To Clean or not to Clean? Improving the Quality of VAA Data[1]

Ioannis Andreadis
Department of Political Sciences, Aristotle University Thessaloniki
Greece

Introduction

A large volume of data is produced by the use of Voting Advice Applications. In order to get the voting advice, VAA users have to express their degree of agreement on a series of political issues. They also give answers to questions about their demographic, social and political characteristics. VAAs are often used by a large part of the population. The large volume of produced data provokes researchers to exploit it. Researchers use this data in multiple ways: to build the profile of VAA users, to evaluate the application, etc. But, what is the quality of these datasets?

The components that affect the quality of VAA data are very similar to the components that affect the quality of web survey data. According to Dillman(2007) [2] the quality of a survey is affected by the overall survey error which consists of four components: coverage error, sampling error, nonresponse error, and measurement error. Coverage error is the error that occurs when some of the elements of the population cannot be included in the sample. Sampling error is the error (inaccuracy) in estimating a quantity based on the sample instead of the whole population. Nonresponse error occurs when some people in the survey sample do not respond to the questionnaire and there is evidence that they differ significantly from those who respond. Measurement error occurs when answers to survey questions are inaccurate or wrong.

The most significant errors associated with web surveys are coverage errors and measurement errors. Coverage errors occur in web surveys because a part of the population does not have Internet access. The probability of measurement error can be larger in all self-administered surveys due to the lack of interaction with a human (the interviewer) who could clarify the meaning of a question in case the respondent needs it. Finally, as Heerwegh and Loosveldt[3] argue, web surveys respondents might have a number of programs running concurrent with the web survey and they might devote their energy to multiple activities (multitasking). This multitasking could increase the probability of measurement error and if the web survey is long it could also lead to drop outs (when another activity requires the entire attention of the user).

Of course VAAs are different from web surveys with regard to two characteristics: access rules and respondent motivation. Access to a web survey is usually prohibited to the general public. In this case, only people who have been sent an invitation can participate to the web survey by entering their unique pin code or token. On the other

---

[1] Please send your comments to: john@polsci.auth.gr
[2] Dillman DA (2007). *Mail and internet surveys: the tailored design method* (2nd edition). New York, NY: John Wiley and Sons, Inc.
[3] Heerwegh, D.; Loosveldt,G. (2008) Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality, *Public Opin.Q.*, 72, 5, 836-846

hand, VAAs are open to anyone with internet access. In addition, a user can participate to a VAA as many times as he/she likes. Another difference between VAAs and Web surveys is the output. Usually, when users complete a web survey, the only output they face is a "Thank you for your participation" screen. In order to get some useful output, web survey participants have to wait for the publication of the analysis of the collected data. People participate to surveys (web or any other mode) by a sense of social responsibility, a self- perception of being helpful, but also express their opinion and affect policy decision-making. On the other hand, people use VAAs because their responses are evaluated immediately and the users get a personalised output, i.e. a personal "voting advice". This VAA feature motivates some users to complete the VAA questionnaire multiple times for various reasons. Some users give their true positions the first time they use a VAA, but then they become curious to find out the answers to various "what if" questions. For instance, they wonder what the output would be if they had answered "Strongly Disagree" (or "Strongly Agree") to all sentences. Other users, the first time they complete a VAA questionnaire, use it as a game; they only want to see the available outcomes, not the outcome for their own positions. As a result, they do not pay too much attention to the questions, or they even give totally random responses without reading the questions. These users want to explore the tool and test how it reacts to their actions; their answers do not correspond to their true positions.

From the previous paragraphs it is obvious that the quality of VAA data suffers in two areas: i) lack of representativeness due to limited coverage, and ii) measurement error due to nonsense answers. With regard to the former problem, the situation will improve as Internet use spreads to groups with lower access rates. The latter problem will not improve and we need to deal with it. The aim of this paper is to address this problem by attempting to answer the following questions: How can we discover these nonsense answers? How serious is the problem, i.e. what is the percentage of nonsense answers? If we analyse the data without removing the invalid cases, what will be the impact on findings and conclusions? The paper concludes with implications and suggestions for VAA designers and researchers working with VAA data.

Response Time

Measuring response time[4] is common in the survey literature. In fact, it is so common that a number of different measuring approaches has been proposed. For instance, there are two types of proposed timers depending on the mode of the survey: active timers and latent timers. Active timers are used when an interviewer is present; the interviewer begins time counting after reading aloud the last word of the question and stops time counting when the respondent answers. This approach assumes that the respondent starts the response process only after hearing the last word of the question.

[4] Time spent to answer a question belongs to a special type of data called “Paradata”. These data do not describe the respondent’s answers but the process of answering the questionnaire. See Stern, M. J. (2008). The Use of Client-side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys *Field Methods* November 20: 377-398 Also, Heerwegh, D. (2003). Explaining response latencies and changing answers using client side paradata from a web survey. *Social Science Computer Review*, 21(3), pp.360-373 and Heerwegh, D. (2004). Uses of Client Side Paradata in Web Surveys. Paper presented at the *International symposium in honour of Paul Lazarsfeld* (Brussels, Belgium June 4-5 2004)

Latent timers are preferred when the questions are visually presented to the respondent (e.g. web surveys). This approach assumes that the respondent starts the response process from the first moment the question is presented to him/her. Another decision to be made concerns the location of time counting. Should counting be done on the server side or the client side? Counting on the server side is feasible by recording a timestamp when a user visits a web page. This means that in order to count time spent on each question, we need to keep each question on a separate web page. Of course this is not a problem for VAAs because usually VAAs present each question on a different page. But there is another problem with server-side time counting. Server-side response time is the result of the sum of the clear response time plus the time between the moment the user submits the answer and the moment the answer is recorded on the server. The second component depends on the type and bandwidth of the user's internet connection, but also on unpredicted, temporary delays due to network load, etc. On the other hand, client-side time counting is done at the level of the respondent's (or client's) computer itself. Consequently, client-side time counting should be preferred because it is more accurate and it does not include any noise.

HelpMeVote 2012 was coded with *jQuery Mobile* and it was built as an *AJAX* application; all 30 pages are downloaded from the beginning to the users' browser. This means that there is no lag time between answering one question and viewing the next question. The time between clicks can be counted accurately. The response times are recorded in hidden input fields.[5] Communication with the server is done in the end, when all questions have been answered and the user has clicked the "Submit" button. When the respondent submits the web page, the content of the hidden fields are stored on the server.

At this point there is another important note that I would like to mention about HelpMeVote 2012. HelpMeVote 2012 allows users to submit only one questionnaire during a session, i.e. after submitting the user cannot go back, change one or more answers and submit again (the system keeps only the initial set of answers). The only way a user can repeat the test is to start from the beginning. This way helpmevote accepts only compete sets of answers and the dataset is already cleaner from the beginning in comparison with the helpmevote application used in 2010 which allowed users to have different sets of answers within the same session.

Tourangeau et al. (2000) [6] divide the survey response process into four major tasks:

1. comprehension of the question,
2. retrieval of relevant information,
3. use of that information to render the judgment, and
4. the selection and reporting of an answer.

The time spent on comprehension and reporting components depends on the characteristics of the questions. Time spent on comprehension depends on the length and the complexity of the question. Time spend on reporting is affected by how many and what type of response categories are offered. For instance, previous results

[5] Of course, with VAAs we can only use a latent timer (no interviewer is present and there is no way to know when the respondent has finished reading the question).

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

indicate that response times are longer when the negative, rather than the positive, end of the scale is presented first. Response time is longer for formats that are difficult for respondents to process.[7] For VAA items, reporting procedure is the same for all questions; thus, it is reasonable to expect a fixed time spent on reporting and it should be short (clicking on a radio button is one of the simplest and fastest ways to report the answer).

Retrieval and judgment may be determined by respondent characteristics[8] (e.g. age, education level, etc) but since I argue that some users give nonsense answers, (and I want to study these users), I suppose that they would also give nonsense answers to the questions regarding their demographic characteristics. Thus, I will not use respondent characteristics in my analysis.

Time dedicated to judgement depends on the existence or not of an attitude on the topic. People with an existing attitude are expected to answer faster than people who make up an attitude on the spot[9]. Even between people who have an attitude, time will depend on the attitude strength. People with unstable attitudes need more time to finalise their answer than people with a stable attitude who do not need to spend more time than the time to retrieve their already processed attitude from their memory. Previous research has revealed a positive relationship between response latency and unstable attitudes (measured as changes of the answer after being exposed to the counterargument).[10] Finally, it has been shown that attitudes expressed quickly are more predictive of future behaviour than attitudes expressed slowly. Bassili (1993) has provided logistic regression evidence supporting the hypothesis that response latency is a better predictor of discrepancies between voting intentions and voting behaviour than self-reported certainty about their vote intention.[11]

Much of the time spent on task 1 involves reading and interpreting the question. One component of this time is related to the complexity of the question. Previous research has shown that badly expressed questions (e.g. double-barrelled questions or questions containing a superfluous negative) take longer to answer than nearly identical questions without these problems[12]. Of course, a well-designed VAA should not include badly expressed sentences; a pilot study should be adequate to spot these questions. Badly expressed sentences should be corrected or replaced.

If all questions included in a VAA have similar complexity, then the most significant factor that affects time spent on Task 1 is the length of the question. These two quantities (length and time) are proportional and their ratio defines the reading speed. VAA users need time to read the sentence using a reading speed suitable for the

---

[7] Christian,Leah Melani; Parsons,Nicholas L.; Dillman,Don A. (2009). Designing Scalar Questions for Web Surveys S*ociological Methods & Research*, 37, 3, 393-425

[8] Yan, T.; Tourangeau, R.(2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22, 1, 51-68

[9] It is also possible that someone who holds no attitude at all is less involved with the issue, he/she does not care about it and gives a quick, unconsidered answer.

[10] Bassili, J.N.; Fletcher, J.F., (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opin.Q.*, 55, 3, 331-346.

[11] Bassili, J.N. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. *Public Opin.Q.*, 57, 1, 54-61.

[12] Bassili, J. N., and B. S. Scott. (1996). Response latency and question problems. *Public Opinion Quarterly* 60 (3): 390–99

comprehension of the thoughts in the sentence. The unit used to measure reading speed in the related literature is "words per minute" (wpm). This unit may be suitable to measure reading speed on large texts, but it is inappropriate unit to measure reading speed on texts of limited size, like the sentences used in a VAA. The number of words in HelpMeVote 2012 sentences ranges from 7 to 24 words. According to the analysis of the Hellenic National Corpus the Average Word Length is 5.33 and the distribution is skewed to the right[13]. This means that it is possible to have a sentence with a limited number of words that is longer than another sentence with more words. For instance, one HelpMeVote 2012 sentence consists of 13 words, 62 characters (74 including spaces) and average word length 4.77. Another sentence consists of 8 words, 67 characters (74 including spaces) and average word length 8.38. The average user has spent 6.24 seconds on the former (13-words) sentence and 7.22 seconds on the latter (8-words) sentence. To avoid similar problems, I have decided to use the number of characters instead of using the number of words. The shorter sentence of HelpMeVote 2012 consists of 44 characters and the longer sentence is 170 characters long.

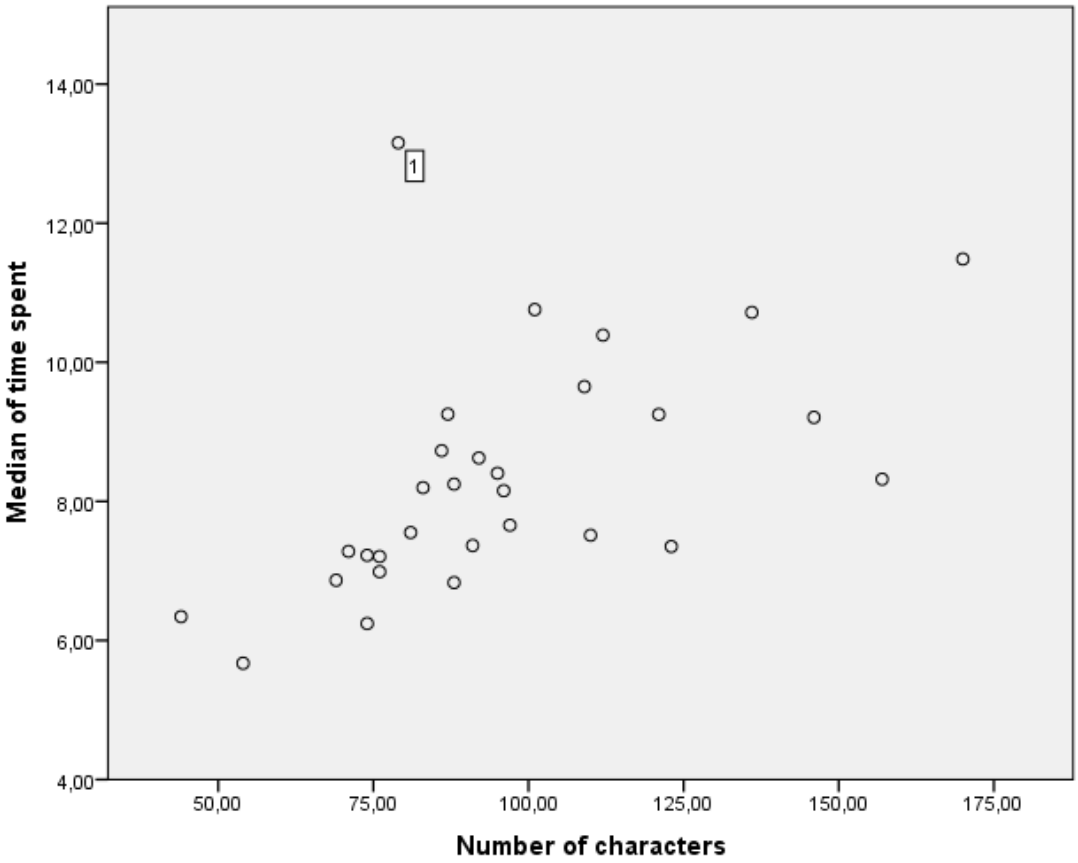

Diagram 1 Scatterplot number of characters – Time spent

For the time spent on each question I need a measure of central tendency, a value that summarizes the time spent (in seconds) by all users. The average value is not the most suitable measure because there are cases with extremely large values (probably by users who have been interrupted by something e.g. phone call, email, chat, etc).

[13] Basic Quantitative Characteristics of the Modern Greek Language Using the Hellenic National Corpus George Mikros, Nick Hatzigeorgiu, & George Carayannis. Journal of Quantitative Linguistics, 2005, Vol. 12, No. 2-3, pp. 167 – 184

Response times are generally right skewed and the average value is sensitive to outliers. Therefore, I use the median value which is robust to extreme values.[14]

Diagram 1 displays the scatterplot of the median time spent on each sentence with the sentence length (counted as number of characters). It becomes obvious that the first case is an outlier, because the time spent (in seconds) on the first question is longer than the time spent on other questions with similar number of characters. This is an expected finding because when users face the first question they need to spend additional time to read the text on the displayed buttons and to understand that they can express their position by clicking on one of these buttons. After answering the first sentence, they are familiar with the procedure and the available options and they can express their position in less time.

After excluding the outlier I apply a linear regression model on these two variables. From Table 1 it is observed that the fitted model is y=4.747+0.036x. This means that for every additional 100 characters in the sentence the time spent on a question increases by 3.6 seconds. According to the fitted model some of the time spent on each sentence depends on the length of the sentence, but there is another amount of time that is constant for all sentences. This constant time is spent by the users to think about the sentence, determine their position and express it by clicking the corresponding button. According to the fitted model, this part of the median time spent is estimated at about 4.7 seconds.

Table 1 Linear regression of time spentd on number of characters including spaces

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,747	,676		7,021	,000
	Number of characters	,036	,007	,715	5,307	,000

a. Dependent Variable: Median of time spent

Table 2 Linear regression of time spent on number of characters without spaces

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,761	,667		7,133	,000
	Number of characters (no spaces)	,041	,008	,718	5,359	,000

a. Dependent Variable: Median of time spent

Table 1 displays the fitted model when the measure used for the length of the sentence is the total number of characters (including spaces). Table 2 shows the same model when the number of characters (without spaces) is used as the independent variable. From a comparison of the tables, it is obvious that there are no significant differences between these two models. It does not matter which variable is used as independent, since both models convey the same information.

[14] van Zandt, T. (2002). Analysis of response time distributions. In H. Pashler, & J. Wixted (Eds.), *Stevens' handbook of experimental psychology* New York: John Wiley & Sons

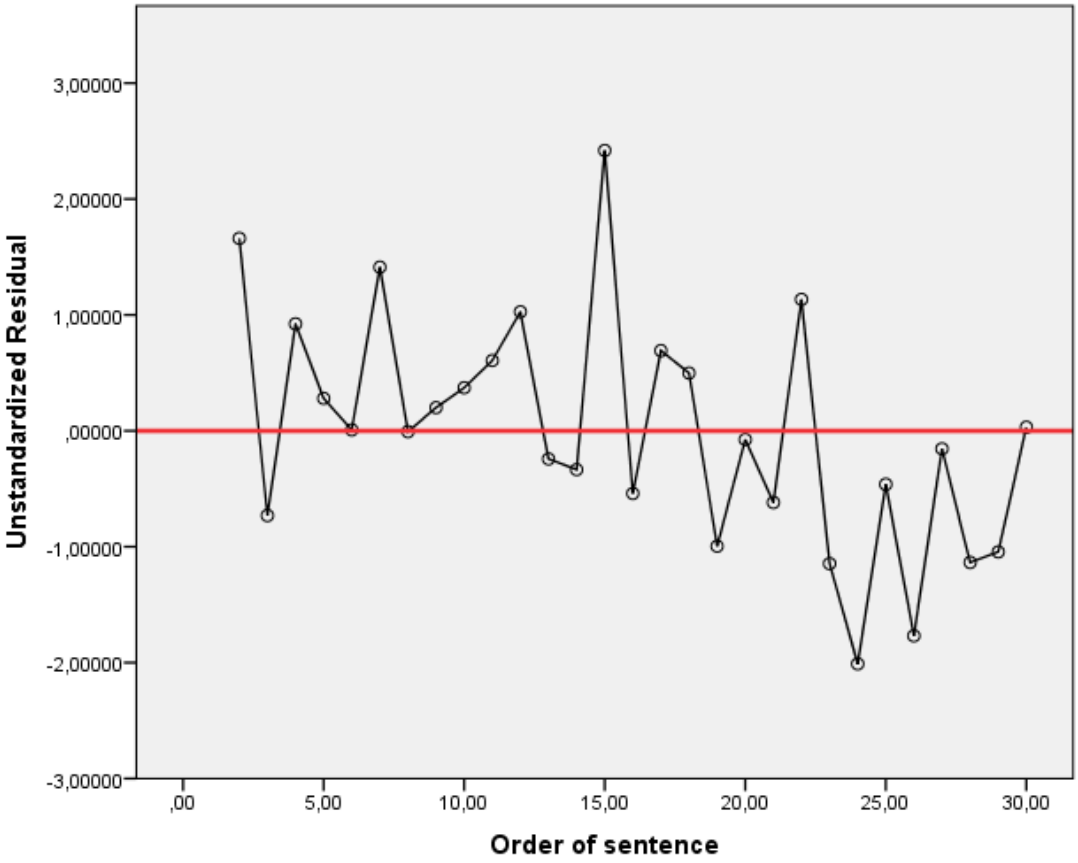

Diagram 2. Line plot of residuals and order of sentence

With regard to time spent on each sentence, one important question that should be answered is the following: "Do users get tired/bored near the end of the test and dedicate less time (pay less attention) to the last sentences?" The analysis of the residuals can shed some light on this issue. The residuals of the fitted model are presented in Diagram 2. The X-Axis is formed by the order of the sentence. A positive residual means that the time spent for the corresponding sentence was more than the time expected according to the model and a negative residual means that the sentence was answered in less time than expected. It seems that until question 22 there are both positive and negative residuals which appear in random order. This is an expected pattern because the time spent on a sentence does not depend only on the length of the sentence.[15] On the other hand, starting from sentence 23 there is a series of negative residuals. This series could be a sign of tiredness but in order to prove this we should run an experiment reordering the questions and measuring if the time spent on the same question depends on the order it appears. This finding is in agreement with similar findings from the analysis of web survey response times. Yan and Tourangeau (2008) classified each question according to the quarter of the questionnaire it was located (1st, 2nd, 3rd, and 4th quarter). They have found evidence that respondents tend to answer more quickly as they get closer to the end of the questionnaire.

[15] As I have already mentioned in previous paragraphs, complexity of the sentence is another significant factor. Studying the complexity of the sentences is out of the scope of this paper. Since VAAs designers try hard to include simple sentences in their VAA, I suppose that all sentences have similar (and limited) complexity.

In the following paragraphs I will try to classify response times in order to find a way to reveal the cases where the response time was so small indicating that the answer is not valid. Fry (1963) classifies readers as good (350 wpm), fair (250 wpm) and slow (150 wpm).[16] Carver (1992) provides a table connecting reading speed rates and types of reading and associates reading rate of 300 wpm with a reading process named rauding which is suitable for comprehension of a sentence, reading rate of 450 wpm with skimming, i.e. a type of reading that is not suitable to fully comprehend the ideas presented in the text and a reading rate of 600 wpm with scanning which is suitable for finding target words. [17]

For English texts the average word length is 4.5 letters[18]. In order to compare the speed of HelpMeVote users with previous findings from studies on English language, I am using a standardized length of a word of five characters. Following the second fitted model that uses as independent variable the number of characters without spaces I estimate that each character requires 0.041 seconds, i.e. a word of five characters requires 0.205 seconds. Converted to the usual units (i.e. wpm) this figure gives 292.7 words per minute. This value positions the median speed of HelpMeVote users near the value 300 which, according to Carver, is the normal speed for rauding, and according to Fry is located between fair (250) and good (350). Thus, the median HelpMeVote user has dedicated enough time to read the sentences using a reading speed that is suitable for comprehension and then allocated enough time (4,76 seconds) to determine and express his/her position.
Using Carver's table, I try to estimate a threshold that will separate answers given after reading and comprehending the sentence from answers given in so little time that there is strong evidence that the user was not able to read and comprehend the sentence, i.e. the answer has no value and it should be discarded. I argue that scanning reading speed is too fast for a VAA user to comprehend the sentence. Thus, I use as a threshold the midway between skimming and scanning i.e. 575 wpm. Converted to characters per second (with 5 characters per word) it gives the value of 43.75 cps. Then, I divide the number of characters (without spaces) in each sentence with this value, I get the minimum time (in seconds) that is necessary to read the sentence. Of course users need some time for all other tasks (2-4) reported by Tourangeau et al. (2000) , i.e retrieval of relevant information, use of that information to render the judgment and the selection and reporting of an answer. The fitted model indicates that the median time spent on this procedure is 4.76 seconds.

Bassili and Fletcher, (1991) using an active timer,[19] have found that on average, simple attitude questions take between 1.4 and 2 seconds, and more complex attitude questions take between 2 and 2.6 seconds. At this point, I do not have access to any other findings of previous research that would help me decide what the minimum time is for a user to determine and express his/her level of agreement with a sentence on a

[16] Fry, E.B. (1963). Teaching faster reading: a manual. Cambridge: Cambridge University Press.
[17] Carver, R.P. (1992) Reading rate: Theory, research, and practical implications, Journal of Reading, 1992, 36, 2, 84-95
[18] n-Grams and their implication to natural language understanding E.J. Yannakoudakis, I. Tsomokos, P.J. Hutton, Pattern Recognit, 1990, 23, 5, 509-528
[19] In their experiment, time counting starts when interviewer presses the spacebar after he has read the last word of the question. Time counting stops with a voice-key (the first noise that comes from the respondent's side triggers the computer to read the clock).

five-point Likert scale. I choose arbitrarily to divide the median value by three, i.e. I argue that someone can be three times faster that the median user and still give a useful answer, but beyond this threshold the answer is given randomly. Dividing the median value by three gives the number 1.587, which is similar to the minimum time reported by Bassili and Fletcher for simple attitude questions (1.4 seconds)[20].

Table 3. Thresholds used to classify answers

Sentence	Number of characters (without spaces)	Threshold1	Threshold2
1	68	4.55	3.14
2	96	5.45	3.78
3	127	6.44	4.49
4	73	4.71	3.25
5	83	5.03	3.48
6	62	4.36	3.00
7	72	4.68	3.23
8	83	5.03	3.48
9	105	5.74	3.98
10	78	4.87	3.37
11	80	4.94	3.41
12	94	5.38	3.73
13	67	4.52	3.11
14	61	4.33	2.98
15	87	5.16	3.57
16	84	5.06	3.50
17	148	7.11	4.97
18	73	4.71	3.25
19	46	3.85	2.63
20	69	4.58	3.16
21	76	4.81	3.32
22	120	6.22	4.33
23	96	5.45	3.78
24	134	6.66	4.65
25	65	4.46	3.07
26	107	5.80	4.03
27	67	4.52	3.11
28	62	4.36	3.00
29	73	4.71	3.25
30	38	3.59	2.45

Consequently, the formula I used to estimate the threshold between valid and non-valid answers is: 1.587+[Characters in sentence without spaces]/43.75. In order to test if the threshold should be different I use another threshold that will help me separate valid answers given by fast users from valid answers given by slow users. Thus, I use as a threshold the midway between rauding and skimming i.e. 375 wpm. Converted to

[20] The readers should keep in mind the different procedures. Bassili and Fletcher use the voice-key that records the time there is some voice from the respondent's side. I record the time the user clicks on one of the available buttons that correspond to answer options. This additional step requires some extra time.

characters per second (with 5 characters per word) it gives the value of 31.25 cps. I argue that fast users will require half of the time the median user needs to decide. Thus, I divide the median value of decision time by two. The formula I have used to estimate this additional threshold between valid answers given by fast users and valid answers given by slow users is: 2.38+[Characters in sentence without spaces]/31.25

Table 3 shows the thresholds used to classify answers. If the time spent on a sentence was more than the time (in seconds) indicated in the column with the label "Threshold 1" I argue that the user dedicated enough time to read the sentence, understand the ideas, and express his/her position. If the time spent was between the two thresholds, I argue that the user has read the sentence with a reading speed around the level of skimming and he has dedicated limited time to determine and express his/her position. The users in this category have acted fast, but within acceptable limits and their answer are probably valid, but this category probably includes both valid and invalid answers. Finally, if a user has spent on a sentence less than the time indicated in the column labeled "Threshold 2", I argue that the dedicated time was not enough for a valid answer; the answer was given either by randomly clicking on any of the available buttons or the user has clicked on a fixed button for all sentences, e.g. the user is playing with the application and he/she wants to see the output it provides when all answers are (supposedly) "Neither agree nor disagree".

As an example of the output of this classification I use the second sentence. As Table 4 shows about 5% of the answers have been given in less than 3.78 seconds, i.e. the users were scanning and the dedicated time was not enough to give a valid answer. The second category (7%) consists of answers that were given in less than 5.45 seconds and more than 3.78 seconds. Users in this category were fast, but it is possible that their answers are valid. Most of the users (about 87%) have spent more than 5.45 seconds. Finally, there are some users (1%) for whom the time spent on sentence 2 was not recorded for various reasons. The most common reason was that some users have tried to skip some questions, i.e. by modifying the URL of the address bar of their internet browser.

Table 4 Distribution of time spent on Sentence 2

	Frequency	Percent
Scanning	24394	5.1
Skimming	33436	7.0
Normal	414516	86.9
Unable to count	4789	1.0
Total	477135	100.0

Pattern of answers

Another way to clean VAA data is to delete records submitted by users who (for various reasons) have given a constant answer to every (or almost every) question (provided that there are questions with opposite directions).

Table 5 Frequencies of fixed answers (rigid: 30 identical answers)

	Frequency	Percent
Strongly Disagree	715	11,6
Disagree	65	1,1
Neither … nor	1486	24,1
Agree	61	1,0
Strongly Agree	300	4,9
No answer	3543	57,4
Total	6170	100,0

Table 5 indicates that there are 6170 records that have the same value in all 30 fields, i.e. the user clicked on the same button for all 30 sentences. The most used constant answer is the "No answer" (57.4% of the constant answer records). The next most used constant answer is the median "neither agree nor disagree" point (24.1%). "Strongly disagree" (11.6%) and "Strongly agree" (4.9%) are next. The preference for "Strongly disagree" can be attributed to the user interface of helpmevote: answering buttons are displayed vertically and the order of appearance is from "Strongly disagree" to "Strongly agree" and last comes the "No answer" button. The other two buttons have been used as constant answers by a very limited number of users.

Of course, it is possible that a user had the intention to click on the same answering button for each question but while he/she was trying to do this at a high speed, he/she accidentally clicked on a different button. In Diagram 3[21] X-axis is formed by categories of records[22] defined by a variable that counts the number of "Strongly Disagree" answers in a record. The Y-axis is formed by the frequencies of these categories. For instance, there are 107 users who have used the answer SD for the 27 of the 30 sentences. We can observe that until the category with 26 SD answers there is a negative correlation between the number of SD answers in a record and the frequency of the record. The minimum frequency (circa 70) is observed for the records with 23-26 SD answers. After point 26 the correlation between the number of SD answers in a record and the frequency of the record is positive. The frequency increases to 107 for the category with 27 SD answers, 166 for the category with 28 SD answers, 278 for the category with 29 SD answers and 715 for the category with 30 SD answers. I argue that records which follow the expected declining trend (i.e. records with a maximum of 26 SD answers) are valid records. For records with more than 26 SD answers the frequency is mounting and goes up to a local maximum for the category with 30 SD answers. This increasing trend is probably an evidence that the records with 27, 28 and 29 SD answers are from users who intended to give the same fixed answer (in this case SD to all questions) but they accidentally clicked on another button in 1-3 cases. Thus, I consider invalid all cases with 27-30 SD, or D, or A, or SA answer.

[21] I use the part of the diagram that includes only the records with more than 15 SD answers because if I had included all cases, the U-shape would have appeared as a straight line, as a result of the first records having very large frequencies (for instance there are 60920 records with one SD answer). The frequency increases until the most frequent category and then it monotonically decreases for all other cases until the minimum is reached.

[22] A record is the set of the answers given by a respondent to all 30 questions.

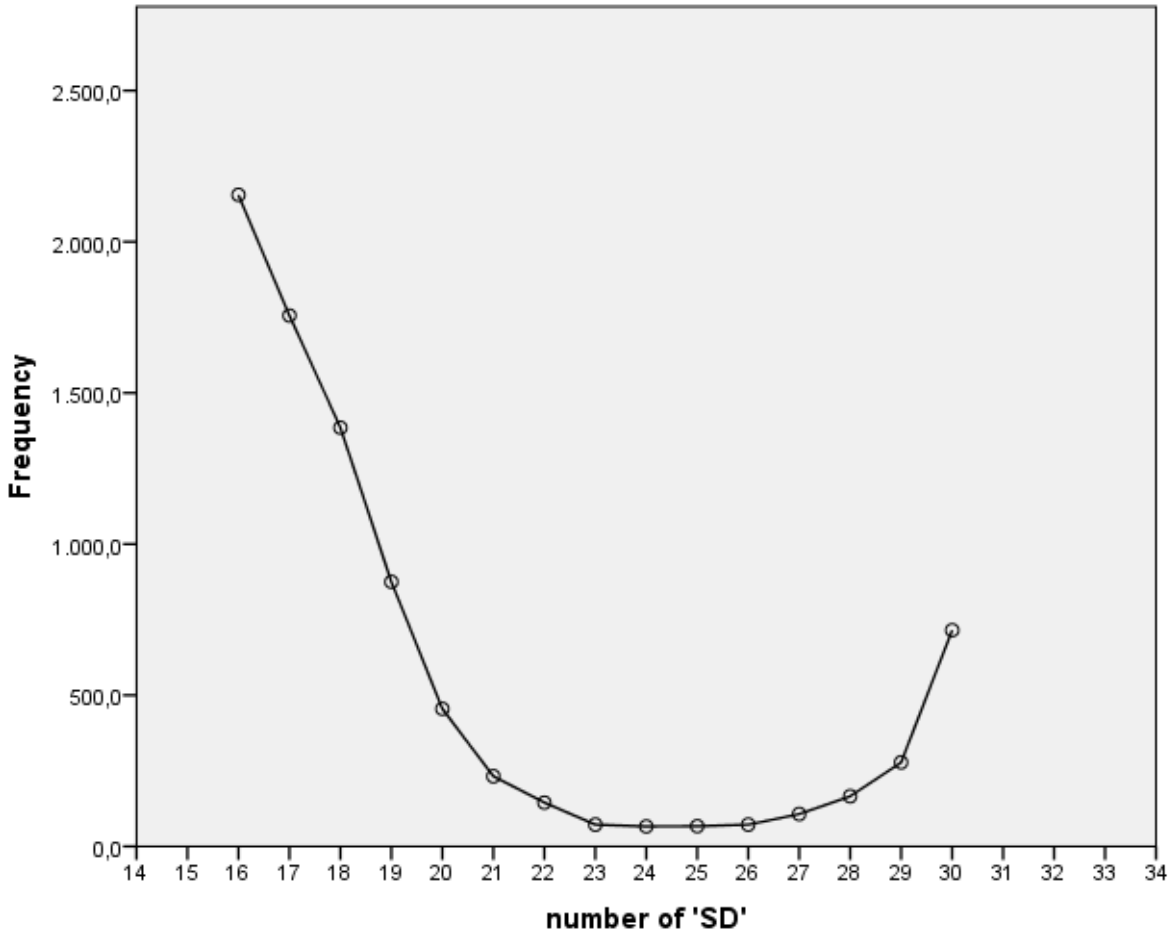

Diagram 3. Frequency of "Strongly disagree" (filtered by frequency>15)

A similar U-shaped curve is observed for the answers "Neither agree nor disagree" (see Diagram 4). The difference with the previous diagram is that the minimum is observed near the category of the records with 20 NN answers. Following a similar logic as I did for the previous diagram, I consider a record as invalid if more than 20 out of 30 questions have been answered with the midpoint "Neither agree nor disagree". Finally, I argue that a record should be considered as invalid if it has more than half of the questions unanswered. Following the aforementioned rules the frequency of records rejected due to the pattern of the answers is shown in Table 6.

Table 6 Records rejected due to the pattern of the answers (flexible)

	Frequency	Percent
Valid (not rejected)	466439	97,8
Strongly Disagree >26	1266	,3
Disagree >26	107	,0
Neither … nor >20	2439	,5
Agree >26	120	,0
Strongly Agree >26	506	,1
No answer >15	6258	1,3
Total	477135	100,0

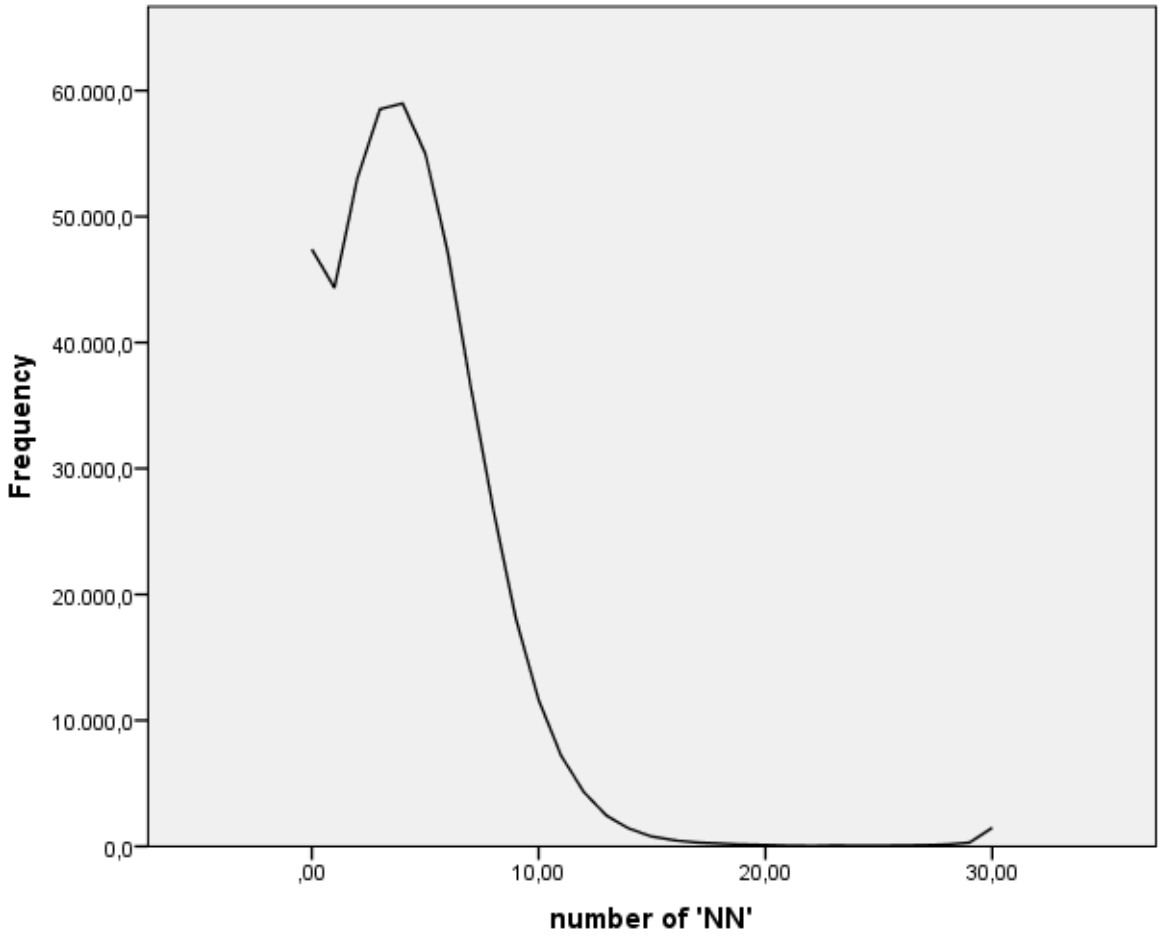

Diagram 4. Frequency of "Neither agree no disagree"

Relation of the two cleaning methods

In this section I will try to provide some answers to the following questions: Is it OK if the cleaning procedure depends on the pattern of the answers only? Are "invalid due to pattern" and "invalid due to time" related to each other? What are the characteristics of this relation?

Table 7 Classification of cases according to time spent on Sentence 2 and cases rejected due to pattern of answers.

timecat2 * Pattern rejected Crosstabulation

			Pattern rejected							
			Valid (not rejected)	Strongly Disagree >26	Disagree >26	Neither ... nor >20	Agree >26	Strongly Agree >26	No answer >15	Total
timecat2	Scanning	Count	18790	1103	74	1833	79	424	2091	24394
		% within timecat2	77,0%	4,5%	0,3%	7,5%	0,3%	1,7%	8,6%	100,0%
		% within Pattern rejected	4,0%	87,1%	69,2%	75,2%	65,8%	83,8%	33,4%	5,1%
	Skimming	Count	32952	48	8	124	9	29	266	33436
		% within timecat2	98,6%	0,1%	0,0%	0,4%	0,0%	0,1%	0,8%	100,0%
		% within Pattern rejected	7,1%	3,8%	7,5%	5,1%	7,5%	5,7%	4,3%	7,0%
	Normal	Count	411927	113	22	474	31	51	1898	414516
		% within timecat2	99,4%	0,0%	0,0%	0,1%	0,0%	0,0%	0,5%	100,0%
		% within Pattern rejected	88,3%	8,9%	20,6%	19,4%	25,8%	10,1%	30,3%	86,9%
	Unable to count	Count	2770	2	3	8	1	2	2003	4789
		% within timecat2	57,8%	0,0%	0,1%	0,2%	0,0%	0,0%	41,8%	100,0%
		% within Pattern rejected	0,6%	0,2%	2,8%	0,3%	0,8%	0,4%	32,0%	1,0%
Total		Count	466439	1266	107	2439	120	506	6258	477135
		% within timecat2	97,8%	0,3%	0,0%	0,5%	0,0%	0,1%	1,3%	100,0%
		% within Pattern rejected	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Table 7 shows that response time alone is a very good indicator to flag invalid cases. Cases rejected due to time (category scanning) include 87.1% of the cases rejected because of fixed “SD” answers, 75.2% of the cases rejected because of fixed “NN” answers and 83.8% of the cases rejected because of fixed “SA” answers. Time checking also flags more than 65% of the cases rejected because of fixed “D” or “A” answers. Finally, time checking has flagged more than 1 out of 3 of the cases rejected due to more than 15 unanswered questions. One the other hand, observing the cases considered as valid according to the time-based criteria, it occurs that 98.6% of the “skimming” and 99.4% of the “normal speed” cases are also valid according to the pattern based criteria. Finally, 41.8% of the cases that I was unable to count the time spent on sentence 2 correspond to records that have more than half of the questions unanswered (probably by users who have jumped directly to one of the following questions without passing through question 2, i.e. by modifying the URL). The relation between the cases rejected due to the pattern of answers and the classification of cases according to the time spent on sentence does not depend on the order of the sentence. For instance, see Table 8, which describes a similar (although a little stronger) relation between cases rejected by time criteria and cases rejected by pattern criteria. The only number that is associated with the order of the sentence is the number of cases that I was unable to count the time spent on the sentence. This number decreases as we move from question to question and it drops to about ½ near the end of the test.

Table 8 Classification of cases according to time spent on Sentence 29 and cases rejected due to pattern of answers.

timecat29 * Pattern rejected Crosstabulation

			Pattern rejected							
			Valid (not rejected)	Strongly Disagree >26	Disagree >26	Neither ... nor >20	Agree >26	Strongly Agree >26	No answer >15	Total
timecat29	Scanning	Count	23637	1210	92	2204	105	474	2469	30191
		% within timecat29	78,3%	4,0%	0,3%	7,3%	0,3%	1,6%	8,2%	100,0%
		% within Pattern rejected	5,1%	95,6%	86,0%	90,4%	87,5%	93,7%	39,5%	6,3%
	Skimming	Count	68129	27	4	97	5	14	316	68592
		% within timecat29	99,3%	0,0%	0,0%	0,1%	0,0%	0,0%	0,5%	100,0%
		% within Pattern rejected	14,6%	2,1%	3,7%	4,0%	4,2%	2,8%	5,0%	14,4%
	Normal	Count	373344	29	8	136	10	16	2069	375612
		% within timecat29	99,4%	0,0%	0,0%	0,0%	0,0%	0,0%	0,6%	100,0%
		% within Pattern rejected	80,0%	2,3%	7,5%	5,6%	8,3%	3,2%	33,1%	78,7%
	Unable to count	Count	1329	0	3	2	0	2	1404	2740
		% within timecat29	48,5%	0,0%	0,1%	0,1%	0,0%	0,1%	51,2%	100,0%
		% within Pattern rejected	0,3%	0,0%	2,8%	0,1%	0,0%	0,4%	22,4%	0,6%
Total		Count	466439	1266	107	2439	120	506	6258	477135
		% within timecat29	97,8%	0,3%	0,0%	0,5%	0,0%	0,1%	1,3%	100,0%
		% within Pattern rejected	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Of course there are a lot of cases which are flagged as invalid by both time and pattern criteria. This shows a strong relationship between the two criteria. Still, there are additional cases that are flagged as invalid by time criteria which are not flagged as invalid by the pattern criteria. For instance, there are 18790 answers to the second sentence of HelpMeVote 2012 which were given in less than 3.78 seconds and when these answers are checked together with the answers given to the rest 29 questions, they do not seem to follow some pattern that would make us suspicious about their validity. This means that **if a voting advice application does not log the time spent to each sentence, the collected data cannot be fully cleaned**.

What are the differences?

In this section I will try to reveal the differences between the answers given by people who have responded the questions at a scanning speed (which I consider as invalid or nonsense answers) and the answers given by people who have dedicated enough time

to give a substantial response. Of curse the distribution of answers depends on the sentence itself. Some issues are widely accepted i.e. the majority of the electorate supports them. On the other hand, there are sentences which are faced with disagreement by the largest part of the electorate.

Table 9 Distribution of answers given to Sentence 2 by response time category

		Sentence 2				
		SD	D	NN	A	SA
Scanning	Count	6628	3743	3578	4056	4222
	%	29.8%	16.8%	16.1%	18.2%	19.0%
Skimming	Count	7572	5264	2251	9497	8515
	%	22.9%	15.9%	6.8%	28.7%	25.7%
Normal	Count	34694	61284	48548	155374	109864
	%	8.5%	15.0%	11.8%	37.9%	26.8%
Unable to count	Count	200	282	200	695	585
	%	10.2%	14.4%	10.2%	35.4%	29.8%
Total	Count	49094	70573	54577	169622	123186
	%	10.5%	15.1%	11.7%	36.3%	26.4%

As Table 9 indicates, most Greek voters agree with the second sentence (together A and SA answers correspond to circa 63% of the total answers) and only 10.5% answer that they strongly disagree. But looking into each category defined by the response time we can observe significant differences. For instance, within the “scanning” group we observe that the most frequent answer is “SD” (29.8%) and all other options are selected with about the same probability (D: 16.8%, NN: 16.1%, A: 18.2%, and SA: 19.0%). This outcome could be the result of primacy effect, i.e. increased likelihood to select the first of the available items. Psychologists argue that when we read the later response alternatives, our mind is already occupied with thoughts about previous response alternatives; consequently, the attention paid to later response alternatives is insufficient (later items are less carefully considered)[23]. Psychologists also support that primacy could be a result of satisficing[24], i.e. respondents choose the first acceptable answer instead of the optimal answer. Previous research shows that response order effects (both primacy and recency) are stronger among respondents low in cognitive sophistication.[25] Order effects are present not only in the frame of surveys using the visual channel; these effects also occur when clicking behaviour is observed with regard to website or email links. It seems that visitors click on the first link more frequently than any other link (primacy effect). The click-through rate

[23] Response order effects depend on the channel used to present the response alternatives (visual presentation vs oral presentation). When oral presentation is used, respondents are able to devote more processing time to the last item because interviewers pause after reading aloud the last available item and wait respondents to give their answer. As a result, when the aural channel is used we observe recency effects instead of primacy effects.

[24] A combination of "satisfy" and "suffice", i.e. to finish a job by satisfying the minimum requirements. Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 2, p. 129

[25] Krosnick, J.A. and Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opin.Q.*, 51, 2, 201-219.

decreases for all subsequent links except the last one, where it increases significantly (recency effect)[26].

Table 10 Distribution of answers given to Sentence 18 by response time category

		Sentence 18				
		SD	D	NN	A	SA
Scanning	Count	3249	3277	3769	2465	2308
	%	21.6%	21.7%	25.0%	16.4%	15.3%
Skimming	Count	14050	16497	3580	6730	2555
	%	32.4%	38.0%	8.2%	15.5%	5.9%
Normal	Count	103902	149411	59082	75406	16233
	%	25.7%	37.0%	14.6%	18.7%	4.0%
Unable to count	Count	415	552	217	302	89
	%	26.3%	35.0%	13.8%	19.2%	5.7%
Total	Count	121616	169737	66648	84903	21185
	%	26.2%	36.6%	14.4%	18.3%	4.6%

The findings from the distribution of responses to Sentence 2 seem to support the hypothesis of a strong impact of primacy effects among the scanning group. But this hypothesis has to be double-checked by observing the distribution of responses to a sentence when the majority does not agree with it (see Table 10 with the distribution of answers to sentence 18). In the total group the sum of SD and D responses to sentence 18 is 62.8%. On the other hand, only 4.6% of the total users select the answer SA. Within the “scanning” group the answers are distributed more uniformly and SA is selected by 15.3%. It seems that among people who are answering with scanning speed, the distribution tends to look like a discrete uniform distribution with five outcomes, i.e. each of the five outcomes is equally likely to be selected (it has probability 1/5). If the hypothesis of the discrete uniform distribution is accepted, this means that the responses of the people in the scanning group are random responses.

Finally, it seems that respondents in the scanning group tend to select extreme answers more often than respondents classified in other groups. For instance, as it is shown in Table 11 in the total population the extreme responses SD and SA to Sentence 26 correspond to 9.3% and 19.8% respectively. In "scanning" group the corresponding percentages are 16.8% and 25.7%. Of course, someone could argue, that the tendency towards the two extreme answers (i.e. SD and SA) among the scanning group could be a result of the aforementioned discrete uniform distribution, i.e. the relative frequencies in the total group is less than 20%, so the observed increased percentages in the scanning group is just a mere outcome of the tendency towards a discrete uniform distribution. But, as Table 12 indicates, the sum of the percentages of strong opinions in the scanning group is larger than the corresponding figure in the Normal speed group, even when this sum in the normal group is larger than 40% (for instance, see sentences 19 and 24).

[26] Murphy, J.; Hofacker, C. and Mizerski, R. (2006). Primacy and recency effects on clicking behaviour, Journal of Computer Mediated Communication, 11, 2, 522-535

Table 11 Distribution of answers given to Sentence 26 by response time category

		Sentence 26				
		SD	D	NN	A	SA
Scanning	Count	8316	5309	7587	15470	12700
	%	16.8%	10.8%	15.4%	31.3%	25.7%
Skimming	Count	10376	10131	12849	41629	26459
	%	10.2%	10.0%	12.7%	41.0%	26.1%
Normal	Count	23948	43529	64088	123916	51656
	%	7.8%	14.2%	20.9%	40.3%	16.8%
Unable to count	Count	104	222	253	519	313
	%	7.4%	15.7%	17.9%	36.8%	22.2%
Total	Count	42744	59191	84777	181534	91128
	%	9.3%	12.9%	18.5%	39.5%	19.8%

Table 12 Comparison of "strong" answer percentages between scanning and normal speed

Sentence	Scanning	Normal	Difference
2	0,49	0,35	0,14
3	0,56	0,31	0,24
4	0,45	0,28	0,18
5	0,40	0,30	0,10
6	0,48	0,36	0,11
7	0,42	0,30	0,12
8	0,41	0,30	0,11
9	0,44	0,34	0,10
10	0,43	0,28	0,15
11	0,42	0,25	0,17
12	0,48	0,33	0,16
13	0,49	0,39	0,10
14	0,36	0,21	0,15
15	0,43	0,29	0,14
16	0,42	0,24	0,18
17	0,40	0,20	0,20
18	0,37	0,30	0,07
19	0,54	0,49	0,05
20	0,43	0,37	0,06
21	0,39	0,27	0,12
22	0,43	0,27	0,16
23	0,48	0,38	0,11
24	0,61	0,54	0,06
25	0,43	0,30	0,14
26	0,43	0,25	0,18
27	0,38	0,30	0,08
28	0,44	0,36	0,08
29	0,47	0,36	0,11
30	0,40	0,36	0,04

Discussion

The present results have both theoretical and practical implications. Theoretically, the results offer support to the importance of recording the time users spent to answer each of the questions in a Voting Advice Application. Recorded response times can be useful in many ways. They can help to identify questions with larger response times than the expected response time for their length. This could be a sign of a badly expressed sentence that should be rephrased, replaced by another question or even totally removed. Response times can also help check if and when users get tired/bored and they start dedicating less time on answering the questions. Some of these ideas have been tested in the context of web surveys.

The main theoretical contribution of this paper is the idea that response times can be used to identify non-valid, unconsidered, incautious answers to VAA questions in order to clean the dataset. Following the notion of four tasks reported by Tourangeau et al. (2000), I have tried to isolate the time requested for the first task and link it with the length of the sentence, in order to classify the users according to their reading speed and total response time. The presented research provides a novel method to identify nonsense answers and demonstrates that VAA data cleaning based only on the pattern of answers is not adequate.

At the practical level, this research presents a series of findings regarding the frequency of the non-valid records and the distribution of answers in these records. It is note-worthy that non-valid answers, identified by the response time criterion, correspond to about 5% of the total answers. With regard to the distribution of the answers in these invalid records, there is a tendency towards a discrete uniform distribution. In addition, there is some evidence for the preference of the extreme answers (SD and SA).

After presenting the aforementioned findings, one final question remains: “If we analyse the data without removing the invalid cases, what will be the impact on findings and conclusions?” In other words, what would be the impact if 5% of a sample consisted of random answers? The answer depends on the analysis that has to be done. For instance, let’s go back to Table 9, and suppose that we need to report the percentage of people who disagree strongly with Sentence 2. If we used the total group (without cleaning) we would report the figure 10.5%, but if we used the “normal reading speed” group, we would give the answer 8.5%. This difference is not very large, but it could change the outcome of (say) a chi square test.

The bottom line is that recording response times can be implemented easily in a VAA environment and it can facilitate data cleaning by removing non-valid answers. Thus, I would like to conclude this paper by suggesting all VAA designers to record response times of their users, since this information could be proved to be really valuable for data cleaning and further research regarding the behaviour of VAA users.

References

Andreadis, I. (forthcoming) Voting Advice Applications: a successful nexus between informatics and political science. BCI '13, September 19 - 21 2013, Thessaloniki, Greece http://www.polres.gr/en/sites/default/files/BCI-2013.pdf

Andreadis, I. (forthcoming) Who responds to website visitor satisfaction surveys? General Online Research Conference GOR13 March 04-06, Mannheim, Germany http://www.polres.gr/en/sites/default/files/GOR13.pdf
Bassili, J.N. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. Public Opin.Q., 57, 1, 54-61.
Bassili, J.N. and Fletcher, J.F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. Public Opin.Q., 55, 3, 331-346.
Bassili, J. N., and B. S. Scott. (1996). Response latency and question problems. Public Opinion Quarterly 60 (3): 390–99
Carver, R.P. (1992) Reading rate: Theory, research, and practical implications, Journal of Reading, 1992, 36, 2, 84-95
Christian, L.M., Parsons, N.L., Dillman, D.A. (2009). Designing Scalar Questions for Web Surveys Sociological Methods & Research, 37, 3, 393-425
Dillman DA (2007). Mail and internet surveys: the tailored design method (2nd edition). New York, NY: John Wiley and Sons, Inc.
Fan, W. and Yan, Z. (2010) Factors affecting response rates of the web survey: A systematic review, Computers in Human Behavior, Volume 26, Issue 2, Pages 132–139 http://dx.doi.org/10.1016/j.chb.2009.10.015
Fry, E.B. (1963). Teaching faster reading: a manual. Cambridge: Cambridge University Press.
Heerwegh, D. (2003). Explaining response latencies and changing answers using client side paradata from a web survey. Social Science Computer Review, 21(3), pp.360-373
Heerwegh, D. (2004). Uses of Client Side Paradata in Web Surveys. Paper presented at the International symposium in honour of Paul Lazarsfeld (Brussels, Belgium June 4-5 2004)
Heerwegh, D. and Loosveldt, G. (2008) Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality, Public Opin.Q., 72, 5, 836-846
Krosnick, J.A. and Alwin, D.F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin.Q., 51, 2, 201-219.
Murphy, J.; Hofacker, C. and Mizerski, R. (2006). Primacy and recency effects on clicking behaviour, Journal of Computer Mediated Communication, 11, 2, 522-535
Simon, H.A. (1956). Rational choice and the structure of the environment. Psychological Review, 63, 2, p. 129
Stern, M. J. (2008). The Use of Client-side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys Field Methods November 20: 377-398
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. New York:
Cambridge University Press
Vicente, P. and Reis, E. (2012) The “frequency divide”: Implications for internet-based surveys. Quality & Quantity: 1-14.
Yan, T. and Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. Applied Cognitive Psychology, 22, 1, 51-68
Yannakoudakis, E.J., Tsomokos, I.and Hutton P.J. (1990) n-Grams and their implication to natural language understanding, Pattern Recognition, 23, 5, 509-528