicity. Monotonicity implies that the probability of a correct answer increases, or at least does not decrease, with the respondent's ability.

To check for monotonicity, we calculate a restscore, which we denote $R_{i,j}$, for each respondent and item. The restscore is the sum of the scores that the respondent achieves

on all items except item i. The restscore is thus a measure of the respondent's ability. To check monotonicity for one item, we can divide the respondents into groups based on the restscore, a procedure referred to as binning (Molenaar and K. S. 2016). If the average score on item i is always higher for a groups of respondents with higher restscores, then we consider item i to pass the monotonicity test. A practical way to do the test is to plot a regressogram with the binned rest scores on the x-axis and the average item scores on the y-axis.

When we construct the test, it is essential to ensure that the scales meet both technical criteria and common-sense criteria for unidimensionality. Formal tests can validate that the scales meet the chosen technical criteria while the common-sense criteria requires that the item response options make intuitive sense as manifestations of the latent trait that the scale measures.

A challenge that comes with the technical validation is that this validation requires empirical data, which is often not available before the implementation of the VAA. To address this difficulty, Germann and Mendez (2016) have suggested what they call Dynamic Scale Validation. When this procedure is used, early respondents are profiled using scales that are not empirically validated. Based on the response data from these early users, the scales can be evaluated and refined for the benefit of subsequent users (Tukey 1961). In the section that follows, we assume that the scales are unidimensional and that the item characteristic curves are non-decreasing.

6.4 Parameter estimation

In an initial stage, we have a set of response patterns but no model parameters. We can start by assuming an ability distribution among the respondents. Although it can be any distribution, a common choice is the standard normal distribution. Based on this assumption, the EM algorithm can be used to estimate the item parameters. Implementations of this algorithm are freely available, for instance through the R mirt package (Chalmers 2012).

6.4.1 The maximum likelihood and maximum a posteriori estimators

Conditioned on the estimates of the item parameters, we can estimate the ability parameters of test takers. One option is to use Maximum Likelihood Estimation (MLE), which may unfortunately produce estimates that are positive or negative infinity. This occurs in the dichotomous case either when all responses are correct or when all responses are incorrect. Thus it occurs with certainty when only one item has been answered, and frequently when only a few items have been answered. For the graded response model, the estimate is infinity when all responses correspond to either the highest level of achievement or to the lowest level.

The infinity problem is very present in an adaptive test situation where the choice of the next item depends on a provisional estimate from previous items. This implies that we recalculate the ability estimate after each response, rather than waiting until all test items have been answered.

To handle the infinity problem, and to get better estimates from small numbers of responses, MAP estimation can be used in the place of MLE. The MAP estimate is the value of the ability parameter that maximizes a posterior distribution function. The prior is often the standard normal distribution.

6.4.2 The choice of prior for the maximum a posteriori estimator

For the adaptive test described in this paper, we recommend the MAP estimate of the ability parameter. Instead of using a standard normal distribution as prior, we can improve accuracy by estimating one dimension with a prior based on information from other dimensions.

To find better priors with this technique we need a reasonably large group of respondents who have already taken the test. The models will be based on data from these early participants. A more technical description of how to calculate model based priors for the MAP estimate is given in the supplementary material.

7 Method

After each response, we want to find the most useful item to present next. Our proposed method ranks the items from an item pool in order of usefulness. For each remaining item $i \in \{1, 2, ..., n_I\}$ that belongs to a scale $k \in \{1, 2, ..., K\}$, we can calculate the item's usefulness as

$$\xi_{i,k} = \eta_{i,k} \cdot \delta_k \tag{6}$$

The parameter $\eta_{i,k}$ is the estimated reduction of the standard deviation in dimension k resulting from selecting item i in scale k. The parameter δ_k measures the importance of the scale to which the item belongs.

To calculate $\eta_{i,k}$, we note that in the univariate case, the maximum likelihood estimate $\hat{\theta}_{(ML)}$ has a normal distribution asymptotically with the true parameter value as its mean and a variance that equals the inverse of the sum of the Fisher information of the answered items (Reckase 2009).

By selecting the item with the highest Fisher information given the provisional location $\hat{\theta}_{j,k}$, we minimize the variance conditioned on the estimated ability. We are not guaranteed to minimize the actual variance, since this depends on the unknown true value of the latent parameter $\theta_{j,k}$. We assume that the item ranking based the accuracy of the ML estimate will be a good approximation for the ranking based on the accuracy of the MAP estimate.

A uniscale setting only requires that we continuously choose the item with the highest Fisher information conditioned on our current estimate of θ_j. In a setting with multiple scales we are also required to choose which scale that we want our next item to measure. We here suggest that an item is better if it gives a greater reduction in the standard deviation for the scale that it measures. This selection criterion favors items in scales where the current variance is high.

The scale importance δ_k measures how important scale k is compared to the other scales. A more important scale is here defined as a scale that has a greater influence on the ranking of the parties in the neighborhood of the current ability estimate.

A detailed description of how item usefulness is calculated mathematically is given in the supplementary materials together with a numerical example. Here we give an intuitive explanation of the meaning of scale importance.

Figure 1 illustrates a scenario where we measure a respondent's location in a two-dimensional space. Each dimension represents a latent trait that we want to measure. A

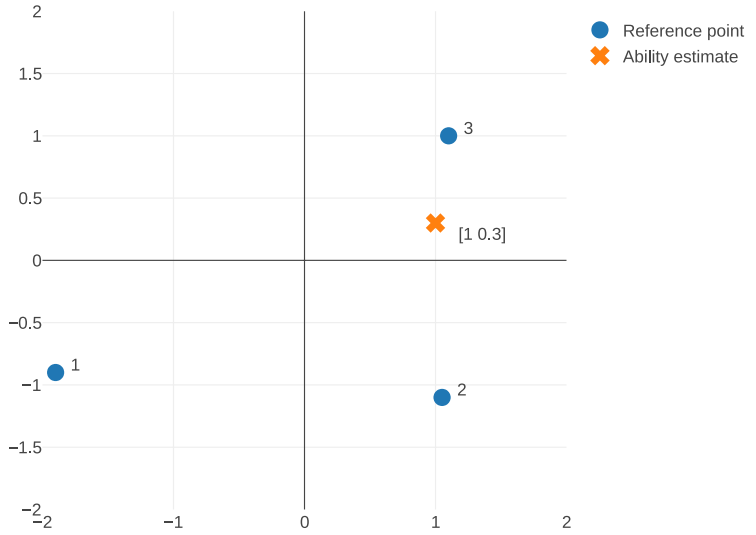

Fig. 1 In this scenario we have a provisional estimate of our ability on two scales. We also have three party positions, and we are interested in how these party positions rank based on the proximity to the respondent

provisional ability estimate shows where we believe that our respondent is located in the political space. In the same space we show the location of three different political parties.

We can see that if the respondent's ability estimate moves in the horizontal dimension it has little effect on the ranking of the parties. Horizontal movements do not change the ranking of party 2 vs party 3. Horizontal movements do change how well the respondent matches against party 1 compared to the other parties, but since party 1 is far away this is less relevant.

If the respondent moves in the vertical direction, this can have an effect on whether party 2 or party 3 will be top ranked. Since both these parties are in the neighborhood of the provisional ability estimate, such movements are considered important. For this reason, we consider the accuracy in the vertical dimension more important than the accuracy in the horizontal dimension in this scenario.

8 Example application

In this section we compare the accuracy of the proposed method to the accuracy of a fixed-order IRT test. We measure accuracy as the proportion of respondents who are matched to their best party. For all results we used R version 4.1.1. The graded response models were fitted with the mirt package version 1.36.1.

The definition of best party differs depending on whether we use simulated data or empirical data. With simulated data, the best party has the shortest Euclidean distance to the true ability used to simulate the data. With the empirical data, the true ability is unknown. In that case we consider the best party to be that which closest matches the ability estimate when all items are included.

For the empirical data, the method used to determine the best party may have an unfair advantage over alternative methods when tests of shorter lengths are compared. To avoid this problem, a shorter test is evaluated against the results of a full-length test done with the same method. Thus, the accuracy of a short-version test with method 1 measures the proportion of respondents who get the same matching as with the full-length version of the test with method 1, and likewise for method 2.

8.1 Comparisons to static-order IRT using simulated data

8.1.1 The simulated data

We generated abilities in two dimensions for 2 000 respondents. The abilities were generated from a bivariate normal distribution with mean 0 and unit variance in both dimensions. The correlation between $\theta_{j,1}$ and $\theta_{j,2}$ was set to 0.5. This resulted in a correlation between the MAP estimates $\hat{\theta}_{j,1}$ and $\hat{\theta}_{j,2}$ of approximately 0.454 when using a standard normal prior. This is comparable to the correlation found in the VAA response data described in Sect. 8.2.1, and we therefore view it as reasonable in a VAA context. The first 1 500 simulated respondents were used as training data and the last 500 respondents were used as test data.

We also simulated 60 items, each scored on a scale with 5 grades. Using GRM, the discrimination parameter was drawn from a uniform distribution with 0.5 as the minimum and 3 as the maximum. The four difficulty parameters were drawn from a uniform distribution with -3 as the minimum and 3 as the maximum. The difficulty parameters for each item were sorted in ascending order, so that $b_{i,1} < b_{i,2} < b_{i,3} < b_{i,4}$. To ensure a minimum of spacing between the difficulty parameters for the same item i, 0.1 was added to $b_{i,2}$, 0.2 to $b_{i,3}$, and 0.3 to $b_{i,4}$.

In our original setup we placed 15 parties in the ability space. The positions of the parties in each dimension were drawn independently from a uniform distribution with -2 as the minimum and 2 as the maximum. We also defined setups with fewer parties. To get a setup with 14 party positions, we started with the 15 random party positions and merged the two closest positions into one. The location of the new point was the average of the two merged points. We repeated this procedure until we had only 2 parties left. For our comparisons, we used setups with 2, 5, 10 and 15 parties respectively.

8.1.2 Comparing the models

We fit the item parameters using the training data. In the next step, we estimated the abilities for the test data, and calculated the proportion of respondents who were matched to the best party with static-order IRT after 1, 2,..., 60 items. For each user, the item ordering was individually randomized.

As a comparison we calculated the proportion who were matched to the best party with our proposed method after 1, 2,..., 60 items. When we estimated the best party after n items, this corresponded to simulating respondents who concluded a 60-item test after n items by clicking a button to end the test. 60 items is longer than most VAAs, but this does not affect the results. With our setup, there is in principle no difference between a simulated respondent who answers 5 out of 60 items and a respondent who answers 5 out of 30 items. The range from 2 to 15 simulated parties covers the range

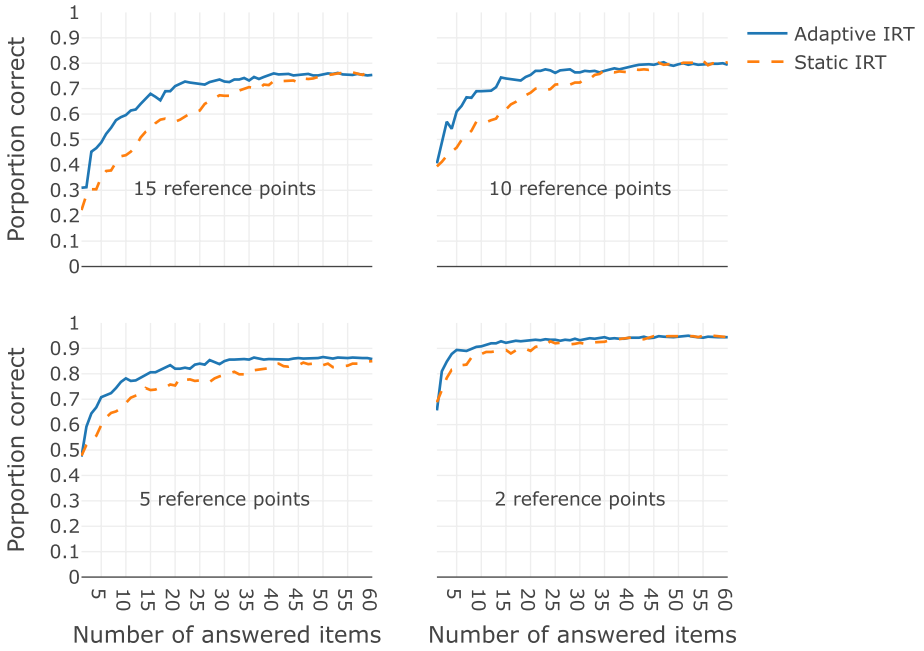

Fig. 2 The proportion of respondents (y-axis) who get the correct top-ranked party when n items (x-axis) have been answered. We compare the proposed adaptive method (solid) to static-order IRT (dashed)

that we would expect in a real scenario. For instance, the VAA EUANDI 2019 has a set of parties from each country. The median number of parties for a country is 9. The maximum is 15 and the minimum is 3.

We repeated the comparison with 2, 5, 10 and 15 parties. For both our proposed method and for static-order IRT, we estimated the ability to 0 in both dimensions when no item had been answered.

8.1.3 Results

The results in Fig. 2 show that the method proposed in this paper gives better estimates than a fixed-order IRT model for tests of shorter length. When the number of answered items is in the range 2–10, our proposed method has a clear advantage regardless of the number of parties. However, the advantage is greater when we match respondents against many parties or candidates.

With a larger number of parties, our proposed method has an advantage even when the user answers as many as 40 items. This in contrast with the scenario where we have only two parties in which case much of the advantage has dissipated after 20 questions.

Thus, when the VAA matches the respondent against three or more candidates, our proposed method can be expected to give substantially better accuracy than a fixed-order test. This makes it a good option for a VAA in a system with multiple parties. A results table that corresponds to Fig. 2 is found in the supplementary materials.

It can be noted that the results from the standard static model and our proposed method are not identical even when all 60 items have been answered. This is because

the item ordering is not the only difference. For the standard IRT method we used a standard normal MAP prior while in our proposed method we used a model-based MAP prior as described in Sect. 6.4.2. Results that isolate the effects of the model-based MAP prior are included in the supplementary material.

8.2 Comparisons to static-order IRT using empirical VAA response data

8.2.1 The VAA data

Here we compare our proposed model to a fixed-order IRT model using real response data collected by the Laboratory of Opinion Research (LORE) at University of Gothenburg between 12 Dec 2017 and 8 Jan 2018. The participants were members of the Swedish citizen panel, a Swedish web panel used for opinion research (Oleskog 2018).

We view the data as equivalent to real VAA data, as the data was generated by members of the general public who evaluated policy statements. The data was collected in a pre-stage for the Swedish Public Service VAA in 2018, but the respondents can be assumed to have interacted with the statements as they would in a VAA setting. In the actual VAA, the items were presented in a certain order, and this is the order that we consider the default order. When we compare our method to a standard IRT model, the standard model uses this default item ordering for the items that were included in the final VAA. In some cases items were slightly rephrased in the actual VAA compared to how they were phrased in the pre-stage survey. The use of the dataset for the purpose of this research project was approved by the Swedish Ethical Review Authority.

The questions were phrased as political proposals, and the respondents could state that a proposal was "Very good", "Quite good", "Quite bad" or "Very bad". They also had the choice of expressing "No opinion". Examples of proposals were "The Swedish Public Employment Service should be abolished", "The Police should have a wider authority to use camera surveillance", and "Profits should not be allowed in the tax-financed welfare sector". LORE invited 3 967 respondents from the Swedish citizen panel to the survey, and 2 680 opted to participate.

The dataset consists of 50 items that were considered relevant in the Swedish national 2018 elections. For 24 of the items we had party positions available. After removing observations that were incomplete with regards to these 24 variables we had 2 582 observations. We randomized the order of the observations and assigned 1 807 to a training set. The remaining 775 observations were assigned to a test set.

A Mokken scale analysis, where we used $c = 0.4$ as the threshold value for the scalability coefficient, resulted in 2 scales with 8 and 7 variables respectively. One scale could be interpreted as measuring traditionalist vs socially liberal sentiments and one scale as measuring pro-market vs pro-government sentiments. We removed three items that did not pass the requirement of intuitively fitting the scale interpretation. This left us with 6 items in each of the two scales. The correlation between the training data MAP estimates $\hat{\theta}_{\cdot,1}$ and $\hat{\theta}_{\cdot,2}$ was 0.451 when using a standard normal prior. We used binning as described in Sect. 6.3 to check the items for monotonicity, dividing the respondents in the training data into 8 bins. All the 12 items that had passed the check for unidimensionality passed the monotonicity check.

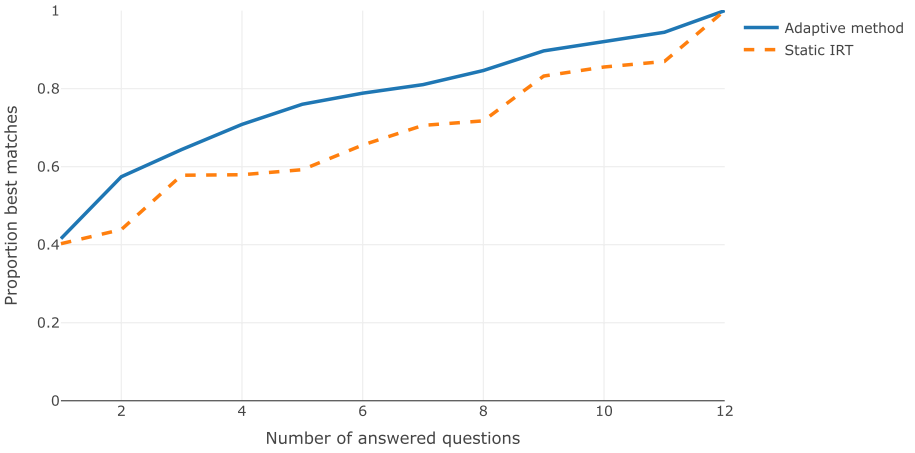

Fig. 3 The proportion of respondents (y-axis) who get the correct top-ranked party when n items (x-axis) have been answered. We compare the proposed adaptive method (solid) to static-order IRT (dashed)

8.2.2 Comparing the models

To identify the best matches according to our definition, we matched the respondents to the parties using tests that included all 12 items. We did this separately for the fixed-order model and for our proposed method. Note that the fixed-order estimates and the estimates for our adaptive method were not always the same, even when all items were included. This was because we used model based MAP priors for the adaptive model and a standard normal MAP prior for the fixed-order model. When we had established the best party for each respondent based on all items, we matched respondents to parties based on 1, 2,.., 12 items.

8.2.3 Results

The results from the VAA data are shown in Fig. 3. We see an advantage of our proposed method over static-order IRT, which is largest for users who answer approximately 5 items and then conclude the test. Of these users, 79 percent find their best match with our method compared to 63 percent with static-order IRT. A table that corresponds to Fig. 3 can be found in the supplemenatry materials.

9 Conclusion and discussion

Long VAA questionnaires can lead to respondent drop-off and low-quality responses due to time constraints and survey fatigue. An option to answer fewer items and still get an accurate result would make VAAs attractive to a broader group of respondents. In this paper we propose a new method, based on Item Response Theory, that makes this type of VAA feasible. The method takes advantage of the fact that some items provide more relevant information than others about the respondent. Which item that is more useful depends on previous answers and is determined dynamically throughout the test.

Thus, different respondents can be presented with the items in a different order. Whereas traditional fixed-order IRT models are primarily useful for VAAs of predetermined length, the method proposed here is more accurate for users who conclude the test without answering all questions.

Transparency requires that the respondent is given a straightforward interpretation of the test results. Our model organizes the question items into multiple unidimensional scales, each of which can be understood in terms of the included question items. This gives the results the interpretability of a unidimensional model.

A consequence of using the proposed method is that the items cannot be organized in topic blocks where a block is held together in the sequence. A reason for using blocks can be that they allow respondents to easier transition from item to item, as considering one item in a block may help in establishing a positions on other related items. It may also encourage consistency in the responses, since a respondent will be reluctant to give responses that are hard to reconcile if the items are close together in the sequence.

The method proposed here prevents the test designer from determining the item order, which is instead determined dynamically based on previous responses. After each response, the method determines which item in the item pool that is most useful, and this item is presented next.

As the method optimizes item selection for the purpose of low-dimensional matching, high-dimensional matching should be treated separately. For instance, a VAA that presents both low-dimensional and high-dimensional results may present only the low-dimensional result when a respondent concludes the test without answering all items. Adding this option to conclude the VAA with fewer answered items can lower the threshold for VAA participation and increase user satisfaction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11135-024-01845-6.

Acknowledgements I wish to express my gratitude to my primary supervisor Frank Miller, who guided me through the research that resulted in this paper, and to my supervisor Ellinor Fackle Fornius, who has greatly contributed to the structure of the paper. Maria Andreasson at the SOM institute provided highly valuable help that gave access to data collected through the Swedish citizen panel. The full and detailed feedback from the anonymous reviewers greatly contributed to improving the paper.

Funding Open access funding provided by Stockholm University. The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Albertsen, A.: How do the characteristics of voting advice application users change over time? Evidence from the German election studies. German Polit. **31**(3), 399–419 (2022)

Ark, L.A.: Mokken scale analysis in R. J. Stat. Softw. **20**, 1–19 (2007)

Baker, F.B., Kim, S.: Item response theory: parameter estimation techniques. (2nd ed., rev. and expanded.) M. Dekker, New York (2004)

Chalmers, R.P.: MIRT: a multidimensional item response theory package for the R environment. J. Stat. Softw. **48**, 1–29 (2012)

Cook, C., Heath, F., Thompson, R.L.: A meta-analysis of response rates in web- or internet-based surveys. Educ. Psychol. Measur. **60**(6), 821–836 (2000)

Crawford, S.D., Couper, M.P., Lamias, M.J.: Web surveys: perceptions of burden. Soc. Sci. Comput. Rev. **19**(2), 146–162 (2001)

de la Torre, J., Patz, R.J.: Making the most of what we have: a practical application of multidimensional item response theory in test scoring. J. Educ. Behav. Stat. **30**(3), 295–311 (2005)

Deutskens, E., de Ruyter, K., Wetzels, M., Oosterveld, P.: Response rate and response quality of internet-based surveys: an experimental study. Mark. Lett. **15**(1), 21–36 (2004)

Fossen, T., Anderson, J.: What's the point of voting advice applications? Competing perspectives on democracy and citizenship. Elect. Stud. **36**, 244–251 (2014)

Ganassali, S.: The influence of the design of web survey questionnaires on the quality of responses. Surv. Res. Methods **2**(1), 21–32 (2008)

Garzia, D., Marschall, S.: Research on voting advice applications: state of the art and future directions. Policy Internet **8**(4), 376–390 (2016)

Garzia, D., Marschall, S.: Voting advice applications. In: Oxford research encyclopedia of politics, Oxford University Press (2019)

Germann, M., Mendez, F.: Dynamic scale validation reloaded. Quality Quant. **50**(3), 981–1007 (2016)

Germann, M., Mendez, F., Gemenis, K.: Do voting advice applications affect party preferences? evidence from field experiments in five European countries. Polit. Commun. **40**(5), 596–614 (2023). https://doi.org/10.1080/10584609.2023.2181896

Germann, M., Mendez, F., Wheatley, J., Serdült, U.: Spatial maps in voting advice applications: the case for dynamic scale validation. Acta Politica **50**(2), 214–238 (2015)

Herzog, A.R., Bachman, J.G.: Effects of questionnaire length on response quality. Public Opin. Q. **45**(4), 549–559 (1981)

Le, A., Han, B.H., Palamar, J.J.: When national drug surveys "take too long'': an examination of who is at risk for survey fatigue. Drug Alcohol Depend. **225**, 108769 (2021)

Lobo M.C., Vink M., Lisi M.: Mapping the political landscape: a vote advice application in portugal. In: Cedroni L., Garzia D. (eds.) Voting Advice Applications in Europe: The state of the art, pp. 143–171. ScriptaWeb, Napoli (2010)

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

Louwerse, T., Rosema, M.: The design effects of voting advice applications: comparing methods of calculating matches. Acta Politica **49**(3), 286–312 (2014)

Manfreda, K.L., Batagelj, Z., Vehovar, V.: Design of web survey questionnaires: three basic experiments. J. Comput-Mediat. Commun. **7**(3), 731 (2002)

Marschall S (2008) The online making of citizens: Wahl-O-Mat. The making of citizens in Europe: New perspectives on citizenship education (S. 137-141). Bonn: Bundeszentrale fur politische Bildung

Michel E, Cicchi L, Garzia D, Ferreira Da Silva F, Trechsel AH (2019) euandi2019 : project description and datasets documentation. Working Paper, European University Institute

Mokken, R.J.: A theory and procedure of scale analysis. The Hague, The Netherlands Mouton. MokkenA Theory Proced. Scale Anal. **62**(3), 331–347 (1971)

Molenaar, I.W.: Mokken Models. In Handbook of Item Response Theory, Chapman and Hall CRC, Boca Raton (2016)

Munzert, S., Ramirez-Ruiz, S.: Meta-analysis of the effects of voting advice applications. Polit. Commun. **38**(6), 691–706 (2021)

Oleskog Tryggvason P (2018) Rapport 2018:4, Utvärdering av Sveriges Televisions valkompassfrågor 2018

Reckase, M.: Multidimensional Item Response Theory. Springer, New York (2009)

Reckase, M.D.: An interactive computer program for tailored testing based on the one-parameter logistic model. Behav. Res. Methods Instr. **6**(2), 208–212 (1974)

Samejima, F.: Estimation of latent ability using a response pattern of graded scores. ETS Res. Bull. Ser. (1968). https://doi.org/10.1002/j.2333-8504.1968.tb00153.x
Samejima, F.: Graded Response Model. In: van der Linden, W.J., Hambleton, R.K. (eds.) Handbook of Modern Item Response Theory, pp. 85–100. Springer, New York (1997)
Sijtsma, K., van der Ark, L.A.: A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. Br. J. Math. Stat. Psychol. **70**, 137–158 (2017)
Sijtsma, K., WMolenaar, I.: Introduction to nonparametric item response theory. SAGE Publications Inc, Thousand Oaks, California, US (2002)
Spaceu2019: What is spaceu2019? http://spaceu2019.eu/about.html (2024). Accessed 22 Feb 2024
Steppat, D., Castro Herrero, L., Esser, F.: Selective exposure in different political information environments–How media fragmentation and polarization shape congruent news use. Eur. J. Commun. **37**(1), 82–102 (2022)
Stout, W.: Psychometrics: from practice to theory and back. Psychometrika **67**(4), 485–518 (2002)
Svenberg, J., Nyman, J., Lindwall, E.: Kompasserna viktigare när medierna tar ut kursen i valbevakningen. SE, DN (2022)
SVTNyheter (2022) Allt du behöver veta om SVT:s valkompasser. *SVT Nyheter*
Trechsel AH (2011) EU-Profiler : positioning of the parties in the European elections
Tukey, J. W.: Curves As Parameters, and Touch Estimation. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 4.1, pp. 681–695 (1961)
Vicente, P., Reis, E.: Using questionnaire design to fight nonresponse bias in web surveys. Soc. Sci. Comput. Rev. **28**(2), 251–267 (2010)
Villar, A., Callegaro, M., Yang, Y.: Where am i? A meta-analysis of experiments on the effects of progress indicators for web surveys. Soc. Sci. Comput. Rev. **31**(6), 744–762 (2013)
Walgrave, S., Nuytemans, M., Pepermans, K.: Voting aid applications and the effect of statement selection. West Eur. Polit. **32**(6), 1161–1180 (2009)
Wall, M., Krouwel, A., Vitiello, T.: Do voters follow the recommendations of voter advice application websites? A study of the effects of kieskompas.nl on its users' vote choices in the: Dutch legislative elections. Party Polit. **20**(3), 416–428 (2010)
Watson, R., van der Ark, L.A., Lin, L.-C., Fieo, R., Deary, I.J., Meijer, R.R.: Item response theory: how Mokken scaling can be used in clinical practice. J. Clin. Nurs. **21**(19–20), 2736–2746 (2012)
Xu, X., Douglas, J.: Computerized adaptive testing under nonparametric IRT models. Psychometrika **71**(1), 121–137 (2006)