Web Survey Response Times
What to Do and What Not to Do

Ioannis Andreadis[1]
[1]School of Political Sciences, Aristotle University of Thessaloniki, 54124, Greece

Abstract
In this paper, I deal with web surveys response times as well as their association with data quality and I provide guidelines and suggestions for web survey scholars and practitioners working with response times. Firstly, I highlight the limited usefulness of the total time spent to complete the whole questionnaire and I argue that our primary focus should be on the response times of each item separately. Furthermore, I show that occurrence of long item response times are usually random and not associated with specific users or items and that extremely long times usually occur after an external distraction (e.g. an incoming email or phone call). Then, I suggest a method that we can use to flag responses that were given in extremely short time (i.e. a response time so short that shows that the flagged respondents instead of reading and comprehending the question, they have scanned the question text). Finally, I compare the suggested method with the method that is most widely used today for speeding detection.

Key Words: web surveys, response times, data quality

1. Introduction to survey response times

Survey response times belong to a special type of data called paradata (Heerwegh, 2003, 2004). These data are non-reactive, and they do not provide information about the respondent's answers. Instead, they provide information about the process of answering the questionnaire, i.e., paradata provide information on how the respondents have interacted with a survey. Item response times and overall survey completion response times of web surveys have attracted the attention of many researchers recently, because longer web surveys suffer from larger break-off rates and greater probability of lower quality responses. Shorter response times can be a sign of burden and an indicator of low response quality. For instance, very short response times may indicate that the respondents have not read the question carefully or that they have not even completely skipped question reading. Not surprisingly, it has been shown that very fast respondents (i.e. when their item response times are below specific thresholds) appear to give random answers, introducing noise to the final dataset (Andreadis, 2012, 2014).

Response times have been measured in various ways the survey literature. For instance, there are two types of proposed timers: 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 variable of collecting response times refers to the side where time is measured (server-side vs client-side). Measurement on the server side is done by taking advantage of the timestamp that is recorded each time a respondent visits a web page. By calculating the difference between the timestamps of two consecutive pages we can get a measure of the time spent on the first page, i.e., with this method the end time of the first page is the same with the start time of the second page. This means that in order to count time spent on each question, we need to keep each question on a separate web 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 spent within the page plus the time spent between the pages. The time spent between the pages is the sum of the transmission time (from the moment the respondent submits the answer and the moment the answer is recorded on the server) and server processing time (from the moment the answer is recorded on the server until the next page is requested). The time spent between pages depends on the type and bandwidth of the respondent's internet connection, but also on unpredicted, temporary delays due to network load, etc. On the other hand, client-side time measurement is done at the level of the respondent's (or client's) computer itself. This time data comes from JavaScript code embedded in each page and it is a more accurate estimate of the time spent on answering a question because it does not include the additional between-pages time.

Couper and Peterson (2017) have used both server- and client-level times in order to disentangle between-page (transmission) times from within-page (response) times and they report that mobile respondents took significantly longer to complete the survey than PC respondents, and that most of this difference is due to within-page times. In compliance with their finding I argue that transmission times are less important than response times for two reasons: i) issues related to the speed of mobile Internet will eventually be eliminated as mobile Internet providers improve their services and ii) new technologies enable web survey designers to download the next pages of the questionnaire to the respondents' browser before these pages are requested, i.e. eliminating any transmission delays.

2. Factors affecting response times

Some of the respondents' characteristics are known to affect response times. For instance, most of the studies in the literature (Andreadis, 2015b; Couper & Kreuter, 2013; Yan & Tourangeau, 2008) tend to agree that age (older people spend more time) and education level (respondents with lower education levels spend more time) have a significant impact on response times. In addition, some studies have found that respondents with clear, pre-existent opinion/position, interested in the survey topic and male respond faster than respondents with more uncertain attitudes, less interest in the topic and female, respectively (Andreadis, 2015a, 2015b; Bassili & Fletcher, 1991). Even between people who have an attitude, time will depend on the attitude strength, because 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. 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.

Response times are also affected by question characteristics. The type of question is one of these characteristics. For instance, prior research has identified grid questions to increase response times, especially for mobile users due to the additional scrolling required for grid questions on mobile devices (Couper & Peterson, 2017). In addition, 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 also 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. More generally, response time is longer for questions and formats that are difficult for respondents to process (Christian et al., 2009). Finally, even for the most simple question type the length of the question text is known to have impact on the response time (Andreadis, 2012, 2014).

Other factors are related to the overall setting and environment of survey participation. Heerwegh and Loosveldt (2008) argue, that 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 response times. In addition, the device used to participate in the survey may have a significant impact on response times. Mavletova (2013) analyzing an experiment with two survey modes conducted using a volunteer online access panel in Russia, reports that the mean time of questionnaire completion for mobile surveys was 3 times longer than the mean time for computer web surveys. On the other hand, Toepoel and Lugtig (2014) offering a mobile-friendly option to respondents of an online probability-based panel organized by a research consultancy agency in the Netherlands, find that the total response times are almost the same across devices. Finally, Andreadis (2015a) estimates that that switching from desktop to smartphone the geometric mean of response times is expected to increase by 17%.

3. Dealing with extreme response times

Extremely long times are the result of an interruption that usually occurs after an external distraction (e.g., an incoming email, phone call, door knocking, etc). Thus, the occurrence of extremely long response times is not associated neither with a respondent nor with an item. Thus, the occurrence of extremely long response times is random, and it can be identified both by looking for extremely long times per item and by looking for extremely long times per respondent.

A good way to look for extreme response times within a respondent is to use the methods of exploratory data analysis and more specifically statistics used for boxplots. Boxplot statistics can identify outliers i.e., values between the inner and the outer fences of the boxplot and extreme values, i.e., values outside off the outer fences. The problem of applying this method on the response times is that it would flag as extreme too many values that are not extreme, because the distribution of response times is usually highly skewed to the right. Thus, the logarithmic function should be applied to the response times before the application of the aforementioned exploratory data analysis method for the identification of extreme values.

After flagging the extremely long response times, they should be recoded as missing values. The logic behind this argument is very simple. We cannot leave them intact, because the recorded time is not the actual time spent on the question, but the sum of the time spent on answering the question, plus an unknown amount of time due to some external distraction. We should not remove the whole record, because we do not have a respondent giving invalid answers. Thus, the best way of dealing with these values is to consider them as missing, because the external distraction that interrupted the respondent has prevented us from recording the actual time spent on the item. By recoding the extremely long response times as missing, we do not allow them to distort the average response times estimated by the sample. At the same time, we do not have to disregard the whole row, and if required, we can impute the missing values.

Since the extremely long response times correspond to a temporary break from the survey, in most cases a respondent spends extremely long times on a very limited number of items. On the other hand, respondents who respond extremely fast to one question they usually rush while responding to most of the questions. This difference has a very good explanation: extremely short times are the result of a decision made by respondents who decide to respond without paying too much (or even any at all) attention to the questions. For this reason, extremely short times should be treated differently as described in the following section.

4. Response times thresholds

The main ideas on minimum response times used in this section have been first published almost twenty years ago (Andreadis, 2012). In this section, I will present the main ideas, and the threshold we can develop based on these ideas. Before answering a survey question, a respondent needs to spend some:

- Time to Read and Comprehend the question and the available response options (TRC), and
- Time to Select and Report an answer (TSR).

The time spent on reading and comprehension depends on respondent characteristics (e.g., age, education level) and the length and complexity of the question. The time spent on selecting and reporting an answer is affected by question type and the number of response options offered. For single choice items, the reporting procedure is very simple; 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).

Much of the time spent on the first task involves reading and interpreting the text. Survey respondents 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 survey, 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.

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 respondent 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 respondent 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 (Yannakoudakis et al., 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, or simply ~40, we can get the minimum time (in seconds) that is necessary to read the sentence. Thus, the Minimum Time to Read and Comprehend (MTRC) can be calculates as **MTRC=NC/40**, where NC is the number of characters (without spaces) of the question text (including response options). The above formula corresponds to the assumption that even the fastest reader would need at least 10 secs to read a question of 400 chars.

Of course, respondents need some time for the second task. 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. Thus, the minimum time reported by Bassili and Fletcher for simple attitude questions (1.4 seconds) can be used as the Minimum Time to Select and Report an answer (MTSR).

Consequently, the minimum response time (MRT) for a simple attitude question is:

$$\mathbf{MRT=MTSR+MTRC=1.4+NC/40}$$

This means that a question of 120 characters would take at least 1.4+120/40=4.4 seconds. Scanning respondents would spend on a question less than MRT. Thus, if a respondent has spent on a sentence less than MRT, the dedicated time was not enough for a valid answer; the answer was given either by randomly clicking on any of the available response options or the respondent has clicked on a fixed button for all sentences, e.g. the respondent was testing the application. Only extremely capable readers would be able to read and comprehend the exact meaning of a statement by just scanning the text. This method has been used as one of the data quality indicator in various studies and the cleaning of various datasets (Andreadis & Kartsounidou, 2020; Hameleers et al., 2018, 2019).

The aforementioned MRT is suitable for simple attitudinal questions. Matrix/Grid/Array questions include a few sub-questions (or items) that share the same response options. In this case, the respondent needs to spend time to select and report an answer for each sub-question. Consequently, the formula should be adapted as follows:

$$\mathbf{MRT=MTSR*NS+MTRC=1.4*NS+NC/40}$$

where, NS is the number of the sub-questions in the matrix question

5. Comparison of Speeding Methods

Zhang and Conrad (2014) have also suggest a speeding detection formula. by Although they follow similar ideas and the same source for the classification of reading speeds, their approach has two significant differences from the method presented here: i) their reading speed threshold is set to 300 milliseconds per word, which corresponds to 200 wpm i.e., much slower that the typical reading speed (300 wpm) and more suitable for learning, and ii) they do not split between TRC and TSR. As a result, if a simple question and a matrix question have the same length, according to their method they should have the same response time.

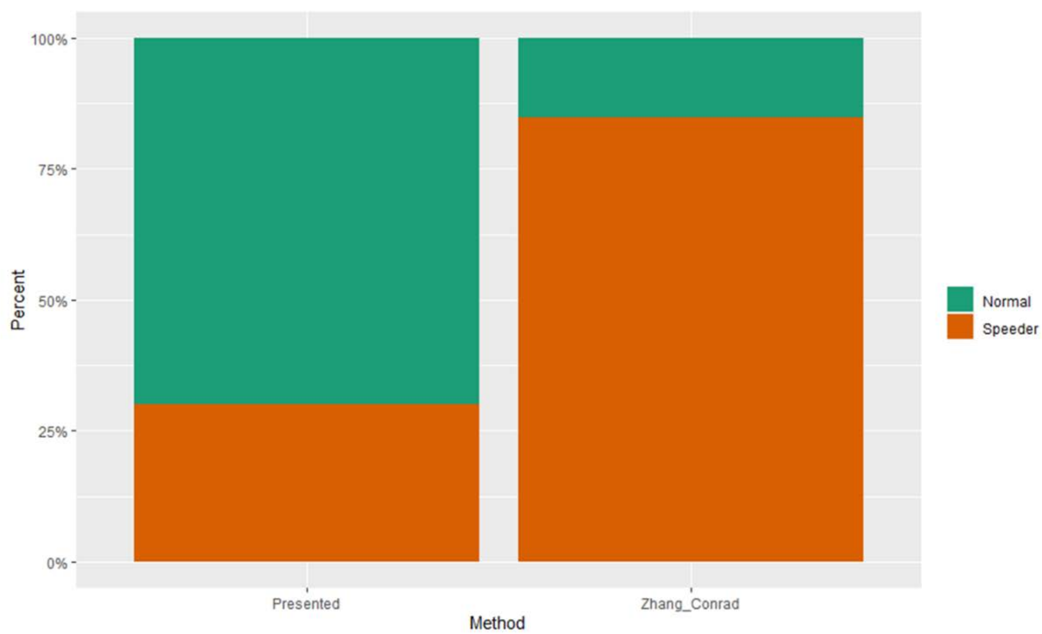

Figure 1: Comparison of the number of speeders (at least once)

These two significant differences between the method proposed by Zhang and Conrad (2014) and the method presented here have an important impact on the quality of speeding detection. Figure 1 shows a comparison of the number of speeders detected by these two methods on a dataset collected from a sample of students at Aristotle University of Thessaloniki using the 2020 questionnaire of the International Social Survey Programme. Using the Zhang-Conrad method almost 8 out of 10 respondents are detected as speeders at least once. This ratio is much higher than the ratio of speeders detected by the presented method. The main reason for this extremely high ratio of speeders is the very slow reading speed that they have used as their threshold.

Figure 2 shows the comparison of the two methods per question type. The ratio of speeders in matrix questions with more than or equal to 5 sub-questions detected by both methods is low and almost identical. On the other hand, when we compare the ratio of speeders in matrix questions with less than 5 sub-questions or in single questions, the Zhang-Conrad gives a much higher rate of speeders.

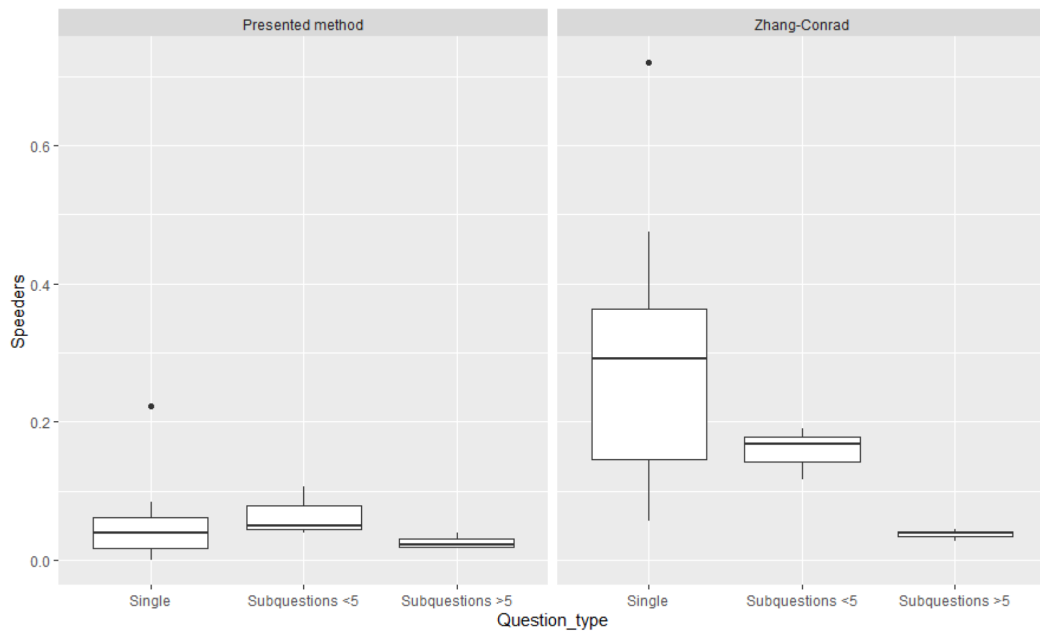

Figure 2: Single vs Matrix questions

5. Discussion

This paper provides a formula that can be used to flag responses which were given so quickly that the response is probably not valid. The method is based on the decomposition of the survey response process into components and a threshold is estimated as the sum of two minimum times: 1) the minimum time for the comprehension of the question and the minimum time for the selection and reporting of an answer. The estimation of the minimum time needed for the comprehension of the question is based on the length of the question text and the methods uses the time the respondents would need to spend on a text of that length if their reading speed was classified as faster than skimming and closer to scanning. For the minimum time for the selection and reporting of an answer the method uses a fixed amount (1.4 seconds) which according to the literature is the minimum time respondents need to select and repot an answer to simple attitude question. The formula is easily adapted for matrix question by multiplying the minimum time for the selection and reporting of an answer by the number of sub-questions in the matrix.

The main theoretical contribution of this chapter is the idea that response times can be used to identify non-valid, unconsidered, incautious answers to web survey questions to clean the dataset. At the practical level, this paper has shown that most widely known method for speeder detection has two serious drawbacks because it uses as a threshold a very slow reading speed and because it does not treat differently simple and matrix questions.

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

Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 3572).

References

Andreadis, I. (2012). To clean or not to clean? Improving the quality of VAA data. *XXII World Congress of Political Science (IPSA)*. http://ikee.lib.auth.gr/record/132697/files/IPSA-2012.pdf

Andreadis, I. (2014). Data Quality and Data Cleaning. In S. Garzia, D. Marschall (Ed.), *Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective* (pp. 79–91). ECPR Press.

Andreadis, I. (2015a). Comparison of Response Times between Desktop and Smartphone Users. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile Research Methods: Opportunities and challenges of mobile research methodologies* (pp. 63–79). Ubiquity Press. https://doi.org/10.5334/bar.e

Andreadis, I. (2015b). Web surveys optimized for smartphones: Are there differences between computer and smartphone users? *Methods, Data, Analysis*, *9*(2), 213–228. https://doi.org/10.12758/mda.2015.012

Andreadis, I., & Kartsounidou, E. (2020). The Impact of Splitting a Long Online Questionnaire on Data Quality. *Survey Research Methods*, *14*(1), 31–42. https://doi.org/10.18148/SRM/2020.V14I1.7294

Bassili, J. N. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. *Public Opinion Quarterly*, *57*(1), 54–61.

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346. https://doi.org/10.1086/269265

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*(3), 390–399.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *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. https://doi.org/10.1177/0049124108330004

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Couper, M. P., & Peterson, G. J. (2017). Why Do Web Surveys Take Longer on Smartphones? *Social Science Computer Review*, *35*(3), 357–377. https://doi.org/10.1177/0894439316629932

Hameleers, M., Andreadis, I., & Reinemann, C. (2019). Investigating the effects of populist communication. In C. Reinemann, J. Stanyer, T. Aalberg, F. Esser, & C. H. de Vreese (Eds.), *Communicating Populism* (pp. 168–182). Routledge. https://doi.org/10.4324/9780429402067-9

Hameleers, M., Bos, L., Fawzi, N., Reinemann, C., Andreadis, I., Corbu, N., Schemer, C., Schulz, A., Shaefer, T., Aalberg, T., Axelsson, S., Berganza, R., Cremonesi, C., Dahlberg, S., de Vreese, C. H., Hess, A., Kartsounidou, E., Kasprowicz, D., Matthes, J., … Weiss-Yaniv, N. (2018). Start Spreading the News: A Comparative Experiment on the Effects of Populist Communication on Political Engagement in Sixteen European Countries. *The International Journal of Press/Politics*, *23*(4), 517–538. https://doi.org/10.1177/1940161218786786

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*(3), 360–373.

Heerwegh, D. (2004). Uses of Client Side Paradata in Web Surveys. *Prispevek Na Simpoziju International Symposium in Honour of Paul Lazarsfeld, Brussels, Belgium*.

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. https://doi.org/10.1093/poq/nfn045

Mavletova, A. (2013). Data Quality in PC and Mobile Web Surveys. *Social Science Computer Review*, *31*(6), 725–743. https://doi.org/10.1177/0894439313485201

Toepoel, V., & Lugtig, P. (2014). What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review*, *32*(4), 544–560. https://doi.org/10.1177/0894439313510482

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. https://doi.org/10.1002/acp.1331

Yannakoudakis, E. J., Tsomokos, I., & Hutton, P. J. (1990). N-Grams and their implication to natural language understanding. *Pattern Recognition*, *23*(5), 509–528. https://doi.org/10.1016/0031-3203(90)90072-S

Zhang, C., & Conrad, F. (2014). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453