Investigating the Adequacy of Response Time Outlier Definitions in Computer-based Web Surveys using Paradata SurveyFocus

Jan Karem Höhne
University of Goettingen (Germany)

Stephan Schlosser
University of Goettingen (Germany)

Abstract
Web surveys are commonly used in social research because they are usually cheaper, faster, and simpler to conduct than other modes. They also enable researchers to capture paradata such as response times. Particularly, the determination of proper values to define outliers in response time analyses has proven to be an intricate challenge. In fact, to a certain degree, researchers determine them arbitrarily. In this study, we use "SurveyFocus (SF)" – a paradata tool that records the activity of the web-survey pages – to assess outlier definitions based on response time distributions. Our analyses reveal that these common procedures provide relatively sufficient results. However, they are unable to detect all respondents who temporarily leave the survey, causing bias in the response times. Therefore, we recommend a two-step procedure consisting of the utilization of SF and a common outlier definition to attain a more appropriate analysis and interpretation of response times.

Keywords: JavaScript, response behavior, response quality, usability study, survey participation

Introduction and Background

Web surveys are increasingly being used for data collection in social research since they offer several benefits: cost-effectiveness, time saving, and the collection of paradata. According to Couper (2000, p. 393) paradata automatically arise during computer-assisted data collection. As a special type of paradata, response times enjoy a long tradition in survey research (Fazio, 1990) and have been used to investigate the mental accessibility and stability of attitudes or to evaluate the complexity of information processing (Mayerl, 2013). Furthermore, they have been used as an indicator of cognitive effort (Höhne, Schlosser, & Krebs, 2017; Lenzner, Kaczmirek, & Lenzner, 2010) or to assess response quality (Höhne, Schlosser, & Krebs, 2017).

As these examples show, the measurement of response times in web surveys is a suitable strategy for shedding light on response behavior and quality of the responses given. However, due to the fact that web surveys are usually based on self-administered modes – implying a spatial distance between respondent and interviewer or researcher – it is difficult or even

---

This document is a preprint and thus it may differ from the final version: Höhne, J.K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata SurveyFocus. Social Science Computer Review, 36, 369–378. DOI: 10.1177/0894439317710450.

impossible to oversee how respondents process surveys (Couper, 2000). Therefore, researchers cannot be sure whether respondents are processing the survey all at once or whether they are being distracted during its completion – e.g., leaving the survey for short periods to check emails, switching browser tabs, or starting applications. Toninelli and Revilla (2016), for instance, show that approximately 13% of respondents read emails, 4% chat with others, and another 33% do other activities while they are participating in surveys. These "multi-tasking" behaviors impede adequate interpretation of response times because they are biased upwards, irrespective of technically correct measurement. Therefore, the evaluation of response behavior and/or response quality by means of response times seems to be somewhat problematic.

When measuring response times, there are also challenges regarding their proper handling. Grubbs (1969), for instance, gives special attention to cases with unusual short or long response times that seem to vary noticeably from other cases in the sample. The survey literature proposes a variety of strategies for dealing with cases such as these. Most commonly, researchers calculate specific "thresholds" based on the response time distributions (see Fazio, 1990; Lenzner, Kaczmirek, & Lenzner, 2010; Mayerl, 2013). Response times below or above these thresholds are called "outliers" and are excluded from data analyses. However, one key problem of defining such response time outliers is the determination of appropriate thresholds.

To determine response time outliers, Mayerl (2013) used two standard deviations below or above the mean. Schnell (1994) proposed using the median plus/minus the interquartile range (IQR) multiplied by 1.5 and Hoaglin, Mosteller, and Tukey (2000) the median plus/minus the upper and lower quartile range multiplied by three. In contrast, Lenzner, Kaczmirek, and Lenzner (2010) applied the upper and lower one percentile as thresholds. As these examples illustrate, the definition procedures result in different amounts of outliers. Furthermore, since response time distributions are typically right-skewed (Fazio, 1990), primarily respondents with "longer" response times will be defined as outliers. Although not affected by the skewness of a distribution, the percentile strategy results in a fixed proportion of outliers and is thus detrimental when no unusual short or long response times exist.

In consideration of how best to overcome the flaws of common outlier definitions based on response time distributions in web surveys, Höhne, Schlosser, and Krebs (2017) proposed a new paradata tool called "SurveyFocus (SF)". SF makes it possible to log the activity of the web-survey page and observe whether respondents leave the page for a certain time period prior to its completion (e.g., switching between browser tabs). If this is the case, respondents are classified as "SF:OFF" (i.e., declared as outliers) for the specific web-survey page and will be excluded from the response time analyses. "SF:ON" denotes that respondents have processed the page continuously. In addition, it can be determined how often and for how long respondents leave the web-survey page. Figure 1 illustrates the SF concept.

Regarding the definition of outliers, SF allows researchers to distinguish respondents who process the survey continuously, called SF:ON respondents, from those who process the survey discontinuously, called SF:OFF respondents. The additional use of SF contributes to a more profound as well as more precise and objective analysis of response times. In this study, we therefore investigate the adequacy of outlier definition procedures based on the response time distributions using SF.

Figure 1. SurveyFocus (SF): active and inactive web-survey pages

Outlier Definitions

For the collection of outlier definition procedures, we took several decision criteria into account to ensure general traceability of the selection process: first, the outlier definitions used must be cited in the literature and/or have been formerly employed in empirical studies. Second, we made sure that the outlier definitions employed are based on different statistical measures. Finally, we ensured that the procedures used vary in their strictness – i.e., in the reduction of case number. For example, the mean is susceptible to extreme values (i.e., comparatively short or long response times) of a distribution and is therefore not necessarily the best measure of the centrality of a distribution. The median, in contrast, is less affected by such extreme values. With respect to the strictness of a procedure, the use of different measures of dispersion is additionally important. For instance, the mean plus/minus two standard deviations (Mayerl, 2013) is less strict than the median plus/minus the interquartile range (IQR) multiplied by 1.5 (Schnell, 1994). Therefore, we use both procedures based on the mean and median combined with different measures of dispersion. We also employ a procedure based on fixed proportions. Table 1 contains the outlier definitions.

Table 1. Response time outlier definitions including the lower and upper thresholds

Definition Procedure	Lower Threshold	Upper Threshold
A	Mean – (2 × SD)	Mean + (2 × SD)
B	$Q_{.50} - (1.5 \times IQR)$	$Q_{.50} + (1.5 \times IQR)$
C	$Q_{.50} - (1.5 \times (Q_{.50} - Q_{.25}))$	$Q_{.50} + (1.5 \times (Q_{.75} - Q_{.50}))$
D	$Q_{.50} - (3 \times (Q_{.50} - Q_{.25}))$	$Q_{.50} + (3 \times (Q_{.75} - Q_{.50}))$
E	$< Q_{.01}$	$> Q_{.99}$

Note. A: Mayerl (2013); B: Schnell (1994); C and D: Hoaglin, Mosteller, & Tukey (2000); E: Lenzner, Kaczmirek, & Lenzner (2010).

Research Hypotheses

Different question presentation modes can be employed in web surveys. In survey practice, questions in single as well as grid presentation mode are most frequently applied (Couper, Tourangeau, Conrad, & Zhang, 2013). However, Dillman, Smyth, and Christian (2009) argue that especially grids imply much more intricate and difficult processing than single questions since they are more complex and thus more demanding. Respondents must pinpoint a number of question stems and response categories vertically and horizontally at the same time (Couper, Tourangeau, Conrad, & Zhang, 2013). In contrast, single questions in web surveys are usually presented individually on the web-survey page, requiring a much less complex task in terms of responding as well as coordination. It can be assumed that (on the basis of web-survey page) responding to grid questions causes longer response times and greater variations in the distributions. Hence, it seems appropriate to investigate single and grid questions separately.

In contrast to common outlier definition procedures, SF provides exact information about the activity of the web-survey page – i.e., how often and for how long respondents are away from the survey. This information allows more elaborated preparation of response time data for statistical analyses. More precisely, the use of SF renders a more differentiated consideration of response times than outlier definitions based only on response time distributions. We therefore hypothesize that common outlier definitions do not detect all SF:OFF respondents (*hypothesis 1*). In addition, we expect that this is more pronounced in grid than in single presentation mode.

From a psychological standpoint, the loss of SF implies an interruption of the response process. Respondents must re-orientate themselves and partially restart this process after returning to the survey. Moreover, assuming that the cognitive effort of answering a survey question is strongly related to the time required to respond (Fazio, 1990), this relation is weakened when the response process is interrupted for whatever reason. In contrast to the first hypothesis, we now compare response times and expect that SF:OFF respondents need significantly longer to respond – after correcting for the "time-out" – than SF:ON respondents (*hypothesis 2*). Again, we expect that this is more distinct for grid than single presentation mode.

Losing SF may also affect the answers of respondents in an undesirable way. Since respondents must re-orientate themselves and begin the cognitive processing anew after coming back to the web-survey page, it can be assumed that this increases the cognitive effort of responding and decreases respondents' motivation. This, in turn, might cause a superficial rather than a conscientious responding, which supports the occurrence of response bias and thus results in low response quality. In this study, we use item non-response as an indicator of "primarily bad" response quality. Based on our previous reasoning, we hypothesize that SF:OFF respondents produce significantly higher item non-response rates than SF:ON respondents (*hypothesis 3*). In line with the former hypotheses, we expect higher item non-response rates for questions in grid than in single presentation mode.

Study Design and Paradata

The data were collected at two German universities in May 2015. Respondents were invited by email. The invitation included an introduction to the topics of the study and a hyperlink that directed respondents to the survey (we sent out n = 58,829 emails). On the first page of the survey, respondents were informed about the procedure of the study and encouraged to read the questions carefully and process them in the given order. Respondents were also informed that different types of paradata are collected. At the end, they additionally received an opportunity to have their data deleted.

To gather response times in milliseconds (the time elapsing between question presentation on the screen and the time the page was submitted by clicking "Next") and the activity of the web-survey page (see Olson & Parkhurst, 2013), we used "Embedded Client Side Paradata (ECSP)" (Schlosser, 2016). ECSP is a JavaScript-based system that can be implemented in Unipark (QuestBack), a web-based survey software, to observe respondents' behavior.

Survey Questions

We used 24 questions: 8 single and 16 grid questions. The 8 single questions were adapted from the *Cross Cultural Survey of Work and Gender Attitudes* (2010) and dealt with achievement motivation. Each of these questions was displayed on a separate screen. The 16 grid questions were partially adapted from the *German General Social Survey* (2006) and dealt with job motivation. These 16 questions were presented in two grids with horizontally shaded rows, respectively. All questions were measured with 5-point, end-labeled decremental response scales (i.e., running from positive to negative) and no numerical labels.[1]

Sample

A total of n = 2,884 students took part, which corresponds to a response rate of 4.9%. Due to technical difficulties, respondents who processed the survey using a smartphone or tablet were excluded (n = 709).[2] We also excluded all respondents who had deactivated JavaScript (n = 28), only visited the first page (n = 123), dropped out of the survey before being asked any study-relevant question (n = 163), or whose mother tongue was not German (n = 110). All in all, n = 1,751 respondents remained for statistical analyses. These respondents were between 17 and 54 years old with a mean age of 24.9 and a standard deviation of 4.2. Of these respondents, 55% were female and at least 93% had previously participated in a web survey.

Analytical Strategy

We contrast SF with response time outlier definitions on web-survey page basis and thus check the activity for each page. According to our first hypothesis, we investigate whether outlier definitions based on the response time distributions are able to catch all discontinuously processing SF:OFF respondents. This will be conducted for each definition procedure A to E (see table 1) and for single as well as grid questions, respectively. In line with Höhne, Schlosser, and Krebs (2017) as well as our research question, we conduct the statistical comparisons on response times (*hypothesis 2*) and response quality (*hypothesis 3*) as follows: we first identify discontinuously processing respondents by means of SF (i.e., inactivity was determined per

---

[1] The wording of all 24 questions and screenshots are available from the corresponding author upon request.

[2] For mobile devices, the combination of different JavaScript events is necessary, which encumbers the determination of how often and for how long respondents are away from the web-survey page. It is only possible to detect whether a focus loss occurs or not.

page). Then, we apply the outlier definition procedures A to E only to the continuously processing SF:ON respondents and compare them to the discontinuously processing SF:OFF respondents. This is to guarantee an adequate and direct comparison, since undetected SF:OFF respondents could affect the determination of the lower and upper outlier definition thresholds.[3]

The analysis of paradata can be conducted on different aggregation levels (Heerwegh, 2003). Since there are no considerable differences on the page-level and to reduce the number of statistical tests as well as efficiently summarize the results, we aggregate the data for the 8 single as well as 16 grid questions, respectively.

The preparation of the paradata, which were collected by means of ECSP (Schlosser, 2016), and respondents' answers were conducted with R version 3.3.3. This also applies to all subsequent statistical analyses reported in this article.

Results

Comparing the Number of Outliers

To investigate our first hypothesis, we began by analyzing the number of response time outliers for each procedure. Afterwards, we contrasted them with SF, which resulted in three different outlier categories: "Procedure-Specific Outliers (PSO)", "SF:OFF", and "Overlap PSO/SF:OFF".[4] Figure 2 contains the descriptive results for single and grid questions. It can be observed that none of the five outlier definitions was capable of capturing all discontinuously processing respondents. Interestingly, the extent of this circumstance varies extremely across the procedures and depends on their strictness: the less strictly a procedure defines the lower and upper thresholds (i.e., the larger the interval of valid defined response times), the more discontinuously processing respondents remain in the dataset. Hence, less strict outlier definitions (e.g., procedures A and E) suffer more from biased response times than stricter definitions (e.g., procedure C). While there are substantial differences regarding the exclusion of SF:OFF respondents, there are small differences regarding the overlap. All in all, outlier definitions based on response time distributions produce slightly insufficient results for single questions.

We now take a look at grid questions. While procedures A and E again result in a large number of SF:OFF respondents, procedure C results in a very small one. However, as suggested in terms of our first hypothesis, the differences are much more pronounced for grid than for single questions. For instance, procedure A suffers from 96% and E from 64% biased response times. Even procedures B and D contain 17% and 20% discontinuously processing respondents, respectively. Hence, the number of respondents with biased response times remaining undetected is considerably higher for grid than for single questions and thus in line with our expectation.

---

[3] If the time-out of a web-survey page is not taking into account response times are additionally biased upwards, which affects the interval of valid defined response times.

[4] In total, 12.62% of all respondents left the web survey at least once with an average time-out of Median = 13.74 seconds. On the page-level, in contrast, the amount of discontinuously processing respondents is only up to 4.94%.

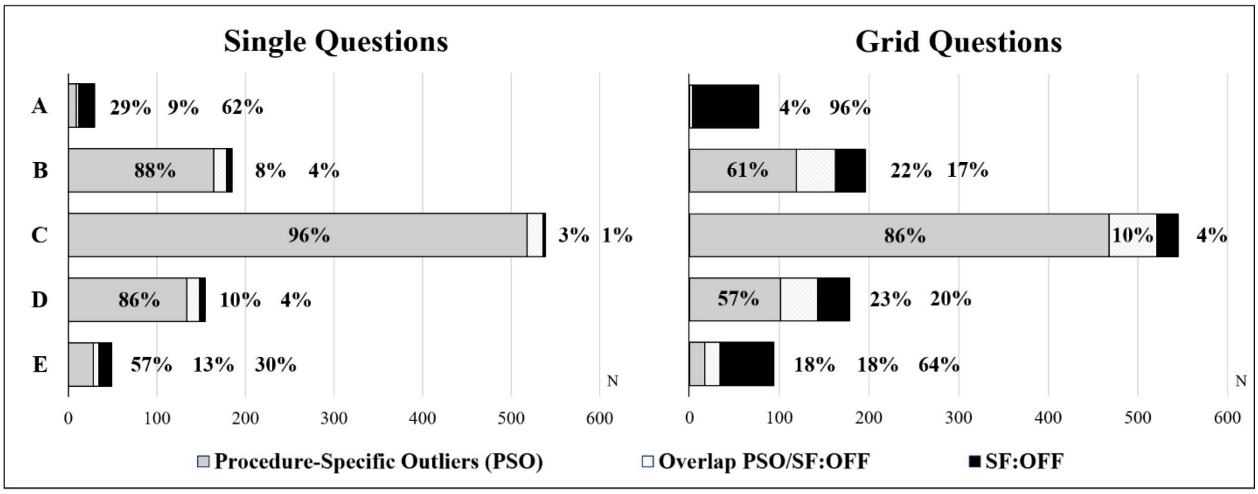

Figure 2. Total number of outliers for each procedure for single and grid questions
Note. The letters A to E on the y-coordinate refer to the outlier definitions in table 1. The numbers on the x-coordinate indicate the frequency of outliers and the numbers in the bars the percentage of outliers detected by the outlier definition procedures, SF, and the overlaps.

Comparison of Processing Time

As regards processing times, we postulated that discontinuously processing respondents would need significantly longer to process survey questions than continuously processing respondents, irrespective of the employed outlier definition procedure. To counteract the bias – due to undetected SF:OFF respondents – in response times, we applied the procedures A to E only on SF:ON respondents and subtracted the time-out of the web-survey page from the response times for SF:OFF respondents. Table 2 contains the results for processing times.[5]

Table 2. Average processing time (in seconds) of SF:OFF respondents and each outlier definition for single and grid questions

	Single Questions			Grid Questions		
Definition Procedure	Mean	Effect Size	*p* value	Mean	Effect Size	*p* value
SF:OFF	17.95			72.52		
SF:ON	9.17	.25	< .001	38.24	.80	< .001
SF:ON + A	7.58	> 1	< .001	35.88	> 1	< .001
SF:ON + B	6.67	> 1	< .001	33.87	> 1	< .001
SF:ON + C	6.61	> 1	< .001	33.82	> 1	< .001
SF:ON + D	6.80	> 1	< .001	34.46	> 1	< .001
SF:ON + E	7.47	> 1	< .001	36.40	> 1	< .001

Note. The comparisons (unpaired t-tests) were calculated between “SF:OFF” and “SF:ON” respondents and each outlier definition procedure (including “SF:ON”). We corrected response times of the SF:OFF respondents for their time-out and calculated Cohen’s d as effect size.

[5] We conducted all analyses with and without a log transformation of the response time data. All analyses remained unchanged and thus we report the untreated solution.

It can be seen that discontinuously processing respondents produce significantly higher processing times than continuously processing respondents; in some cases, they are even twice as high. This applies across all outlier definition procedures and both presentation modes. However, contrary to our expectations, the differences between single and grid questions are relatively small. We also calculated Cohen's *d* coefficient as a measure of effect size and observed values equal to or higher than $d = .80$ (except for single questions SF:ON); implying very strong effect sizes. Moreover, there are some differences with respect to the outlier definition procedures, which, in turn, implies that the less strictly a procedure defines the lower and upper thresholds, the higher the procedure-admitted processing time. Accordingly, procedure A and E admit the highest and procedure C the lowest processing times.

Item Non-Response

We postulated that respondents who left the web survey during completion provide lower response quality in terms of item non-response. Figure 3 illustrates the statistical results. It can be observed that, in line with our expectation, discontinuously processing respondents produce a significantly higher item non-response rate than continuously processing respondents, irrespective of the definition procedure A to E. This applies for single and grid questions. However, contrary to our expectation, this is much more pronounced for single than for grid questions. Interestingly, there are almost no differences in the amount of item non-responses across the outlier definition procedures, which applies to both presentation modes. Hence, in contrast to our previous findings on the number of outliers as well as the processing time, the results do not differ substantially across the outlier definition procedures.

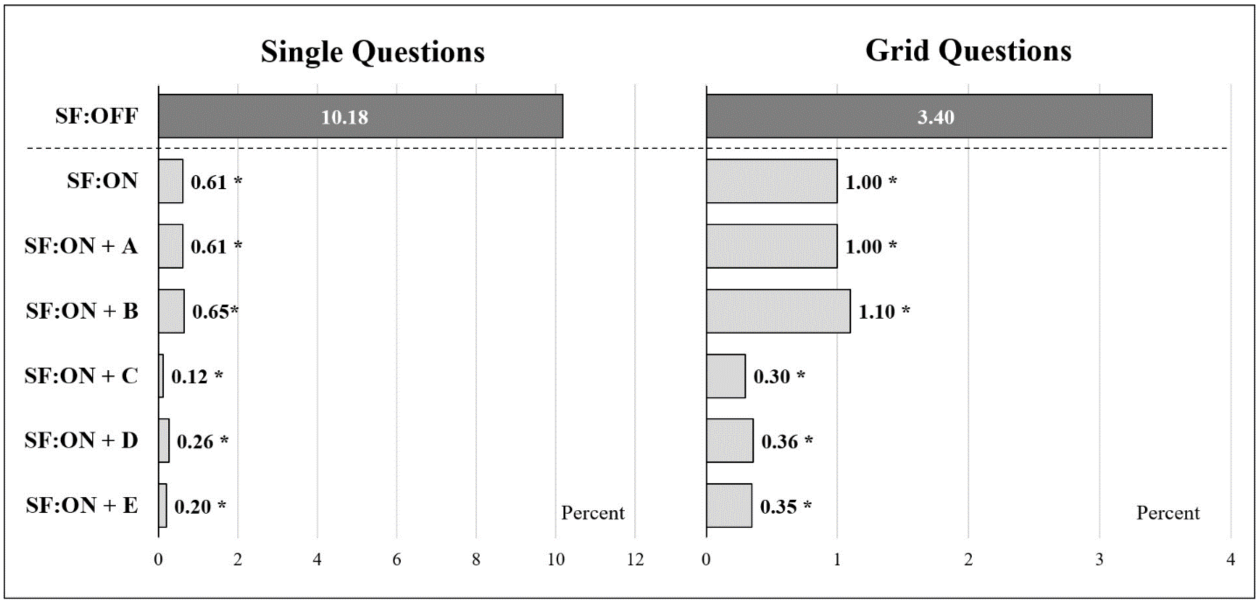

Figure 3. Proportion of item non-response for SF:OFF respondents and each outlier definition procedure for single and grid questions

Note. $^*p < .001$. We calculated chi-square tests between SF:OFF and SF:ON respondents and each outlier definition procedure (including SF:ON). For single questions, the expected frequencies were partially smaller than 5 so that we conducted Fisher's exact test to validate the results of the chi-square tests. All results remained unchanged.

Discussion and Conclusion

The goal of this article was to investigate the performance of different outlier definition procedures based on response time distributions by means of SF (Höhne, Schlosser, & Krebs, 2017). Our results indicate that all outlier definitions used have something in common: they are not capable of capturing all discontinuously processing respondents. This circumstance is more pronounced for grid than for single questions and might be attributable to their complexity. The extent of biased response times depends on the outlier definition procedure itself. Implying that the less strict the definition procedure is, the more SF:OFF respondents remain undetected. In particular, these procedures are unable to catch respondents with comparatively short than long time-outs. The strictness, however, also affects how many respondents will be excluded as outliers from the subsequent response time analyses. Hence, researchers are caught in a conflict between catching respondents with biased response times and reducing their case numbers when only employing outlier definition procedures based on response time distributions. As suggested by one of the anonymous reviewers, it would be desirable if future research takes a closer look at the reasons why respondents leave the web-survey page and how this corresponds to the acquisition of these respondents by outlier definitions in response time studies.

The analyses on processing times reveal significant differences for all comparisons. This supports our assumption that leaving the web-survey page causes an interruption so that respondents must re-orientate themselves and restart their response processes after returning. We can also observe that stricter definition procedures result in on average shorter response times than less strict ones. However, the average response time of SF:ON respondents without applying any definition procedure does not appear to be the optimal way because they are comparatively high. This implies that common outlier definitions are still necessary to deal with the "noise" in response time data due to different forms of multi-tasking and/or distraction (e.g., watching TV). Nevertheless, our findings indicate that response times of SF:OFF respondents are considerably biased upwards even if they are corrected and thus it seems reasonable to exclude them to achieve an appropriate analysis and interpretation of response times.

As suggested by another anonymous reviewer, it would be interesting if future studies combine SF with replacing strategies such as multiple imputation. Furthermore, they could also investigate the effects of outlier definitions on the power of experimental treatments.

Next, we investigated response quality between discontinuously and continuously processing respondents. We can observe significant differences for both presentation modes. Unexpectedly, item non-response rates are higher for single than for grid questions. One explanation could be that it is easier for respondents to decide to skip single instead of multiple questions in a grid. This is only a theoretical consideration that requires further research. Although not reported, SF:OFF respondents also selected more often middle response categories than SF:ON respondents indicating an error of central tendency.

If we compare the outlier definition procedures used, we observe that they produce quite similar results, except as it concerns the detection of discontinuously processing respondents. This fact is directly linked to the number of outliers defined by the procedure itself. Hence, it depends on the strictness of defining the lower and upper thresholds. While procedures A and E exclude only few or even no "very short" response times, procedure C very rigorously excludes many "short" ones. In consequence, these procedures seem to be somewhat dubious because they could harm the connection between response times and cognitive effort in opposite directions. Therefore, procedures B and D seem to be a good compromise because they produce reasonable thresholds. Procedure D, on the other hand, represents a more distribution-sensitive definition than procedure B since it uses the lower and upper quartile range as a measure of dispersion instead of the interquartile range (IQR). Hence, procedure D seems to be more appropriate.

The results of this study also provide new insights on the continuity of web-survey completion. For instance, about 13% of the respondents processed the web survey discontinuously and the average time of absence was about 14 seconds (see footnote 4). In our opinion, it would be worthwhile to use SF in future studies to investigate the effects of a continuous and discontinuous completion on response behavior of respondents.

Altogether, there are two limitations: first, our research is based on a student sample, which might hamper the external validity of the results. However, conducting research with students does not restrict the generalizability in general. Furthermore, our study is conceptualized as a usability study to introduce a new paradata tool that contributes to a more elaborated evaluation of response times. Particularly, the findings on the number of outliers (*hypothesis 1*) and differences in processing times (*hypothesis 2*) should not solely depend on the sample. In contrast, since it is based on university students presumably with above-average cognitive skills it is to expect that especially the differences in processing times would be more pronounced in a general population sample. Second, due to technical limitations combining JavaScript events, this study focuses on PCs and neglects mobile devices. These limitations, however, do not necessarily preclude the practical application of SF for mobile devices in response time outlier definitions (see footnote 2). Nevertheless, it would be desirable for future research to overcome these limitations and replicate this study for mobile devices.

Our findings show that the determination of outliers in response time analyses is an indispensable strategy for dealing with the "noise" in response time data. In other words, researchers must apply appropriate procedures to ensure their correctness. Although outlier definitions based on response time distributions produce relatively sufficient results, they are not exhaustive. The reason for this is that they are unable to capture all discontinuously processing respondents, which biases response times and affects the determination of the thresholds. Therefore, we encourage survey researchers to use SF for more sophisticated data cleaning. More precisely, we suggest they apply a two-step procedure: first, SF:OFF respondents are excluded on the basis of the web-survey page. Second, a common outlier definition procedure (preferentially D) for the remaining SF:ON respondents can be applied. Altogether, our findings suggest that the two-step outlier definition procedure seems to be superior: it excludes an appropriate number of cases and allows a more elegant as well as subtler calculation of thresholds for response time analyses.

References

Couper, M. P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. *Social Science Computer Review*, *18*(4), 384–396.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The Design of Grids in Web Surveys. *Social Science Computer Review*, *31*(3), 322–345.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method (3rd ed.)*. New York, NY: John Wiley & Sons.

Fazio, R. H. (1990). A Practical Guide to the Use of Response Latency in Social Psychological Research. In C. Hendrick, & M. S. Clark (Eds.), *Research Methods in Personality and Social Psychology* (Vol. 11, pp. 74–97). London, UK: Sage Publications.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21.

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.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. New York, NY: John Wiley & Sons.

Höhne, J. K., Schlosser, S., & Krebs, D. (2017). Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys. *Field Methods, 29*(4).

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology, 24*(7), 1003–1020.

Mayerl, J. (2013). Response Latency Measurement in Surveys. Detecting Strong Attitudes and Response Effects. *Survey Methods: Insights from the Field*. Retrieved on October 20, 2016 from http://surveyinsights.org/?p=1063

Olson, K., & Parkhurst, B. (2013). Collecting Paradata for Measurement Error Evaluations. In F. Kreuter (Ed.), *Improving Surveys with Paradata. Analytic Uses of Process Information* (pp. 43–72). Hoboken, NJ: John Wiley & Sons.

Schlosser, S. (2016). *Embedded Client Side Paradata (ECSP)*. Retrieved on October 27, 2016 from https://zenodo.org/record/55169

Schnell, R. (1994). *Graphisch gestützte Datenanalyse*. München: Oldenbourg.

Toninelli, D, & Revilla, M. (2016). "Is the Smartphone Participation Affecting the Web Survey Experience?". *Proceedings of the 48th Scientific Meeting of the Italian Statistical Society*. Salerno. ISBN: 9788861970618.