

# Methodological challenges in the analysis of voting advice application generated data

Fernando Mendez<sup>1</sup>, Kostas Gemenis<sup>2</sup> and Constantinos Djouvas<sup>3</sup>

<sup>1</sup>University of Zurich

<sup>2</sup>University of Twente

<sup>3</sup>Cyprus University of Technology

**Abstract**—Voting advice applications (VAA) have become an increasingly popular feature of electoral campaigns. VAAs are online tools that use survey techniques to measure the degree to which the policy preferences of citizens match those of political parties or candidates. In some cases, such as The Netherlands, VAA's can attract millions of respondents providing an incredibly rich source of mass public opinion data. As a result political scientists have begun to exploit such datasets and this is fuelling a burgeoning literature on the topic. To date, however, there has been surprisingly little research on the cleaning techniques used to filter out the many rogues entries that are known to be present in VAA generated datasets. This paper presents the various methods used for cleaning VAA generated datasets that have been used for empirical research. Two main techniques are used based on item response timers and pattern recognition techniques. We show why cleaning matters and the problems that flow from not establishing rigorous cleaning techniques. The problem as such is not exclusive to VAA data but is common to all web based research involving self-administered surveys. To that end the techniques we present could be generalisable beyond the specific case of VAA-generated datasets.

## I. MOTIVATION

Voting Advice Application (VAA) is a generic term for the, freely available on the internet, tools that match the preferences of voters to that of candidates, political parties, or other stakeholders. The mechanism of the VAA is simple. Before an election, a team of researchers formulates a number of statements on policy issues that are politically salient and determines the positions of parties and/or candidates on each of these issues [8], [10], [25]. Citizens (hereafter referred to as VAA 'users') are then invited to answer the same questionnaire of policy issues, and their responses are compared to those of parties and/or candidates. VAAs use various matching algorithms and visualization techniques to present how well the users' policy preferences match to those of different parties and/or candidates [11], [17], [20], [26].

VAAs have enjoyed a growing popularity across Europe, where university consortia, media organizations, NGOs, and government-sponsored civic education institutes have designed over two dozen such applications. In some countries VAAs have become an integral part of election campaign with an estimated number of users reaching 40 per cent of the electorate (The Netherlands 2012), or as high as 6.7 million in (Germany 2009) [18]. From the perspective of VAA designers, the presumption is that, by communicating information

about the policy positions of parties and/or candidates and by providing the matching output, VAAs can work as voter information tools that could help citizens make more informed decisions when casting a vote [2], [7], [23]. Moreover, it is often argued that the information VAAs provide may boost political participation such as electoral turnout [6], [9], [16], [19].

Unsurprisingly, many inferences regarding the design and consequences of VAAs are drawn from analyzing VAA users' log-files (hereafter referred to as 'VAA-generated data'). Despite the normative and practical importance of these inferences, researchers do not put the quality of VAA-generated data under close scrutiny very often. This is rather unfortunate, considering that, unlike other online surveys, VAAs are freely available and open to the public which implies that anyone can answer the VAA questionnaire more than once, often by generating responses at random.

We label such behaviour as 'rogue' responses and argue that they cannot be fully explained by the measurement errors associated with the cognitive demands placed on survey respondents [14]. As the respondents are rewarded with a rich output in terms of visuals (bar charts, scatterplots, spider plots) after completing the VAA questionnaire, they have an incentive to 'play around' with the tool and generate multiple (meaningless or meaningful) response patterns in order to see how different responses affect the generated outcome. This tendency results in a considerable over-reporting of the number of users but, most importantly, has implications regarding the inferences made by analyzing VAA-generated data. Our paper looks into this problem by identifying the strategies in trying to filter out such 'rogue' responses in the empirical VAA literature and exploring different methods for doing so more effectively.

## II. REVIEW OF CURRENT PRACTICES

We begin with a review of current practices with regards to cleaning VAA-generated data. In a literature search we identified 14 published studies in peer-reviewed journals that have employed VAA-generated data in their analyses. For reasons of space, we excluded studies published in edited volumes and conference proceedings. These 14 studies are summarized in Table I. In half of the studies the VAA data are used to study VAAs themselves, namely to study the demographic and

attitudinal characteristics of VAA users [4], [12], or to evaluate different measures and outputs of presenting the matching results to VAA users [2], [11], [13], [17], [20]. The other half of studies addresses substantive questions in political science pertaining voter turnout [5], [6], party-switching [1], [15], [27], and the dimensionality of political space [28], [29]. All of the studies, however, use VAA-generated data to draw inferences about the population of VAA users more generally, or in a particular country.

| Study                  | Data source (N)                    | Cleaning approach |    |    |    |    |    |    |
|------------------------|------------------------------------|-------------------|----|----|----|----|----|----|
|                        |                                    | IP                | GL | TT | QT | NO | SA | OI |
| Alvarez et al. [1]     | EU Profiler (15,234)               |                   |    |    |    | X  |    | X  |
| Alvarez et al. [2]     | EU Profiler (15,528)               |                   |    |    |    | X  |    | X  |
| Çarkoğlu et al. [4]    | Oypusulasi (58,579)                |                   | X  |    |    |    |    |    |
| Dinas et al. [5]       | EU Profiler (19,828)               |                   |    |    |    | X  |    | X  |
| Fivaz and Nadig [6]    | Smartvote (27,320)                 |                   |    |    |    |    |    | X  |
| Germann et al. [11]    | Smartvote (20,954)                 |                   |    |    |    |    |    | X  |
| Hooghe & Teepe [12]    | Wij kiezen partij voor u (205,811) | X                 | X  |    |    | X  |    |    |
| Katakis et al. [13]    | Choose4Greece (75,292)             | X                 |    | X  |    | X  |    |    |
| Ladner et al. [15]     | Smartvote (18,290)                 |                   |    |    |    |    |    | X  |
| Louwerse & Rosema [17] | Stemwijzer (10,000)                |                   |    |    |    | X  |    |    |
| Mendez [20]            | See Wheatley [28] below (53,399)   | X                 |    | X  | X  |    | X  |    |
| Wall et al. [27]       | Kieskompas (4,257)                 |                   |    |    |    |    |    | X  |
| Wheatley [28]          | Meuvoto (19,069)                   |                   |    |    |    |    |    |    |
|                        | Perú Escoge (40,627)               |                   |    |    |    |    |    |    |
|                        | Scottish Vote Compass (14,864)     | X                 |    | X  | X  |    | X  |    |
|                        | Cyprus Vote Compass (5,470)        |                   |    |    |    |    |    |    |
| Wheatley et al. [29]   | Scottish Vote Compass (12,053)     | X                 | X  | X  | X  |    | X  |    |

Notes: IP: subsequent entries from the same IP-based identifier; GL: entries from outside the country of interest; TT: total response time across the questionnaire; QT: response time per question; NO: number of no-opinions; SA: number of identical consecutive answers; OI: data from an opt-in survey. In the Mendez and Wheatley studies no timers used in the Brazilian data.

TABLE I  
PUBLISHED STUDIES USING VAA-GENERATED DATA.

Turning to how the VAA data was cleaned in these studies, one can identify at least seven different approaches that are not mutually exclusive. In practice, these approaches are often used in certain combinations which reflect constraints in the data generation process. Four of these approaches require the use of what is known as ‘paradata’ in the computer/online

survey literature [21]. Paradata in the context of VAAs can be further distinguished into IP identifiers and response latency timers. The former are used by most VAAs and consist of a cookie that is installed in the user web browser. The cookie transforms the IP number of the user into a unique identifier which is in turn recorded in VAA datasets. This identifier can be used to clean out multiple responses coming from the same computer (IP), or users coming from a geographic location other than the country that is being launched (GL). Of course, none of these two approaches is perfect. The same user might be using different computers (for example one at home and another at work) to access the VAA and therefore the IP-based identifier will not be able to detect the subsequent entries as responses coming from the same individual. Moreover, the reverse can also occur. Different individuals can use the same computer to respond to the VAA questionnaire, such as different students in a university library computer, or different members of a family in a home computer or tablet. Some studies aim to prevent filtering out these legitimate responses by looking at the demographic variables responses in entries from the same IP-identifier [28]. Including responses coming from a geographic location other than the country which the VAA is launched largely depends on the electoral context and the nature of the research question addressed by the analysis. Responses from outside the country may be analytically meaningful and useful in contexts such as the elections to the European Parliament where citizens can vote in a member-state other than the one mentioned in their passport, or in countries with large diaspora population and postal voting rights. Nevertheless, it is always possible for foreigners to produce an overwhelming number of irrelevant entries especially when VAAs are available in widely-spoken languages and are reported in foreign media. In such instances it would be advisable to clean out such observations.

A second type of paradata used to clean VAA-generated datasets are response latency timers [3]. These can be further differentiated in timers that measure the total time a user has spent in answering the 30-odd questions in a VAA (TT), and timers associated with each of these questions (QT). Andreadis [3] recently concluded on the basis of a thorough analysis that it is impossible to properly clean VAA data out of rogue entries without the use of individual question timers. The logic of this argument is quite simple and intuitive. The psychology of survey response tells us that there is a certain cognitive process that respondents go through when they are confronted with the task of giving answers to questions about their attitudes [21]. Since surveys are often a burdensome task, respondents might skip some of these steps by engaging in what is known as satisficing, giving an answer that merely appears to be satisfactory [14]. Reading the question is the very first task of this cognitive process. It is safe to assume therefore that entries from respondents who have not even spent the necessary time in reading the VAA question, should be filtered out as rogue entries. Andreadis proposed the idea of a threshold for latency timers based on the time needed to ‘skim’ through a question taking into account the question’s length and

linguistic complexity. Given that most VAA questions are relatively short, there is little variation in readability measures typically used in linguistics to measure such complexity. This means that, in practice, responses to VAA questions under two or three seconds can be considered rogue. The use of a total timer implies that one can establish a cleaning threshold by has to multiplying the number of questions in the main VAA questionnaire by two or three. A threshold of 100–120 seconds to answer 30 questions, however, may miss several rogue entries. Respondents are known to engage in satisficing towards the end of a questionnaire. Given that answering 30 attitudes answers in a row can be burdensome for many, respondents might take time to answer the first two dozen or so questions and rush through the last few. Individual timers would have classified this response as a rogue, but a total timer would not.

Unfortunately, most VAAs do not collect response latency timers (let alone individual question timers). This means that studies that do not have access to such paradata can only hope to be able to clean VAA datasets by looking at user response patterns. Studies employing this strategy often use the number of missing ‘no opinion’ responses (NO), or the number of consecutive identical responses (SA). Of course, cases with missing responses might be excluded from statistical analyses that rely on listwise deletion automatically. Nevertheless, researchers may be also motivated to filter out some cases that exceed a certain number of ‘no opinions’. In some extreme cases, users might respond to all questions in the main VAA questionnaire as ‘no opinion’. Louwerse & Rosema [17] for instance, filter out such cases, whereas Hooghe & Teepe [12] and Katakis et al. [13] set a threshold at 7 (out of 42) and one (out of 30) no opinions respectively.<sup>1</sup> Looking the number of identical answers is another pattern-based strategy to identify rogue entries [14]. Provided that VAA designers have produced a balanced questionnaire to avoid acquiescence bias effects, a large number of successive identical answers could identify an observation in the dataset as a rogue entry. Some studies have therefore used a 15 (out of 30) successive identical statements as a threshold to determine rogue entries [20], [29], [28].

Despite all the above measures to identify rogue entries, several studies use a totally different strategy. Most VAA datasets, in addition to the 30-odd questions that are used for matching users to parties and/or candidates, feature responses to additional optional questionnaires. We refer to those responses as ‘opt-in’ data. VAAs use different approaches to generating opt-in data. Some, such as the EU Profiler simply include a ‘Help Our Research’ button visible on the main and results page. Users who click on that button are redirected in an additional questionnaire. Other VAAs such as Smartvote and Kieskompas allow users the option to give our their personal

email. VAA designers then send users invitations to follow-up questionnaires which users have the option to complete. The argument here is that VAA users are not going to fill-in such an optional questionnaires if they are just playing around with the VAA website in order to generate matching output. The responses of these users are often considered legitimate without any further qualification [11], [27]. Moreover, relying exclusively on dataset observations of users who have opted-in these additional questionnaires is often necessitated by the nature of the research question. Very often, the variables that are of interest to the researchers are only available in these opt-in surveys (e.g. [1], [2], [5], [27]).

Having described the established practices in using/cleaning VAA-generated data, our next task is to think of the implications of these strategies in terms of drawing inferences about VAA users. Considering that rogue entries usually represent random clicks through the VAA interface, one could conclude that using a full uncleaned VAA dataset would only add some noise in the statistical analysis. Moreover, considering that rogue entries usually represent less than 5% observations in VAA datasets, and that these datasets usually contain between 50,000–200,000 entries, one could conclude that the inclusion of rogue entries would leave any statistical analysis virtually unchanged. There is another side to this argument, however. Rogue entries may not always contain random responses. For instance, one of the studies reported that the dataset they employed contained 30,000 entries from the same IP address [17]. Moreover, these entries had responses that were exactly identical to one another. Fortunately, these were part of a largest VAA-generated dataset from a single country to date that contained a total of 4,2 million entries. One can easily imagine, however, that had these entries been part of a smaller dataset, they would have introduced considerable bias in any statistical analysis. We therefore conclude that VAA-generated data cannot be left uncleaned.

Relying exclusively in opt-in data promises to solve this problem of cleaning by downsizing the VAA dataset to responses coming from the most motivated users. We argue that this could be equally questionable in the sense that studies who do this might be ‘throwing out the baby with the bathwater’. To give an indication of the size of such downsizing, one could consider the studies based on EU Profiler opt-in data where a dataset of nearly 900,000 respondents is reduced to less than 20,000. The problems introduced by such downsizing are considerable. Even though the statistical power of analyses is not affected considerably, the representativeness of the resulting dataset is. Since not all citizens are able or willing to use VAA websites, this self-selection implies that inferences drawn from VAA-generated data cannot be generalized to the population of a country. End users of VAA-generated data are well aware of this limitation, but the reliance on opt-in data implies a double, or even triple self-selection in the case of opt-in surveys that are sent via emails that have been previously solicited from users (see [22]). Responses in opt-in surveys are less likely to be rogue, but also more likely to come from highly motivated, politically interested and

<sup>1</sup>More curiously, Carkoğlu et al. [4] claim that their VAA had more than 190,000 visitors but include in their analysis only the 73,041 visitors who ‘provided answers of any sort’. Needless to say, this process does not qualify as data cleaning. Those who visit a website but do not provide any answer whatsoever (i.e. leave after viewing the entry page) do not qualify as ‘VAA users’. Unfortunately, some VAA designers often present website visitors as ‘users’ in a deliberate effort to boost their VAA’s ‘usage’ figures.

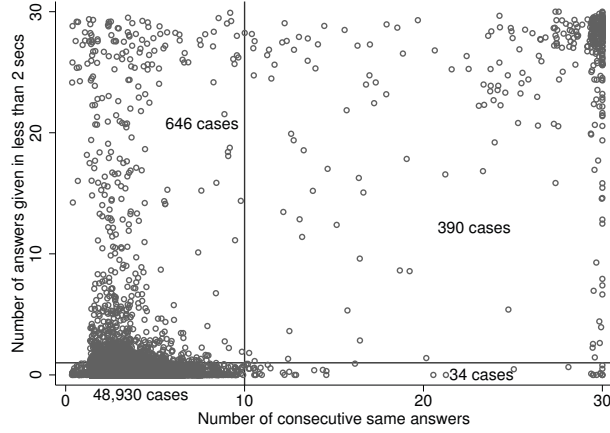


Fig. 1. Identifying rogue entries using timers and response patterns in Parteienavi data.

educated users. To put it simply, it is questionable whether the inferences drawn from opt-in surveys can be even generalized to the level of VAA users. Surprisingly, the problem of double self-selection is only addressed in one study that relies on opt-in data. Dinas et al. use a Heckman selection model to investigate whether their inferences are robust to such double self-selection [5]. Unfortunately, the instrument they use to explain self-selection into the opt-in survey—‘number of no opinion responses’—is invalid because it does not satisfy the exclusion criterion.<sup>2</sup> If VAA-generated data has to be cleaned, but reliance in opt-in surveys can lead to biased inferences, timers and response patterns are the only viable strategies left. In the following section, we further explore pattern-based strategies for cleaning in an attempt to increase the usefulness of a large number of VAA datasets that have not employed response latency timers on individual questions and are therefore difficult to clean.

### III. EMPIRICAL ANALYSIS

Our data come two VAAs: Parteienavi launched for the 2013 German federal election and Choose4Greece launched for the 2014 regional elections. Both VAAs were designed and implemented by the PreferenceMatcher consortium (<http://www.preferencematcher.org>) in collaboration with researchers at the University of Konstanz (Parteienavi) and the University of Macedonia (Choose4Greece). In Fig. 1 we present the distribution of entries according two variables that are often used to identify rogue entries. We adopt a minimal definition of a rogues as those entries that include more than one (out of 30) questions answered in less than 2 seconds.

As evident from Fig. 1, using response patterns such as the number of identical consecutive answers can identify

<sup>2</sup>See [24]. The number of no opinions is a well-known proxy of political sophistication and political sophistication is known to be related to turnout (the dependent variable in Dinas et al.). This renders the instrument invalid and the results of the Heckman selection model approach questionable.

only approximately one third of the rogue entries defined by the individual question timers. In order to investigate the usefulness of response patterns further, we conducted an exploratory analysis on the first 50,000 users to asses whether users’ answer patterns across the entire range of VAA policy statements could be used to predict rogues according to our definition above. We used the ‘caret’ package in R to model a binary classification problem with rogue versus valid user as the output to predict and users’ answer responses to the 30 policy questions as input features. Two machine learning classifiers were used to train the models, a binary logistic model and decision trees. The model training was performed on a training set that represented 70 per cent of the samples, leaving 30 per cent as the test set. The results based on users’ answer patterns were disappointing as can be seen in the confusion matrix. Although overall accuracy was high, 94 per cent for two classifiers respectively, sensitivity was very poor. Neither classifier was able to correctly identify more 5 per cent of rogues. This is not to imply that machine learning techniques cannot be applied to identify rogue pattern behaviour, for instance by using more sophisticated classifier, but rather to illustrate that the exercise is not at all straightforward.

| Prediction | Reference |       |
|------------|-----------|-------|
|            | Rogue     | Valid |
| Rogue      | 48        | 12    |
| Valid      | 770       | 14169 |

Accuracy : 0.9479; CI : (0.9442, 0.9514);  
Kappa : 0.1027; Sensitivity : 0.05868;  
Specificity : 0.99915

TABLE II

CONFUSION MATRIX AND STATISTICS BASED ON A BINARY LOGISTICS REGRESSION MODEL

| Prediction | Reference |       |
|------------|-----------|-------|
|            | Rogue     | Valid |
| Rogue      | 45        | 8     |
| Valid      | 773       | 14173 |

Accuracy : 0.9479; CI : (0.9443, 0.9514);  
Kappa : 0.0973; Sensitivity : 0.055012;  
Specificity : 0.999436

TABLE III

CONFUSION MATRIX AND STATISTICS BASED ON A DECISION TREES CLASSIFIED

In a second attempt to analyze rogue users’ online behavior we drew on the Choose4Greece dataset where we have trialled recording users click through patterns. By tracking and recording users’ actions in the results page, one can extract useful information about different aspects of their behavior through a detailed analysis of the exhibited navigation patterns. In our case, the actions recorded consist of all mouse clicks a user performs in the results page. Each such action is captured by the triplet: tab, item, and timestamp. Tab is used to identify the webpage’s tab where the user performed a click and was used to disambiguate item under different tabs (e.g., legend item appears in multiple tabs), the item is a webpage object under the active tab, and finally the timestamp is the time elapsed since the beginning of the session, i.e., the time the results



page is loaded.

Clicks are split into two distinct categories, the receptive and the unreceptive clicks. The former includes all clicks on items that trigger an event resulting in webpage appearance change; the latter contains all clicks that have no effect and are ignored by the web browser. In this section, we use the aforementioned events to analyze the valid and invalid (rogue) user behavior in the result page as the first step for identifying invalid data entries using users' behavior in the results page.

For doing that, we used three different statistics:

- Clicks: This is the number of clicks a user performed in the results page. As mentioned above, these clicks are split into receptive and unreceptive.
- Time (Seconds): This is the total time a user spends in the results page. This is the same for both event types since this is the timestamp of the last click.
- Interval: This is the frequency (time interval) between two successive clicks.

All the above results are grouped by event type (receptive, and both receptive and unreceptive), and by user type (valid and invalid). The results are presented in the table in Annex 1 from which we can extract the following:

- The number of clicks (of both event groups) the two user types is similar.
- The time valid users spend in the results page is considerably longer.
- The time interval between successive receptive events performed by both types of users is similar. However, the time interval between any event type (both receptive and unreceptive) depicts different behavior between the two user types. That is, invalid users have considerably smaller time interval between two successive events. This suggests that invalid users (in contrast to valid users) in-between receptive events perform other unreceptive events. At a first glance, one can say that this is because invalid users do not pay sufficient attention towards understanding the webpage and simply rush through clicking on any item of the webpage, a behavior somewhat similar to the one they exhibited during answering the questions.

Our preliminary conclusions on this first attempt to map users' behaviour in the results section show some promise since it can be used to identify different forms of rogue behavior. However, more analysis is required for identifying patterns that will identify invalid data with high certainty.

#### IV. CONCLUSION

Although users' behavior showed promise with regard to identifying rogue entries, our exploratory analysis drawing on two machine learning classifiers was unable to identify rogue entries. Further analysis using alternative classifiers, such as anomaly detection algorithms, is one obvious path where we would like to take forward this type of research. For the time being there seems to be little alternative to using response latency timers as the optimal benchmark for identifying rogues.

To that end, VAA designers should consider implementing question-specific timers in VAAs so as to assist end-users in the cleaning of VAA-generated data. In the absence of such paradata, most end users will continue to rely on the opt-in data, a tendency that could compromise the validity of inferences based on VAA-generated data.

#### REFERENCES

- [1] R. Michael Alvarez, Ines Levin, Peter Mair, and Alexander H. Trechsel. Party preferences in the digital age: The impact of voting advice applications. *Party Politics*, 20:227–236, 2014.
- [2] R. Michael Alvarez, Ines Levin, Alexander H. Trechsel, and Kristjan Vassil. Voting advice applications: How useful and for whom? *Journal of Information Technology & Politics*, 11:82–101, 2014.
- [3] Ioannis Andreadis. Data cleaning and data quality. In Diego Garzia and Stefan Marschall, editors, *Matching voters with parties and candidates*. ECPR Press, Colchester, 2014.
- [4] Ali Çarkoğlu, Thomas Vitiello, and Mert Moral. Voter advice applications in practice: Answers to some key questions from turkey. *International Journal of Electronic Governance*, 2012.
- [5] Elias Dinas, Alexander H. Trechsel, and Kristjan Vassil. A look into the mirror: Preferences, representation and electoral participation. *Electoral Studies*, doi: 10.1016/j.electstud.2014.04.011, 2014.
- [6] Jan Fivaz and Giorgio Nadig. Impact of voting advice applications (VAAs) on voter turnout and their potential use for civic education. *Policy & Internet*, 2(4):167–200, 2010.
- [7] Thomas Fossen and Joel Anderson. What's the point of voting advice applications? competing perspectives on democracy and citizenship. *Electoral Studies*, doi: 10.1016/j.electstud.2014.04.001, 2014.
- [8] Kostas Gemenis. Estimating parties' positions through voting advice applications: Some methodological considerations. *Acta Politica*, 48:268–295, 2013.
- [9] Kostas Gemenis and Martin Rosema. Voting advice applications and electoral turnout. *Electoral Studies*, doi: j.electstud.2014.06.010, 2014.
- [10] Kostas Gemenis and Carolien van Ham. Comparing methods for estimating parties' positions in voting advice applications. In Diego Garzia and Stefan Marschall, editors, *Matching voters to parties and candidates*, pages 33–47. ECPR Press, Colchester, 2014.
- [11] Micha Germann, Fernando Mendez, Jonathan Wheatley, and Uwe Serdült. Spatial maps in voting advice applications: The case for dynamic scale validation. *Acta Politica*, doi: 10.1057/ap.2014.3, 2014.
- [12] Marc Hooghe and Wouter Teepe. Party profiles on the web: An analysis of the logfiles of non-partisan interactive political internet sites in the 2003 and 2004 election campaigns in belgium. *New Media & Society*, 9:965–985, 2007.
- [13] Ioannis Katakis, Nicolas Tsapatsoulis, Fernando Mendez, Vasiliki Triga, and Costas Djouvas. Social voting advice applications: Definitions, challenges, datasets and evaluation. *IEEE Transactions on Cybernetics*, 44:1039–1052, 2014.
- [14] Jon A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5:213–236, 1991.
- [15] Andreas Ladner, Jan Fivaz, and Joëlle Pianzola. Voting advice applications and party choice: Evidence from smartvote users in switzerland. *International Journal of Electronic Governance*, 5:367–387, 2012.
- [16] Andreas Ladner and Joëlle Pianzola. Do voting advice applications have an effect electoral participation and voter turnout? Evidence from the 2007 Swiss federal elections. In E. Tambouris, A. Macintosh, and O. Glassey, editors, *Electronic participation*, pages 211–224. Springer, Berlin, 2010.
- [17] Tom Louwerse and Martin Rosema. The design effects of voting advice applications: Comparing methods of calculating results. *Acta Politica*, doi: 10.1057/ap.2013.30, 2013.
- [18] Stefan Marschall. Profiling users. In Diego Garzia and Stefan Marschall, editors, *Matching voters with parties and candidates*, pages 93–104. ECPR Press, Colchester, 2014.
- [19] Stefan Marschall and M. Schultze. Voting advice applications and their effect on voter turnout: The case of the german wahl-o-mat. *International Journal of Electronic Governance*, 5:349–366, 2012.
- [20] Fernando Mendez. Matching voters with political parties and candidates: An empirical test of four algorithms. *International Journal of Electronic Governance*, 5:264–278, 2012.

- [21] Kriste Olson and Bryan Parkhurst. Collecting paradata for measurement error evaluations. In Frauke Kreuter, editor, *Improving surveys with paradata: Analytic uses of process information*, pages 43–72. John Wiley & Sons, London, 2013.
- [22] Joëlle Pianzola. Selection biases in voting advice application research. *Electoral Studies*, 10.1016/j.electstud.2014.04.012, 2014.
- [23] Martin Schultze. Effects of voting advice applications (vaas) on political knowledge about party positions. *Policy & Internet*, 6:46–68, 2014.
- [24] Allison J. Sovey and Donald P. Green. Instrumental variables estimation in political science: A reader’s guide. *American Journal of Political Science*, 55:188–200, 2010.
- [25] Kirsten van Camp, Jonas Lefevere, and Stefaan Walgrave. The content and formulation of statements in voting advice applications: A comparative analysis of 26 vaas. In *Matching voters to parties and candidates*. ECPR Press, 2014.
- [26] Marcus Wagner and Outi Ruusuvirta. Matching voters to parties: Voting advice applications and models of party choice. *Acta Politica*, 47:400–422, 2012.
- [27] Matthew Wall, André Krouwel, and Thomas Vitiello. Do voters follow the recommendations of voter advice application websites? A study of the effects of kieskompas.nl on its users’ vote choices in the 2010 Dutch legislative elections. *Party Politics*, doi: 10.1177/1354068811436054, 2012.
- [28] Jonathan Wheatley. Using vaas to explore the dimensionality of the policy space: Experiments from brazil, peru, scotland and cyprus. *International Journal of Electronic Governance*, 5:318–348, 2012.
- [29] Jonathan Wheatley, Christopher Carman, Fernando Mendez, and James Mitchell. The dimensionality of the scottish political space: Results from an experiment on the 2011 holyrood elections. *Party Politics*, doi: 10.1177/1354068812458614, 2012.

| Statistic | Event Type | User Type | Min.  | 1st Qu. | Median | Mean  | 3rd Qu. |
|-----------|------------|-----------|-------|---------|--------|-------|---------|
| Clicks    | Receptive  | Valid     | 0     | 0       | 3      | 5.661 | 8.5     |
|           |            | Invalid   | 0     | 0       | 4      | 7.495 | 10      |
|           | Both       | Valid     | 0     | 0       | 6      | 10.23 | 13      |
|           |            | Invalid   | 0     | 0       | 4      | 9.017 | 13.5    |
| Time      | Both       | Valid     | 10    | 67      | 123    | 375.4 | 263     |
|           |            | Invalid   | 7     | 41      | 80     | 200.2 | 182.2   |
|           | Receptive  | Valid     | 3     | 215.5   | 575    | 537.6 | 808.8   |
|           |            | Invalid   | 1     | 231     | 505    | 524.8 | 814     |
| Interval  | Both       | Valid     | 1.36  | 8.3     | 13     | 28.42 | 22.8    |
|           |            | Invalid   | 1.889 | 4.393   | 7.024  | 12.78 | 14.61   |

## ANNEX 1