Data quality and data cleaning

Ioannis Andreadis

Although VAAs are different from web surveys on various aspects, 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) 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 or they have Internet access but they never use it. Moreover, people who use the Internet more frequently are more prone to visit a VAA in a similar way they are more prone to participate to a web survey. Finally, even among frequent users there are differences regarding the type of use. For instance, Internet users who get online having as their primary task to play games are less probable to visit a VAA than people who get online to search for information (see Andreadis, 2013; Fan and Yan , 2010; Vicente and Reis, 2012).

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 (2008) 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 hand, VAAs are open to anyone with internet access. In addition, users can participate to a VAA as many times as they like. Web surveys that are open to the public (i.e. a pin/token is not required), may suffer by the same problem (multiple submissions by a single user). Some people may be motivated to participate to a survey multiple times by their intention to influence the findings of the survey by inflating the frequency of their views (e.g.

to make their favourite political party appear as more popular than it really is). We may observe this behaviour on users of unprotected web polls (usually with one question only) which publish the frequencies of the answers instantly. But when users complete a normal web survey, the only output they usually see is a "Thank you for your participation" screen. In order to learn the findings, web survey participants have to wait for the publication of the analysis of the collected data. Thus, people participate to surveys (web or any other mode) by a sense of social responsibility. 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. This process of playing with the Voting Advice Application can be called VAA testing.

From the previous paragraphs it is obvious that the quality of VAA data suffers of two major shortcomings: i) lack of representativeness due to limited coverage, and ii) measurement error due to VAA testing. More information on the difference between the group of VAA users and the general population can be found on the following chapter of this book that presents the profile of VAA users. As Internet use spreads to groups with lower access rates the difference between the group of VAA users and the general population becomes smaller. The aim of this chapter is to address the error that results from VAA testing by answering the following questions: How can we discover the nonsense answers submitted by users who were just testing the VAA? How serious is the problem, i.e. what is the percentage of nonsense answers? What are the differences between VAA testing cases from the rest of the cases? The chapter concludes with implications and suggestions for VAA designers and researchers working with VAA data.

Response Time

Item response time, i.e. time spent to answer a survey 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, 2008; Heerwegh, 2003; Heerwegh, 2004). Measuring response time is common in the survey literature. In fact, it is so common that many different measuring approaches have 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. 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 net 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. Of course, client-side time counting depends on the settings of the users' browser, i.e. if the settings prevent the execution of any script, then it is not possible to run anything on the client-side. Thus, in order to minimize the number of cases with missing values, response time should be measured with the simplest and most widely installed scripting language.

Estimating the thresholds

Tourangeau et al. (2000) 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.

For the common respondent, 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 spent on reporting is affected by how many and what type of response categories are offered. For instance, previous results 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 (Christian, Parsons, and Dillman, 2009). 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).

According to Yan and Tourangeau (2008), retrieval and judgment may be determined by respondent characteristics (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 do not use respondent characteristics in the analysis presented in this chapter.

Time dedicated to judgement depends on the existence or not of an attitude on the topic. People with a pre-existent opinion/position are expected to answer faster than people who decide on the spot. Even between people who have an attitude, time will depend on the attitude strength. People with unstable positions need more time to finalise their answer than people with a stable position who do not need to spend more time than the time to retrieve their already processed opinion from their memory. Bassili and Fletcher (1991) have found a positive relationship between response latency and unstable positions (measured as changes of the answer after being exposed to the counterargument). 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.

Much of the time spent on Task 1 (comprehension of the question) involves reading and interpreting the text. One component of this time is related to the complexity of the question. As Bassili and Scott (1996) have shown, 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. 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 comprehension of the ideas 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, because it is possible to have a sentence with a small number of lengthy words that is longer and requires more reading time than another sentence with more but shorter words. To avoid similar problems, I have decided to use the number of characters instead of using the number of words.

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). 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. Thus, if we want to classify a reading rate to one of the three aforementioned categories, we can use the following rule:

- reading rate ≤ 375 wpm → rauding,
- 375 wpm < reading rate ≤ 525 wpm → skimming
- 525 wpm< reading rate → scanning

Using these rules, 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. 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.

For English texts the average word length is 4.5 letters (see Yannakoudakis, Tsomokos and Hutton, 1990). Thus the above rules converted to characters per second (with 4.5 characters per word) give the following:

- reading rate ≤ 28.125 cps → rauding,
- 28.125 cps < reading rate ≤ 39.375 cps → skimming
- 39.375 cps < reading rate → scanning

If we divide the number of characters (without spaces) in each sentence with the number 39.375, we can 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.

Bassili and Fletcher (1991), using an active timer, 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. In their experiment, time counting starts when the interviewer presses the spacebar after reading 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). For VAAs and web surveys time counting stops when the user clicks on one of the available buttons that correspond to answer options. This additional step requires some extra time. Thus, the minimum time reported by Bassili and Fletcher for simple attitude questions (1.4 seconds) can be used as the minimum time for Task 4 (selecting and reporting the answer).

Consequently, the item response time of scanning respondents should be less than: Threshold1=1.4+[Characters in sentence without spaces]/39.375 and the corresponding time of skimming respondents should be between Threshold1 and Threshold2=1.4+[Characters in sentence without spaces]/28.125. Users, who have spent on a sentence less time than Threshold2, are suspected of answering without understanding the statements. For most people the time given by the formula of Threshold2 is not enough, but there may be some VAA users who are very fast readers and they are capable of understanding the statement just by

skimming the text. Thus, if a more strict rule is to be preferred, this is given by Threshold1: if a user has spent on a sentence less than the time of Threshold1, 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 was testing the application (e.g. to see the output it provides when all answers are "Neither agree nor disagree"). Thus, Threshold1 will catch a smaller number of cases suspected of being invalid, but the probability of these cases to be invalid is higher. Only extremely capable readers would be able to read and comprehend the exact meaning of a statement by just scanning the text.

Using the thresholds

In this section I apply the methodology described in the previous section on a dataset from the Greek VAA *HelpMeVote* which was used for the Greek Parliamentary Elections of 2012. For the election of May 2012 *HelpMeVote* includes 30 statements displayed on separated pages, but 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 one. The time between clicks can be counted accurately. The response times are recorded in hidden input fields. 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. A presentation of all the technical details of *HelpMeVote* (including the statements that have been used) can be found in Andreadis (forthcoming). Table 7.1 shows the thresholds used to classify the answers to each question/sentence.

As an example of the output of this classification I use the second sentence. As Table 7.2 shows about 5 per cent of the answers have been given in less than 3.838 seconds, i.e. the users were scanning and the dedicated time was not enough to give a valid answer. The second category (3.4 per cent) consists of answers that were given in less than 4.813 seconds and more than 3.838 seconds. Users in this category were fast, but it is possible that some of these answers are valid. Most of the users (about 90.3 per cent) have spent more than 4.813 seconds. Finally, there are some users (1 per cent) 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.

Up to this point I have used the thresholds developed in the previous section to classify a single answer in one of the following groups: scanning, skimming or normal. But if a user has answered only one or two questions with a scanning or skimming speed, this does not mean that all 30 answers are invalid. In order to classify a total row as invalid we need at least half of the answers to belong to one of the first two categories. Following this rule I find that more than 1 out of 20 cases of the dataset have been submitted by users who have not spent enough time to read and comprehend the VAA sentences (Table 7.3). The next question

that should be answered is the following: Are these cases which are classified as invalid different from the rest of the cases?

Table 7.1 Thresholds used to classify answers

Sentence	Number of characters (without spaces)	$Threshold_1$	$Threshold_2$
1	68	3.127	3.818
2	96	3.838	4.813
3	127	4.625	5.916
4	73	3.254	3.996
5	83	3.508	4.351
6	62	2.975	3.604
7	72	3.229	3.960
8	83	3.508	4.351
9	105	4.067	5.133
10	78	3.381	4.173
11	80	3.432	4.244
12	94	3.787	4.742
13	67	3.102	3.782
14	61	2.949	3.569
15	87	3.610	4.493
16	84	3.533	4.387
17	148	5.159	6.662
18	73	3.254	3.996
19	46	2.568	3.036
20	69	3.152	3.853
21	76	3.330	4.102
22	120	4.448	5.667
23	96	3.838	4.813
24	134	4.803	6.164
25	65	3.051	3.711
26	107	4.117	5.204
27	67	3.102	3.782
28	62	2.975	3.604
29	73	3.254	3.996
30	38	2.365	2.751

Table 7.2 Distribution of time spent on Sentence 2

	Frequency	Percent
Scanning	25,095	5.3
Skimming	16,427	3.4
Normal	430,786	90.3
Unable to count	48,27	1.0
Total	477,135	100.0

Table 7.3 Valid and invalid cases according to response time

	Frequency	Percent
Normal	438,132	91.8
VAA testing	25,051	5.3
Unable to count	13,952	2.9
Total	477,135	100.0

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 very fast speed (which I have classified as invalid or nonsense answers) and the answers given by people who have dedicated enough time to give a substantial response. Of course 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 7.4 Distribution of answers given to Sentence 2 by response time category

	SD	D	NN	A	SA
VAA testing	23.8%	14.5%	17.6%	22.2%	21.8%
Normal	9.8%	15.1%	11.4%	37.1%	26.6%

As Table 7.4 indicates, most Greek voters agree with the second sentence (together A and SA answers correspond to more than 63 per cent of the total answers) and only 9.8 per cent answer that they strongly disagree. But, within the invalid group we observe that the most frequent answer is SD (23.8 per cent) and

all other options are selected with about the same probability (D: 14.5 per cent, NN: 17.6 per cent, A: 22.2 per cent, and SA: 21.8 per cent). 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)[1]. Psychologists also support that primacy could be a result of satisficing[2], i.e. respondents choose the first acceptable answer instead of the optimal answer (see Simon, 1956). Krosnick and Alwin (1987) have shown that response order effects (both primacy and recency) are stronger among respondents low in cognitive sophistication. 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 (see Murphy, Hofacker and Mizerski, 2006). It seems that visitors click on the first link more frequently than any other link (primacy effect). The click-through rate decreases for all subsequent links except the last one, where it increases significantly (recency effect).

Table 7.5 Distribution of answers given to Sentence 18 by response time category

	SD	D	NN	A	SA
VAA testing	21.2%	25.1%	23.2%	18.1%	12.4%
Normal	26.5%	37.2%	13.9%	18.3%	4.2%

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 7.5) with the distribution of answers to sentence 18). In the normal group the sum of SD and D responses to sentence 18 is 63.7 per cent. On the other hand, only 4.2 per cent of the users select the answer SA. Within the VAA testing group the answers are distributed more uniformly and SA is selected by 12.4 per cent. It seems that among VAA testers, 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 testing group are random responses.

[1] 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.

[2] A combination of "satisfy" and "suffice", i.e. to finish a job by satisfying the minimum requirements.

Table 7.6 Comparison of Cramer's V between Normal and VAA testing

	Cramer's V	
Sentence	Normal group	VAA Testing
q1	0.326	0.194
q2	0.259	0.085
q3	0.153	0.160
q4	0.146	0.081
q5	0.206	0.078
q6	0.244	0.133
q7	0.212	0.115
q8	0.292	0.103
q9	0.352	0.188
q10	0.131	0.065
q11	0.174	0.049
q12	0.146	0.061
q13	0.356	0.134
q14	0.189	0.107
q15	0.322	0.096
q16	0.224	0.111
q17	0.235	0.081
q18	0.280	0.111
q19	0.304	0.255
q20	0.300	0.138
q21	0.139	0.090
q22	0.184	0.145
q23	0.152	0.074
q24	0.498	0.263
q25	0.304	0.166
q26	0.272	0.104
q27	0.358	0.163
q28	0.292	0.159
q29	0.352	0.156
q30	0.372	0.195

The usual test for the null hypothesis that a sample follows a particular theoretical distribution is the Chi-Square Goodness of Fit test. For the group of VAA testers, it seems that the observations tend to follow a discrete uniform distribution, i.e. all answers seem to occur with equal frequency. Since, there are 5 substantial answers, the expected relative frequency of each category under the null hypothesis is 0.2. We can test both the normal group and the VAA testing group against the null hypothesis and observe which of the two groups is closer

to the theoretical distribution. If the number of cases was equal in both groups I could directly compare the Chi-square values. But since the number of cases in the normal group is much larger than the number of cases in the VAA testing group, it is better to compare the values of Cramer's V which does not depend on the number of cases. In Table 7.6 I compare Cramer's V statistics calculated for the Goodness of Fit to the uniform discrete distribution between the Normal group and the VAA testing group. It becomes obvious that for all questions (except one where the coefficients are practically equal) the distribution of answers in the VAA testing group is closer to a uniform discrete distribution than the distribution of answers in the Normal group.

Pattern of answers and relation with response time

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 7.7 Frequencies of fixed answers (rigid: 30 identical answers)

	Frequency	Percent
Strongly Disagree	715	11.6
Disagree	65	1.1
Neither … nor	1,486	24.1
Agree	61	1.0
Strongly Agree	300	4.9
No answer	3,543	57.4
Total	6,170	100.0

Table 7.7 indicates that there are 6,170 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 per cent of the constant answer records). The next most used constant answer is the median "neither agree nor disagree" point (24.1 per cent). "Strongly disagree" (11.6 per cent) and "Strongly agree" (4.9 per cent) 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 some users had the intention to click on the same answering button for each question but while they were trying to do this at a high speed, they accidentally clicked one or more times on a different button. VAA researchers, who want to have their VAA data as clean as possible can follow a method to identify these cases that is available in Andreadis (2012). In the same paper it is shown that 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.

This means that if a voting advice application does not log the time spent to each sentence, the collected data cannot be fully cleaned.

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 chapter 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 five per cent 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.

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 five per cent 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 4, and suppose that we need to report the percentage of people who disagree strongly with Sentence 2. If we used the total sample (without cleaning) we would report the figure 10.5 per cent, but if we used the "normal reading speed" group, (i.e. what remains from the total sample after removing the invalid cases) we would give the answer 8.5 per cent. 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.

Bibliography

Andreadis, I. (2013) 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. (2013) 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/GOR2013.pdf

Andreadis, I. (2012) To Clean or not to Clean? Improving the Quality of VAA Data, Presented at the XXIInd World Congress of Political Science, Madrid 8-12 July 2012 http://www.polres.gr/en/sites/default/files/IPSA-2012.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