Survey Research Methods (2019)
Vol. 13, No. 2, pp. 181-194
doi:10.18148/srm/2019.v13i2.7392

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

Does mode of administration impact on quality of data? Comparing a traditional survey versus an online survey via a Voting Advice Application

Vasiliki Triga
Cyprus University of Technology

Vasilis Manavopoulos
Cyprus University of Technology

This paper compares two modes of administering an election survey: a traditional, door-to-door survey and an identical online version promoted via a Voting Advice Application. Whereas online political surveys are known to suffer from self-selection bias of politically interested respondents, traditional surveys are plagued with socially desirable responding and are susceptible to the effects of satisficing and other fatigue-related effects. Using a propensity score matching methodology, we examine the extent to which such differences exist between the two modes of administration. While we report mixed findings regarding the structure of respondents' answer patterns, significant differences emerged in relation to social desirability bias with the offline group being more"affected" than the online group.

Keywords: survey modes; online surveys; propensity score matching; mode effects; satisficing; social desirability; voting advice applications

1 Introduction

Obtaining public participation in surveys and ensuring the quality of responding have been a mainstay concern since the advent of structured social inquiry. John Sinclair in the 18th century, to take the perennial example, found that beyond necessitating up to 23 reminders to ensure full participation of all UK ministers in his happiness-related inquiry (de Leeuw, 2005), he had to organize an expedition of "statistical missionaries" to ensure the quality of the answers he received. The problem in its current form regards both increasing costs of traditional survey techniques (de Leeuw & Collins, 1997), decreasing response rates (Holbrook, Krosnick, & Pfent, 2007) and new challenges in obtaining probability samples (Keeter, 2006). On the other hand, advances in communications technology are providing researchers with large amounts of data, potentially alleviating some of the aforementioned problems; relevantly here, in the realm of politics, Voting Advice Applications (hereafter VAAs) are increasingly being used as a potential new data source for analyzing political opinion (Garzia & Marschall, 2014; Germann & Mendez, 2016; Mendez, 2017; Wheatley, Carman, Mendez, & Mitchell, 2014).

VAAs are online tools that provide users who visit the website and fill in the policy questionnaire with measures of how "close" they are to political parties or candidates. The political parties/candidates have been typically positioned on the policy issues by experts[1]. Policy items, which are usually formulated as Likert items, span a range of issues such as "the legalization of same-sex marriage" or 'the need to increase taxation for higher incomes'. Beyond their putative helpfulness for the electorate, VAAs can additionally be useful to researchers, as they involve collecting large amounts of information from willingly-involved participants, presumably motivated to be as accurate or "truthful" as possible with their responses, as the latter affects the feedback they receive. Although not without detractions (see Walgrave, Nuytemans, & Pepermans, 2009; Walgrave, Van Aelst, & Nuytemans, 2008), the large datasets generated by VAAs can have certain analytical advantages including the potential for more in-depth analyses (e.g. focusing on voters of smaller political parties) but also wider analyses, the examination, for example, of understudied populations, such as those of non-English speaking or smaller countries (here for the case of Cyprus, see Marzuca, Serdült, and Welp (2011) for the case of Brazil etc.).

VAAs then involve collecting large amounts of information from prospective voters regarding matters of political interest. "Large amounts of information", however, does not always constitute large amounts of data, that is, information on the basis of which trustworthy inferences regarding social matters can be drawn. This is because a certain amount of Total Survey Error (TSE) can be expected to be involved in VAA-generated datasets. TSE refers to mis-estimation

Contact information: Vasiliki Triga, Dept. of Communication and Internet Studies, 30 Archibishop Kyprianou St., CY-3036, Lemesos, Cyprus (email: vasiliki.triga@cut.ac.cy).

[1] In some VAAs it is the candidates that position themselves rather than experts.

of statistical properties of a population arising from the design, collection, processing and analysis of survey data (see Biemer, 2010; Groves and Lyberg, 2010 for reviews).

An obvious aspect for example is that, since VAAs involve self-selected individuals (i.e. with both access to the internet, related skills and interest in responding to the relevant questionnaire) the datasets produced are affected by a number of sampling-related measurement errors. Indicatively, VAA datasets frequently suffer from non-coverage (i.e. segments of the population with a known to be smaller probability of inclusion, e.g. the elder – Marschall, 2014, non-representativeness (Pianzola & Ladner, 2011) and selection effects (e.g. VAA users tend to be more interested in politics, to be younger etc. – see Marschall, 2014). Although these discrepancies in the profiles of populations reached through online and offline surveys remain large, their magnitude can be reasonably expected to become smaller as new communications technologies proliferate among different parts of the population. Moreover, the use of data pre-processing techniques, such as poststratification weighting to population parameters or propensity score matching to survey weights known to be reliable, can serve to ameliorate some of these problems (e.g. Mendez & Wheatley, 2014; Popp, Horvath, Banducci, Coan, & Krouwel, 2016; Wheatley et al., 2014).

Yet, even though the populations from which VAAs draw their data may gradually become, or made to, converge with those from traditional surveys, some residual TSE is bound to remain, as non-sampling related Error also needs to be taken into account. In this respect, differences between traditional surveys and VAAs can be summarised as arising from two distinct but related sources: differences in how respondents are recruited to participate and differences in the mode of administration (or completion) of the respective questionnaire. While, for example, participation in a traditional survey requires some form of solicitation of the respondent and, at least some minimal engagement with the person conducting the study, VAAs users are self-selected individuals involved in a self- completing capacity. As such, complete control of the mode of administration is being relinquished from the researcher to the participants, individuals who, presumably interested in the feedback that they would receive, complete the VAA. These sources of TSE can lead to differences in outcomes and potential distortion, usually termed in the methodological survey literature Mode-related measurement error (Groves et al., 2004; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). It is important to note that this type of error is not only unknown, if not to an extent unknowable, but is also less amenable to corrective measures than sampling-related error, although some precautionary measures can be taken prior to analyses (e.g. data cleaning).

The aim of this paper is to compare responses of VAA-users to those of individuals recruited using a standard sampling frame and completing a pen-and-paper equivalent to the VAA questionnaire in the presence of an administrator. To explain whether the different mode of administration had an effect in how participants answered the questionnaire, we focus on detecting three types of differences: a) overall response tendencies, for example over-preference for particular response categories in one or the other mode, b) aspects of response quality, namely non-differentiation of answers and random responding and c) overall agreement between offline and VAA respondents to the same policy-related questions. In order to attribute any of these type of differences to the different mode of administering the questionnaire to the online and the offline groups, we undertook a pre-processing analytical step. This involved matching offline respondents to VAA users on a number of demographic and political identity characteristics using propensity score matching. This process enabled us to make the two samples equal in terms of their demographic composition for stricter comparisons between the two modes.

2 Theoretical Background

The reasons for expecting the presence of mode-related discrepancies between the VAA and its pen-and-paper counterpart can be broadly split into two types: cognitive (or psychological) reasons and reasons of a normative nature (Dillman, 2000). We discuss both briefly below.

2.1 Reasons of cognitive nature

In responding to a question using predetermined response options, individuals are involved in a task that cannot be considered cognitively effortless or straightforward. Even granting that respondents merely report internal dispositions, rather than construct attitudes on the spot, a number of steps is postulated to be involved: comprehension of the material presented, recall from memory of relevant information, integration of retrieved information and question to be answered, use of "an appropriate estimation or judgment strategy", formation of a response and reporting (Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000). These requirements, commonly combined with lack of extrinsic motivation, may lead to the exhaustion of respondents' available cognitive resources. In such cases, individuals may withdraw participation or elect to continue completion using cognitive shortcuts instead of careful thinking and accurate responding. This implicit or explicit employment of heuristic-based rather than optimal responding processes triggered by the disparity between the cognitive resources required and those available, is known in the survey literature as "satisficing" (Krosnick, 1991; Vannette & Krosnick, 2013).

Satisficing, notionally an umbrella concept, does not predict a single type of behavior but rather suggests a series of different strategies upon depletion of cognitive resources (see

Vannette and Krosnick, 2013 for a review). Some individuals may withdraw participation, while others may make increased use of the "Don't know" or "N/A" option, if available. Yet others may complete the questionnaire providing answers that are irrelevant to their "true" opinions or behavior by always choosing the first or neutral option offered or consistently "Agree"-ing, regardless of the content of the question. Finally, some respondents may provide arbitrary answers that are reflected in an observable (e.g. non-differentiated, similar responses to all questions), or non-observable pattern, (e.g. providing completely random responses).

Since satisficing is postulated to be dependent upon cognitive resources, we might expect that self- completion survey modes (e.g. mail surveys, online surveys) would encourage less satisficing behavior, as participants can complete the task at their discretion. Empirical findings from comparisons of online and offline modes of administration involving close-ended question however, do not lend ubiquitous support for this hypothesis. Depending on the specific responding behavior under examination, some studies report increased satisficing behavior online, others exhibit the opposite pattern, while others produce mixed results. Starting with "non-differentiation", a satisficing type of behavior that refers to the tendency to provide similar responses to all survey items, Chang and Krosnick (2009, 2010) found that participants interviewed face-to-face provided more undifferentiated responses than participants self-completing an online survey. Heerwegh and Loosveldt (2008), on the other hand, report the opposite finding comparing face-to-face and online administration, in addition to increased non-response rates for their online group. Similarly, Fricker, Galesic, Tourangeau, and Yan (2005) reported less differentiation of responses online, compared to interviews conducted over the telephone.

Regarding other types of satisficing behavior, Dillman, Smyth, and Christian (2009) reported an increased tendency to "Agree" to any given statement ("Acquiescence bias") under interview conditions, compared to both internet and mail respondents. However, Fricker et al. (2005) report no differences in acquiescence bias between online and telephone surveys. Moreover, studies comparing telephone and online surveys found that telephone interviewees responded using extreme response options (e.g. "Completely Agree") more frequently (Dillman et al., 2009; Oosterveld & Willems, 2003). Similarly, other researchers found that online survey respondents provided more neutral and less extreme answers (Duffy, Smith, Terhanian, & Bremer, 2005; Frippiat, Marquis, & Wiles-Portier, 2010). A finding that does seem to exhibit some consistency according to Ye's meta-analysis (Ye, Fulton, & Tourangeau, 2011), is an increased tendency toward extreme positive but not extreme negative responses in telephone and interview conditions compared to online administration.

Although satisficing has become central in the survey literature, the phenomenon remains understudied, especially for data from the general population rather than students (e.g. Heerwegh & Loosveldt, 2008; Kreuter, Presser, & Tourangeau, 2008) or professional groups (Converse, Wolfe, Huang, & Oswald, 2008; Couper & Triplett, 1999). The lack of consensus that characterizes the existing body of literature is not surprising, due to the difficulty in detecting the phenomenon as well as its context specificity. Satisficing can be postulated to affect the universal process of responding throughout or at a specific point in time, making identification of when respondents are engaged in non-optimal responding a difficult task. Moreover, upon engagement of satisficing processes, a number of different responding strategies are available, each of which are connected to a different behavioral pattern and may act synergistically or antagonistically. The well-known increased use of the response categories that appear last ("recency effect") when employing auditory data collection techniques, for example, may help the detection of acquiescence bias or may mask it, depending on whether "Completely Agree" or "Completely Disagree" appears as the first or the last response option offered.

2.2 Reasons of Normative Nature

Most traditional methods of data collection require that participants are involved in some solicited form of interaction with an interviewer or an administrator of the survey. Depending on its magnitude and whether verbal or non-verbal, this interaction can lead to invocation of cultural or societal norms which can constrain responses (de Leeuw, 1992, pp. 29–30). This phenomenon is known as social desirability bias (Crowne & Marlowe, 1964, p. 109). It is postulated to affect responding either through: a) impression management, i.e. conscious editing of responses in order to appear more favorable to the person conducting the study, or even b) self-deception, biased processing of information in a self-fulfilling manner during the response formation stages (Holden, Wood, & Tomashewski, 2001). Whatever the mechanism, this tendency suggests the expectation that some participants may give culture or society-compliant rather than "honest" responses, skewing the results towards more mainstream and accepted positions.

Social desirability bias is generally thought to be related to the social distance between respondent and administrator and/or to the trust toward the conductor of the study. It is accentuated by the more active involvement of the administrator (Green & Tunstall, 1999), and public rather than private administration (Sullman & Taylor, 2010). Therefore, we can expect that increasing the extent of the interaction would invoke more socially-desirable responding since respondents will be more susceptible and sensitive to signs of disapproval (Holbrook, Green, and Krosnick, 2003, pp. 86–87; Tourangeau and Yan, 2007). Consequently, it is unsurprising that surveys employing self-completion methods (e.g. mail surveys) tend to invoke less social desirability bias compared to face-to-face and telephone interviews (de Leeuw, 1992, 2005). A similar pattern is observed in studies collecting data online through PC web and mobile web devices (Mavletova & Couper, 2013) in contrast to both face- to-face (Heerwegh, 2009) and telephone interviews (Chang & Krosnick, 2009; Kreuter et al., 2008). The finding persists even in studies that focus on (nominally) non-sensitive issues, such as politics (e.g. Duffy et al., 2005; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010).

It should be noted that although social desirability bias is presented separately from satisficing, this is not to imply that the two are separate phenomena. Providing socially acceptable instead of deliberative responses is itself a heuristic and is, thus, reflective of satisficing behavior (Krosnick, 1991; Vannette & Krosnick, 2013). Moreover, in the absence of some independent measure of social desirability (e.g. concurrent administration of the Balanced Inventory of Desirable Responding [Paulhus, 1991]) or without having established that a particular question tends to invoke social norms, it is often not possible to attribute any differences to the depletion of cognitive resources or social desirability, which is also the case in the present paper.

With these observations in mind, our interest is in comparing responses obtained from individuals using traditional offline survey methodology and from online VAA users in two respects: a) overall patterns of responding, some of which can be postulated to be indicative of non-optimal responding (e.g. an overall tendency to "agree" to all policy-related questions or use the first response option offered) and b) individuals who are similar to one another in a number of respects provide similar responses online and when using the VAA.

3 Method

3.1 Datasets

The findings reported in this paper are based on analyses of data collected through two distinct routes:

The offline dataset consisted of questionnaires collected between January 23rd and January, 30th, 2013 from 818 individuals whose participation was elicited through door-to-door stratified random sampling with replacement of individuals who could not be contacted. The strata were designed to take into account the regional population density, gender, age and whether the area was urban or rural. Offline participants completed a pen-and-paper questionnaire version of the VAA in a self- completing capacity, although the person administering the questionnaire remained present throughout. In order to impress the anonymity of the procedure, participants were provided with an envelope, which they were invited to seal after enclosing their questionnaire.

The online dataset involved responses from 11,102 VAA users who completed the Choose4Cyprus VAA (approximately two percent of the electorate). The VAA was launched three weeks before the elections, on January, 30th, 2013 through various media channels and online social networks and lasted throughout the pre-electoral period leading to the first round of the 2013 Cypriot Presidential election in February, 17th, 2013. VAA users freely visited the VAA website and completed the relevant questionnaire in a self-completing capacity. Prior to the data pre-processing that is described below, we decided on a perfunctory cleaning of the data from both sources on the basis of a large number of incomplete responses (30 percent) to the respective questionnaire. For the case of the VAA data, we additionally eliminated users with very quick response times (under two seconds in any of the policy items and over three responses in under three seconds) and multiple data entries from the same computer. This left us with 786 respondents offline and 10,408 VAA users overall.

3.2 Stimuli

Both the offline and VAA version of the questionnaire consisted of a number of questions pertaining to policy-related preferences and a number of supplementary questions regarding demographics, political identity (e.g. self-placements on a Left-Right axis) and political behaviour (e.g. vote intention for the upcoming election). The total of policy items in common between the two versions of the questionnaire was 30 although the number of policy questions that were used for our analysis was 19, since these appeared in the same order in both the offline and online version (see Appendix I for details). In both cases the supplementary questions came prior to the policy- related questions. The focus of comparisons are the policy-related questions which were presented as Likert-type items, e.g. question 1 was 'Fiscal deficits should be mainly covered by the additional taxation of wealth'. Six response options presented from left-to-right in the following order (Completely Agree, Agree, Neither agree nor disagree, Disagree, Completely Disagree, plus No Opinion).

3.3 Data pre-processing

As expected the two datasets differed in a number of respects. In accordance with previous findings (e.g. Boogers & Voerman, 2003; Marschall, 2014), the online VAA sample was significantly younger ($t(850) = 13.9$, $p < 0.01$), predominately male ($x^2(1) = 66.6$, $p < 0.001$), more likely to have a university degree ($x^2(1) = 742.5$, $p < 0.01$) and declared themselves to be more progressive ($t(900) = 3.2$, $p < 0.01$). As such, any differences observed could be attributable to sample composition effects, rather than mode-related differences.

In order to enable direct comparison, a number of different data pre-processing techniques were attempted (Exact Matching, Propensity Score Matching, Entropy Balancing Scores weighting). Although these processes are assumed to lead to unbiased estimates and equivalent datasets, in practice this is not always feasible (Duffy et al., 2005; Malhotra and Krosnick, 2007, pp. 293–296), nor was it the case here. Balancing the need to create as "similar" as possible datasets and the desire for, as minimal as possible, information loss, we used Propensity Score Matching (Rosenbaum & Rubin, 1983). More specifically we used its 1:3 variable ratio without replacement variant with a caliper of 0.1 (Caliendo & Kopeinig, 2008) and with exact matching employed for all categorical variables (see below). This technique matches every individual in the offline sample to three (or less) respondents from the online dataset with the most similar propensity score, i.e. the most similar person(s) when considering the matched-for variables. This process culminates in a combined dataset that allows within-subject (pairwise) analyses, where each offline respondent is compared only to their own weighted matches from the VAA sample, a stricter alternative to, e.g. balancing the two groups as to matched-for variables or accounting for the latter's effects by entering them as covariates in a regression model.

The matched-for variables included: a) Demographics: Age, Sex and Education[2]; b) Political Identity: Party Identification, Previous Vote in the previous Parliamentary election (2011) and Vote Intention in the upcoming Presidential election; and c) Ideological Affinity: Respondent self-placement on two eleven-point Left-Right economic and Progressive-Conservative social values scales.

Any participant without valid responses to all of the above variables was eliminated prior to the analyses, as they could not be fully matched, which is the main reason why nonresponse rates in the two modes are not examined in this study. In sum, the procedure yielded an offline (n = 332) and an online (n = 773)[3] subset, with near-zero differences in all "matched-for" variables (see Appendix II). All matching described was conducted using the MatchIt package for R statistical software (Ho, Imai, King, & Stuart, 2007).

3.4 Comparison variables

This paper is concerned with three inter-dependent but distinct aspects of mode-related measurement error which are explained in this sub-section.

Overall Response Tendencies. We first focus on the examination of differences in the univariate distributions of responding behaviours between offline and VAA respondents, in order to examine systematic preference for one type of response category or another in the two modes. We do this by comparing the absolute number of each type of response (i.e. Completely Agree ("CA"), Agree ("A") etc.) in all 19 policy-related items under examination for the two modes. Another tendency we explore is that toward Extremeness, which is the systematic selection of extreme answers. To measure Extremeness tendency, we calculate the sum of "Completely Agree" and "Completely Disagree" responses. Finally, we measure also the Acquiescence bias by calculating the sum of "Completely Agree" and "Agree" responses.

Response Quality Indicators. We then consider two specific aspects of satisficing: providing undifferentiated responses and providing "random" responses, assuming an underlying latent dimension in some of the items. To measure non-differentiation of responses we used the following indices: i) Non-Differentiation Index: The tendency of individuals to respond to all items using the same or a similar response category. The calculation of the non-differentiation index follows Mulligan, Krosnick, Smith, Green, and Bizer (2001) (as reported in Chang and Krosnick, 2009[4]) and takes values between 0 and 1, larger values suggesting more similar responses overall. ii) Maximal length of same consecutive answers. The aforementioned index designed by Mulligan et al. (2001) assumes that identical (or similar) responses to all questions examined are to a degree mutually exclusive. Since the 19 policy-related items examined here are to a much larger degree independent from each other, a larger Non- Differentiation Index might indicate a strong ideological position, instead of satisficing. So, we additionally calculated per respondent the maximal number of same consecutive answers they provided, so that someone who responded "Completely Agree" nine times consecutively would get a score-count of 9, while someone who also used "CA" nine times but only thrice consecutively would get a score of 3.

Since satisficing may also be expressed with random answering behavior, we attempted to construct an index to detect such behavior. Although the 19 policy items of the main questionnaire are theoretically independent, it is natural to expect that they can be grouped into a smaller set of underlying constructs, identified through dimension-reduction statistical techniques. If this is the case, we can expect that individuals respond in a similar manner to all the questions pertaining to the dimension. Someone, for example, strongly against increasing taxation will also be against state interven-

[2] Dichotomised as with or without having completed a University degree.

[3] As it tends to better reduce bias (Ming & Rosenbaum, 2000), variant ratio matching was employed, so not all offline respondents had three VAA matches. After determining the number of VAA matches per offline respondent, each pair was assigned a "matching weight" to reflect the number of matched VAA users (i.e. a weight of 1 for offline respondents with only one match, 0.66 for those with 2 matches).

[4] $\text{NonDiff} = \sqrt{|Q1 - Q2|} + \sqrt{||Q1 - Q3|} + \cdots + \sqrt{|Qi - Q(i-1)|}$ Subsequently, the calculated value per individual is rescaled as $\text{NonnDifferentiation} = \frac{\text{NDi}-\max(\text{ND})}{-\max(\text{ND})}$, in order for higher values close to 1 to indicate most varied responses, while 0 would indicate least non-differentiation (i.e. same response to all questions).

tion in the economy and in favour of privatization of government agencies. It is plausible that deviation from this responding pattern may be indicative of random responding (although this is not necessarily the case).

A single such dimension was identified by applying Mokken scale analysis for ordinal level data, a non- parametric Item Response Theory model (Van der Ark, 2012), using the "mokken" package for R (Van der Ark, 2007). Mokken Scaling Analysis is well-suited for analysing policy preferences of respondents and has been frequently applied to VAA data (Gemenis, 2013; Germann, Mendez, Wheatley, & Serdult, 2015; Katsanidou & Otjes, 2017; Mendez & Wheatley, 2014; Wheatley et al., 2014). The Mokken Scaling Analysis was performed using the original datasets ($n = 11,102$, offline and online combined), after it was reduced to be representative of population parameters as to demographics (sex, age, education) and vote in the 2011 Parliamentary election. The analysis yielded a single dimension. This incorporated issues closely related to what Kriesi et al. (2006) describe as "cultural" dimension of political conflict. Such issues can include the separation of church and state, the decriminalization of drugs, the institutionalization of same-sex relationships as well as the Cyprus conflict, which refers to the division of the Island between the Greek Cypriot and the Turkish Cypriot communities (see Appendix III). The scale included items that needed to be reversed in order to fit with the underlying dimension[5].

Following Meijer (1994), we calculated the weighted Guttman errors (G^*) per respondent, a person-fit statistic that indicates how consistent responses are with respect to responses to all other questions in the scale (i.e., to the dimension itself). It should be noted, however, that such scale-aberrant responding reflects behaviour that is simply non-fitting (not necessarily guessing or random responding). Yet, although inconsistent responding does not necessarily reflect random answers (Krosnick, Narayan, & Smith, 1996), a consistently higher index in one group may reflect increased non-optimal responding (random responding, social desirability effects or otherwise).

3.5 Offline-VAA models of agreement and association

Finally, we examined the joint distributions of offline and VAA responses to examine how similar the responses given in each mode were. Treating the responses to the 19 Likert-type items as ordinal, all offline respondents and their VAA pairs were considered independent raters of the same object and a two-way 5×5 square contingency matrix was calculated per question containing the joint responses by offline and online participants (e.g. CA online ×- CA offline, CA online × A offline etc.). Subsequently we summed up the contingency matrices from each of the 19 questions to create a single matrix of joint distributions of response pairs in order to examine the overall tendency of offline-VAA pairs to agree with each other.

To examine association within this overall contingency matrix, we employed the log linear modelling approach of agreement and association as described in Agresti (2010, pp. 247–250). The logic of this three-step hierarchical approach is to first fit a baseline model that assumes no relationship ("correlation") between responses of offline and VAA respondents in a pair (Independence model). This model's ability to explain patterns in the data is assessed through its deviance (G^2) from the "Saturated model", which perfectly accounts for all cells in the contingency matrix but adds no new information for inference. Subsequently, a "perfect agreement" parameter is added accounting for the cells where offline and VAA respondents have provided an identical response (i.e. the elements of the main diagonal of the contingency matrix) to create the "Agreement model". In addition to assessing the fitness of this model against the Saturated one, the amount of improvement in predictive capacity from the previous step (Independence model) is also tested, using a chi-square test. Finally, an "Agreement plus Linear-by-linear Association" model is calculated by adding an extra parameter to account for similarity of responses off the main diagonal, i.e. where pairs of participants give similar but not identical responses, accounting for the ordinal nature of the data.

The agreement parameter for the models is constructed by simply inserting in the model a vector assigning 1 for all response pairs on the main diagonal (i.e. CD/CD, D/D etc.), while 0 on all other cells of the table (see Agresti, 2010). The linear-by-linear association parameter is constructed taking advantage of the structure of the square matrix by assigning numbers to responses (so CD=1, D=2, N=3) and multiplying the two responses. So, for example, the response pair CD/CD is $1 \cdot 1 = 1$, the response pair D/N is $2 \cdot 3 = 6$, the same as N/D, A/CA is $4 \cdot 5 = 20$ etc. This parameter is, in essence, used to account for pairs of offline and VAA respondents with neighboring but not identical responses.

4 Results

We begin the presentation of results by examining differences in responding behaviour manifested at an omnibus level, which would suggest an overall preference for a response category offline or online (through the VAA). Subsequently, we compare offline and VAA data as to two aspects of satisficing, namely non-differentiation and random responding and finally, we examine the tendency of offline

[5] Without including items that require a reversal of responses, in order to fit with the underlying scale, a respondent who provided completely undifferentiated responses (e.g. all "Completely Agree") might be someone who is ideologically consistent or someone who always responded with the last option offered without reading the content of the questions they were responding to.

respondents to provide similar answers to the same questions as their VAA-counterparts to whom they have been matched.

4.1 Differences in overall response patterns

As detailed in the introduction, we expected mode-related differences in the preference of response categories between the offline and the online mode of administrating the questionnaire. Indeed, we find substantial differences between the two modes in the tendency to employ all response categories over others (see Table 1), with the exception of "Neither agree nor disagree".

Analyses at this omnibus level suggest a greater tendency for offline respondents to provide more "Completely Agree" (CA) ($t(772) = 13, p < 0.001$) and "Agree" (A) responses ($t(772) = 10.4, p < 0.001$) than their VAA-user counterparts to a statistically significant degree. As a corollary, VAA-users tended to provide more "Disagree" (D) ($t(772) = -13.6, p < 0.001$) and "Completely Disagree" (CD) responses ($t(772) = -15.8, p < 0.001$). Providing a different reading of the same results, participants who responded through the VAA in this case exhibited increased tendency toward primacy (since "CA" was the first response option offered) and more significantly toward Acquiescence ($t(770.6) = 26.8, p < 0.001$). Considering Extremeness whoever, the offline-online pairs did not differ to a statistically significant extent ($t(772) = -0.2, p = 0.284$).

A note is warranted on the magnitude of these discrepancies. Although we find some clear systematic over-preference for some response categories in each mode, some care needs be taken to avoid over-interpreting these differences. Not only was there significant agreement between offline and online pairs in general (see Section 4.3), but the magnitude of differences was generally small; in the case of the strongest effect observed for a single response category, that for systematic preference for "CA" in the online condition, the average difference between pairs was two more such responses online out of a total of 19 (10.5 percent more "CAs" online), as significant agreement existed between offline and online pairs in general (see Section 4.3 below).

4.2 Non-differentiation and random responding

Moving on to other measures of satisficing, we find an increased tendency for offline respondents to provide less differentiated responses ($t(771.4) = 7.8, p < 0.001$—see "NonDifferentiation"-Table 1), when taking into account responses to all 19 policy items. Simultaneously, offline participants seem to have provided longer strings of identical responses to subsequent questions (avg. rank where offline>online 334.1; where online>offline 315.1), although it should be noted that this was a marginally non-significant tendency ($Z = -1.93, p = 0.054$) and that online respondents provided longer strings of identical responses more often (8.7 percent) than their paired offline counterparts (see Table 1).

Turning to the examination of the Conservative-Progressive scale obtained from applying a Mokken Scaling Analysis, we find that the responses of offline participants were less likely to adhere to the latent underlying cultural dimension of politics ($t(611.9) = 18.1, p < 0.001$). So offline responses resulted in twice as many Guttman "errors", that is scale-aberrant behaviour. We should reiterate however, that this only reflects scale-aberrant and not necessarily random responding and we find it very plausible that online respondents simply tended to have more structured political attitudes, having decided to engage with a political tool such as a VAA in the first place; in the absence of some way to control for political sophistication, it is impossible to be certain that this result indicates more random responding offline. As an aside, offline participants also provided responses that suggested more conservative or traditional values ($t(741.5) = 5.41, p < 0.001$) albeit to a very small degree (0.2 points on a 10-pt.scale, 2 percent).

4.3 Agreement models

As described in the method section, we constructed a contingency matrix of joint responses of offline respondents and VAA-users in pairs, for all questions altogether (see Table 2). Although some discrepancy can be expected in how two different individuals respond, even if they are similar as to demographics and political characteristics, we also expect substantial agreement between offline respondents and their VAA counterparts. Given this expectation we fit a number of models in order to examine the presence of agreement and association between them. The first model attempted (Independence model) predicts observed frequencies on the basis of the assumption that offline and online responses are independent (i.e. there is no association between responses in the two modes). Although the independence model cannot be reasonably expected to fit the data well, it does establish a baseline to judge the fit of subsequent association models.

Table 3 shows the fit statistics for the independence and sequentially augmented models for agreement and agreement plus linear-by-linear association. The likelihood ratio statistic for the independence model was $G^2 = 189.9$ with 16 degrees of freedom and $p < 0.001$, not a tenable model for predicting the joint distribution of responses by offline and VAA pairs. Adding the agreement parameter to the model produces a G^2 of 81.3 ($\text{df} = 11, p < 0.001$), a clear improvement on the independence model ($\Delta G^2 = 108.6, \Delta\text{df} = 5, p < 0.001$), although still a substantial departure from the saturated model. Particularly noteworthy are higher levels of agreement obtained for extreme ends of the scale while substantial absence of agreement exists for D/D pairs of responses. Further adding the parameter for uniform association yields a G^2 of 24.3 ($\text{df} = 10, p < 0.001$), a statistically significant improvement to both Independence and to the +Agreement model, indicating a positive association be-

Table 1
Differences in response categories between the offline and the online (VAA) modes of administering the questionnaire

	test-statistic				Avg. pair		Means	
Response Pattern	df	t	d[a]	p	difference	Std. Dev.	Offline	Online
numOfCDs	772	−15.8	−0.800	< 0.001	−2.00	0.130	1.90	3.90
numOfDs	772	−13.6	−0.710	< 0.001	−1.80	0.140	3.60	5.50
numOfNs	772	1.6	-	0.761	0.20	0.110	2.70	2.50
numOfAs	772	10.4	0.510	< 0.001	1.40	0.140	5.90	4.50
numOfCAs	772	13.0	0.660	< 0.001	1.90	0.150	4.30	2.40
Acquiescence	771	26.8	0.095	< 0.001	3.40	0.130	10.30	6.90
Extremeness	772	−0.2	-	0.284	−0.05	0.220	6.26	6.31
Scale score	742	5.4	0.400	< 0.001	0.26	0.030	1.70	1.50
NonDifferentiation	771	7.8	0.300	< 0.001	0.04	0.006	0.22	0.18
Guttman Errors	612	18.1	0.980	< 0.001	13.9	0.770	26.80	12.90
MaxSameConsec		−1.9[b]					286[c]	360[d]

[a] Effect size, Cohen's d calculated following Dunlap, Cortina, Vaslow, and Burke (1996). [b] Wilcoxon's signed ranks test used since distribution of differences between modes vary much non-normal. Mean rank of online < offline = 334.1; mean rank of online > offline = 315.11; 204 ties. [c] Number of pairs where offline respondents provided longer strings of identical responses to subsequent policy questions (204 ties). [d] Number of pairs where online respondents provided longer strings of identical responses to subsequent policy questions (204 ties).

Table 2
Models testing agreement and association between offline and VAA respondents

Model	G^2	df	p	ΔG^2	Δ df	p	AIC
Independence	189.9	16	< 0.001				408.7
+Agreement	81.3	11	< 0.001	108.6	5	< 0.001	310.1
+Agreement	24.3	10	< 0.001	57.0	1	< 0.001	255.1

Model	Parameter (off-on)	Estimate	Z-statistic	p	Odds Ratio
+Agreement	Rater agreement (CD)	0.29	4.20	< 0.001	1.33
+Agreement	Rater agreement (D)	−0.14	−3.00	0.003	0.87
+ Linear-By-Linear Association	Rater agreement (N)	0.10	1.46	0.140	1.11
	Rater agreement (A)	0.09	2.03	0.042	1.09
	Rater agreement (CA)	0.02	0.32	0.750	1.02
	Linear-By-Linear Association	0.05	7.58	< 0.001	-

Null model deviance: 3554.8 on 24 degrees of freedom.

tween similar responses between online and offline even off the main diagonal (e.g. CA-offline/A-online, N-offline/D-online etc.).

Overall then we find statistically significant increased probability for higher levels of complete agreement between online and offline respondents and a significant positive association off the main diagonal over and beyond what can be accounted for by the null-association model. Despite the improvement in prediction offered by the final model, it still remains a statistically significant departure from the saturated model. The residuals from the final model allows for cell-wise detection of instances where the model fails in particular. Examining these (see Table 2), we find high positive standardized adjusted residuals for the following pairs of responses D-online/CD-offline, N-online/A-offline and CD-online/CA-offline.

Table 3
Contingency matrix of joint responses of offline respondents and VAA respondents in pairs

	Offline				
Online	CD	D	N	A	CA
CD	452.1[a] -[b]	485.5 (2.80)	173.5 (-0.16)	268.9 (-1.00)	114.5 (-2.20)
D	635.2 (-1.49)	785.8 -	370.2 (-0.05)	648.4 (1.39)	314.7 (-0.05)
N	415.2 (-0.39)	598.4 (-0.73)	288.7 -	454.4 (-0.44)	269.7 (2.03)
A	823.1 (-1.90)	1333.3 (1.28)	598.4 (0.80)	1178.1 -	594.5 (-0.27)
CA	647.3 (3.90)	876.2 (-2.54)	423.0 (-0.69)	838.2 (-0.17)	496.7 -

[a] Weighted observed frequencies in offline-VAA pairs.
[b] Standardised adjusted residuals from Log-Linear model of +Agreement +Linear-by-linear association.

The large presence of the first two pairs of responses (D-online/CD-offline and N-online/A-offline) are less problematic in the sense that they are reflective of the overall responding tendencies reported above in Section 4.1. These refer to an increased Acquiescence in the offline condition and larger tendency for CD-responses than D-responses in the online condition (see the effect sizes of Table 1). Of interest too is the anti-diagonal part of the matrix (e.g. CD-online/CA-offline, etc.), which indicates disharmony between online and offline pairs. While we did find discrepancies for some of the anti-diagonal cells it should be noted that this was the result of disharmony among 40 offline-VAA pairs.

5 Discussion

This study sought to investigate differences in responding behavior on a questionnaire about voters' policy preferences using two distinct modes of survey administration. The first group was solicited in the traditional manner through stratified sampling and completed the survey using a pen-and-paper questionnaire in the presence of an administrator, while the second completed the questionnaire online –a so-called Voting Advice Application– and responded in conditions of their own choosing. This relatively large number of online users (11,102) allowed us to use propensity score analysis to match respondents from the traditional offline survey to the online sample The resulting pairs of respondents (offline and online) were very similar as to their socio-demographics, political orientation and self-reported ideological preferences so that any differences observed would be as close as possible to being able to indicate "pure" mode-related error.

The analysis revealed some interesting results and differences. In terms of overall response patterns, we observed a systematic over-preference of the online respondents for both "Completely Disagree" and "Disagree", the last two response options offered. As a corollary, offline respondents tended to use "Completely Agree" and "Agree" more, i.e. exhibited Acquiescence bias, in line with the findings of Dillman et al. (2009).

Considering responding behaviour more clearly indicative of satisficing, that is, non-optimal responding due to depletion of cognitive resources, we provide findings in line with Chang and Krosnick (2009, 2010) (though cf. Fricker et al., 2005; Heerwegh & Loosveldt, 2008). Namely, individuals in the offline condition tended to provide both less varied responses throughout the 19-item long policy questionnaire and longer strings of identical responses to subsequent questions, though the latter to a marginally non-significant degree. Simultaneously, we find that online respondents provided less scale-aberrant responses, which, assuming a latent ideological dimension, may be (but not necessarily is) indicative of random responding. In a further exploration we applied a Mokken Scaling Analysis to extract an ideological scale resembling a Progressive vs. Conservative dimension from the policy items included in the questionnaire. Here we found that the online group tended to be placed at a slightly more Conservative end of the scale, with offline respondents closer to the middle of the scale, though still on the Conservative pole.

A more positive result concerning the comparability of traditional and online surveys emerged from log linear models which confirmed that matched offline and online respondents provided both more identical and similar responses (e.g. CA/A) than would be expected by chance. In cases of disagreement in the pairs the results partially reflect the aforementioned tendency for offline responses to be more acquiescent. However, we also obtained high residuals in the anti-diagonal elements, i.e. a high number of cases where offline and online respondents provided diametrically different responses, particularly CA-offline and CD-online. This is indeed a more worrisome finding with regards to the possibility of mixing different survey methods.

Ultimately, we cannot know which of the two modes of survey administration produces responses closer to the "real" opinions of the respondents. Our tentative conclusion is that online data provided by the VAA respondents is more likely to be of "better" quality in terms of more accurate responding and less satisficing type answering patterns not to mention a reduced scope for socially desirable responding. In many respects, this is not too surprising since the online group opted-in to the survey and had intrinsic incentives to answer more accurately since they responded in the expectation of feedback based on their responses. On the other hand, VAA data has clear disadvantages with regard to representativeness and non-coverage. As a concluding remark, it is important to keep in mind that it is not unlikely that the reported effects are highly context sensitive and literature, both in support and against of the reported results, suggests the need for more research addressing the issue of comparability between online and offline or traditional survey administration.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. New Jersey: Wiley.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Boogers, M. & Voerman, G. (2003). Surfing citizens and floating voters: Results of an online survey of visitors to political web sites during the Dutch 2002 General Elections. *Information Polity*, *8*(1-2), 17–27.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Chang, L. & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678.

Chang, L. & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, *74*(1), 154–167.

Converse, P. D., Wolfe, E. W., Huang, X., & Oswald, F. L. (2008). Response rates for mixed-mode surveys using mail and e-mail/web. *American Journal of Evaluation*, *29*(1), 99–107.

Couper, M. P. & Triplett, T. (1999). A comparison of mail and e-mail for a survey of employees in US statistical agencies. *Journal of Official Statistics*, *15*(1), 39–56.

Crowne, D. P. & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: John Wiley & Sons.

de Leeuw, E. D. (1992). *Data quality in Mail, Telephone, and Face-to-face surveys*. Amsterdam: TT-Publicaties.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

de Leeuw, E. D. & Collins, M. (1997). Data collection method and data quality. An overview. In L. E. Lyberg, P. B. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwartz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 199–221). New York: John Wiley & Sons.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail and Mixed-Mode surveys*. New York: John Wiley & Sons.

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*(6), 615–639.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Frippiat, D., Marquis, N., & Wiles-Portier, E. (2010). Web surveys in the social sciences: An overview. *Population*, *65*(2), 285–311.

Garzia, D. & Marschall, S. (2014). *Matching voters with parties and candidates. Voting advice applications in comparative perspective*. Colchester: ECPR Press.

Gemenis, K. (2013). Estimating parties' policy positions through voting advice applications: Some methodological considerations. *Acta Politica*, *48*(3), 268–295.

Germann, M. & Mendez, F. (2016). Dynamic scale validation reloaded: Assessing the psychometric properties of latent measures of ideology in VAA spatial maps. *Quality and Quantity*, *50*(3), 981–1007.

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

Green, C. & Tunstall, S. (1999). A psychological perspective. In K. G. Bateman I. J.and Willis (Ed.), *Valuing environmental preferences: Theory and practice of the contingent valuation method in the US, EU, and developing countries* (pp. 207–258). Oxford: Oxford University Press.

Groves, R. M., Fowler, J. F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: John Wiley & Sons.

Groves, R. M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111–121.

Heerwegh, D. & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79–12.

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2007). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, . . . R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 499–528). New York: John Wiley & Sons.

Holden, R. R., Wood, L. L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology*, *81*(1), 160–169.

Katsanidou, A. & Otjes, S. (2017). Beyond Kriesiland: EU integration as a super issue after the Eurocrisis. *European Journal of Political Research*, *56*(2), 301–319.

Keeter, S. (2006). The impact of cell phone non-coverage bias on polling in the 2004 presidential election. *Public Opinion Quarterly*, *70*(1), 88–98.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

Kriesi, H., Grande, E., Lachat, R., Dolezal, M., Bornschier, S., & Frey, T. (2006). Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research*, *45*(6), 921–956.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *70*, 29–44.

Malhotra, N. & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 293–296.

Marschall, S. (2014). Profiling user. In D. Garzia & S. Marschall (Eds.), *Matching voters with parties and candidates: Voting advice applications in comparative perspective* (pp. 93–104). Colchester: ECPR Press.

Marzuca, A., Serdült, U., & Welp, Y. (2011). Questão Pública: First voting advice application in Latin America. In E. Tambouris, A. Macintosh, & H. de Bruijn (Eds.), *Electronic participation. epart 2011. lecture notes in computer science* (pp. 216–227). Berlin: Springer.

Mavletova, A. & Couper, M. (2013). Sensitive topics in PC web and mobile web surveys: Is there a difference? *Survey Research Methods*, *7*(3), 191–205.

Meijer, R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314.

Mendez, F. (2017). Modeling proximity and directional decisional logic: What can we learn from applying statistical learning techniques to VAA-generated data. *Journal of Elections, Public Opinion and Parties*, *27*(1), 31–55.

Mendez, F. & Wheatley, J. (2014). Using VAA-generated data for mapping partisan supporters in the ideological space. In D. Garzia & S. Marschall (Eds.), *Matching voters with parties and candidates: Voting advice applications in comparative perspective* (pp. 161–173). Colchester: ECPR Press.

Ming, K. & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, *56*(1), 118–124.

Mulligan, K., Krosnick, J., Smith, W., Green, M., & Bizer, G. (2001). Nondifferentiation on attitude rating scales: A test of survey satisficing theory. Unpublished Paper.

Oosterveld, P. & Willems, P. (2003). *Two modalities, one answer? Combining internet and CATI survey effectively in market research*. Amsterdam: ESOMAR.

Paulhus, D. L. (1991). Measurement and control of response bias. In *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego: Academic Press.

Pianzola, J. & Ladner, A. (2011). *Tackling self-Selection into treatment and self-Selection into the sample biases in VAA Research*. Paper presented at the 6th ECPR General Conference; 24-27 August, Reykjavik, Iceland.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. (2003). Common methods biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Popp, R., Horvath, L., Banducci, S., Coan, T., & Krouwel, A. (2016). *What voting advice applications can teach us about voters and elections*. Paper presented at the 10th ECPR General Conference; 7-10 September, Prague, Czech Republic.

Rosenbaum, P. L. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 1–55.

Sudman, S., Bradburn, N. N., & Schwarz, N. (1996). *Thinking about answers: The Application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Sullman, M. J. M. & Taylor, J. E. (2010). Social desirability and self-reported driving behaviours: Should we be worried? *Transportation Research Part F: Traffic Psychology and Behaviour*, *13*(3), 215–221.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

Van der Ark, L. A. (2007). Mokken scale snalysis in R. *Journal of Statistical Software*, *20*(11), 1–19.

Van der Ark, L. A. (2012). New developments in mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27.

Vannette, D. L. & Krosnick, J. A. (2013). A comparison of mindlessness and survey satisficing. In A. l. C. T. Ngnoumen & E. J. Langers (Eds.), *The wiley-blackwell handbook of mindfulness* (pp. 312–327). West Sussex: John Wiley & Sons.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

Walgrave, S., Nuytemans, M., & Pepermans, K. (2009). Voting aid applications and the effect of statement selection. *West European Politics*, *32*(6), 1161–1180.

Walgrave, S., Van Aelst, P., & Nuytemans, M. (2008). Do the vote test': The electoral effects of a popular vote advice application at the 2004 Belgian elections. *Acta Politica*, *43*(1), 50–70.

Wheatley, J., Carman, C., Mendez, F., & Mitchell, J. (2014). The dimensionality of the Scottish political space: Results from an experiment on the 2011 Holyrood elections. *Party Politics*, *20*(6), 864–878.

Ye, C., Fulton, J., & Tourangeau, R. (2011). More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, *75*(2), 349–365.

Appendix
Tables

(Appendix tables on next pages)

Table A1
The 19 common questions of the VAA questionnaire and the offline survey

q#	Question
q1	Fiscal deficits should be mainly covered by the additional taxation of wealth
q2	It is necessary to extend the time limit for unemployment benefit even if it burdens the deficit
q3	The quasi-governmental agencies should be privatized regardless of whether they are profitable
q4	The institution of ATA[1]should be repealed
q5	Labor rights of public employees should be equated with the rights of private sector employees
q6	Allowances for political refugees should be cut
q7	The increase in unemployment is mainly due to the uncontrolled influx of foreign (EU and non-EU) workers
q8	Holders of bonds should be compensated for the full value of these
q9	Access by T/Cs[2]in free medical care should be limited except for residents of the free areas
q10	A bi-zonal bi-communal federation for Cyprus will be sustainable
q11	The new President of the Republic should be bound by previous agreements in negotiations for Cyprus
q12	Cyprus should raise the issue of abolition of the British bases prior to a comprehensive settlement of the Cyprus problem
q13	The closing of the checkpoints (between the two communities) should be used as leverage to solve the Cyprus problem
q14	Military service should be reduced to 18 months
q15	Crime will be tackled effectively if the number of non-EU migrants is limited
q16	The role of the church should be focused on spiritual matters rather than matters of general policy of the State
q17	Possession of soft drugs (i.e. marijuana) for personal use should be decriminalized
q18	Same-sex couples should be institutionalized in the form of a civil partnership
q19	The creation of casinos should be allowed

[1] ATA: Automated Wage Indexation
[2] T/Cs: Turkish Cypriots

Table A2
Differences in matched-for variables before and after data pre-processing

		Original data		Matched dataset	
	Variable	test-statistic	p	test-statistic	p
Continuous	Age	*t*(849.9) = 13.88	< 0.001	*paired t*(772) = 0.66	0.50
	SelfPlace LR	*t*(899.9) = −0.96	0.335	*paired t*(772) = 0.38	0.70
	SelfPlace ProgCons	*t*(903.7) = 3.17	0.002	*paired t*(772) = -0.56	0.58
Categorical	Sex	$x^2(1) = 66.6$	< 0.001		
	Education	$x^2(1) = 742.5$	< 0.001	All pairs with identical values	
	partyId	$x^2(6) = 33.4$	< 0.001		
	prevVote	$x^2(6) = 43.3$	< 0.001		
	voteIntention	$x^2(3) = 36.3$	< 0.001		

Note. Matching technique: Propensity score matching, variant ratio 1:3, without replacement, with a caliper of 0.25 times the standard deviation. Of propensity scores and removing respondents outside of common support.

Table A3
Mokken Scale Analysis on the 19 policy items included in the offline and online questionnaires

q#	Question	Scalability Coefficient
q13r[1]	The closing of the checkpoints should be used as leverage to solve the Cyprus problem	0.34
q14r[1]	Military service should be reduced to 18 months	0.30
q16	The role of the church should be focused on spiritual matters rather than matters of general policy of the State	0.39
q17	Possession of soft drugs (e.g. marijuana) for personal use should be decriminalized	0.36
q18	Same-sex couples should be institutionalized in the form of a civil partnership	0.36
q19	The creation of casinos should be allowed	0.39
	Overall scalability coefficient	0.36

[1] The "r" suffix indicates that the original responses needed to be reversed for the item to fit with the scale.

Table A4
Models testing agreement and association between offline and VAA respondents

Model	G^2	df	p	ΔG^2	Δdf	p
Independence	109.367	16	< 0.001			
+Agreement	46.901	11	< 0.001	62.466	5	< 0.001
+Agreement + Linear-By-Linear Association	19.264	10	0.037	90.103	6	< 0.001

Model	Parameter (off-on)	Estimate	Z-statistic	p	Odds Ratio
+Agreement	Rater agreement (CD)	0.290	4.06	< 0.001	1.34
	Rater agreement (D)	- 0.200	−3.50	0.003	0.82
	Rater agreement (N)	0.110	1.38	0.167	1.20
	Rater agreement (A)	0.080	1.64	0.102	1.09
	Rater agreement (CA)	0.340	5.20	< 0.001	1.41
+Agreement + Linear-By-Linear Association	Linear-By-Linear Association	0.028	3.79	< 0.001	NA