Can timestamp analyses show the bottlenecks in web surveys?[1]

Ioannis Andreadis

Introduction, Data and Methods

This paper analyses the response times recorded in June 2013 from the multi-country, multilingual *WageIndicator* [2] web-survey on work and wages. The national *WageIndicator* web-surveys are adapted to peculiarities in the country. The analysis includes data from Germany, the Netherlands, Brazil, Belgium, Great Britain, Portugal and Spain. For Belgium there are two datasets because the web survey is available in two languages: French and Dutch. The selection of the countries was based on the number of useful cases in each dataset. I have started with the dataset that has the maximum number of useful records (1706) and I have stopped with Spain because (79 cases) because I have considered that dataset with fewer cases are not adequate for the analysis that follows.

Response time (Andreadis, 2012; Couper and Kreuter 2013; Heerwegh, 2003; Heerwegh, 2004) belongs to a special type of data called "Paradata". Paradata do not describe the respondent's answers but the process of answering the questionnaire. Paradata can provide useful information on how the respondents have interacted with the web survey. In this paper, I analyse the relationship between the time spent on the survey and dropout, i.e. the outcome of the respondent's decision to abandon the web-survey.

I have used the time the survey has started the survey (paradata variable: created) the times the user has finished with the Wageindicator

- introductory questions (e.g. What is your employment status? Do you have one paid job or more? etc); paradata variable: intro
- occupation page (e.g. What is your occupation? Do you want to specify your occupation in greater detail? etc); paradata variable: occupation1
- page regarding main business activity of the organisation, number of people employed and percentage of female employees; paradata variable: firm1
- page regarding the locations where the organisation has businesses / plants; paradata variable: firm2
- education page (e.g. What year did you finish full-time education? What is the highest level of education you have attained? etc) paradata variable: education
- page regarding the kind of employment contract and the working hours; paradata variable: contract_workinghours
- first page regarding the wage (e.g. Do you receive your salary per MONTH? Do you know your GROSS and YOUR NET wage? etc); paradata variable:wage1

---

[1] Paper presented at the WageIndicator Conference, August 26-28 2013, Amsterdam The Netherlands. The Author would like to acknowledge the contribution of the COST Action IS1004 WebDataNet http://www.webdatanet.eu

[2] http://www.wageindicator.org

- second page regarding the wage (e.g. Did your last wage include any of these allowances?); paradata variable:wage2

If the response time is missing for a page of the WageIndicator web survey and it is not missing for the previous page, then we can define the corresponding case as a new dropout of the survey that has occurred while the user was on the page with the missing response time.[3]

By calculating the difference between two consecutive timestamps we can get a measure of the time spent on a page. In addition, by calculating the difference between a timestamp of a specific page and the time the survey has been started, we can have a measure of the overall time spent on the questionnaire until this page. The overall time is associated with the order of the page in the questionnaire (i.e. pages that appear near the end of the questionnaire will be associated with longer overall times than pages that appear near the beginning of the questionnaire).

Overall response time can be an indicator of burden to the respondent, i.e. respondents who need more time to arrive to a page have probably faced more difficulties in answering the previous questions and the probability of them getting tired can be larger. Galesic (2006) has shown that there is a link between the overall survey burden and dropout and she argues that there is a cumulative effect of all previous questions in a questionnaire on dropouts: "*But as the survey continues, the influence of negative aspects of participation (such as fatigue and boredom) becomes stronger and so does the preference to stop participating. However, until this change crosses their inhibitory threshold and affects behavior, respondents will continue to participate despite the growing tendency to stop. The preference to complete the questions may nevertheless decline, and that can result in lower quality of their answers*".

In addition, Yan and Tourangeau (2008) and Andreadis (2012) have found evidence that respondents tend to answer more quickly as they get closer to the end of the questionnaire. This could be a sign that respondents get tired/bored near the end of the survey and they dedicate less time (pay less attention) to the last questions.

Based on the aforementioned findings, I use the response times of the WageIndicator web survey to test the following hypotheses:

i) Larger overall response time near the beginning of the web questionnaire could be an indicator of respondents facing difficulties with the questionnaire. These respondents are expected to dropout more frequently than people who have spent less time on the same questions.

ii) Smaller overall response time near the end of the questionnaire could be a sign of tiredness and/or boredom. These respondents have probably started giving lower quality responses and they are more probable to dropout than people who have spent more time on the same questions.

[3] Dropouts are usually defined by the data itself instead of the using missing response times as I describe here. The former method (keeping track of the fist unanswered question) has the disadvantage that if answering the questions is not obligatory, the respondent may have abandoned the survey several pages after the last unanswered question.

Findings

Table 1 displays the dropout rates per page in each country. The rate of respondents who leave the survey before they finish the introductory questions ranges from 5% (Netherlands) to 15% (Portugal and Spain). This difference may have various reasons: Probably, Dutch visitors are more interested in the topic of the survey than the respondents of other countries. This could imply that Portugal and Spain teams should target to a more focused group of respondents. Time differences could also be attributed to differences between languages (for some languages it takes more time and more effort to ask the same thing). Finally, since national web-surveys have been adapted to peculiarities in each country, it is possible to have more introductory questions in one country, than in another.

From the group of respondents who have finished the introductory questions, a significant part abandons the survey before finishing the occupation page. The rate of new dropouts in the occupation page ranges from about one out of five (for most of the countries) to about one out of three (in Brazil and in Portugal). Given that both the latter surveys are in the Portuguese language, this could mean that the Portuguese occupation tree should be checked again. These two surveys continue to suffer by many new dropouts in the following page. The dropout rates are low for the following pages until the wage1 page.

Table 1. Dropout rates per page in each country/language

	N*	Intro	Occupation	Firm1	Firm2	Education	Contract - working hours	Wage1	Wage2
DE	1706	8%	23%	12%	3%	5%	3%	10%	5%
NL	1645	5%	20%	9%	5%	10%	3%	8%	6%
BR	999	9%	31%	22%	11%	10%	5%	15%	13%
fr_BE	198	13%	22%	13%	7%	9%	1%	24%	20%
nl_BE	179	10%	19%	5%	6%	9%	1%	24%	20%
GB	178	10%	20%	11%	8%	7%	4%	11%	5%
PT	116	15%	34%	25%	6%	9%	7%	38%	17%
ES	79	15%	18%	4%	8%	12%	5%	15%	0%

Source: WageIndicator (June 2013). The presented dropout rate is the ratio of the number of respondents who have not completed the current page (i.e. they have a missing timestamp for the current page) divided by the number of respondents who have completed the previous page (the timestamp for the previous page is not missing).

* The presented N refers to the number of respondents at the beginning of the survey

Table 2 Median overall time spent on the survey until the page

	N*	Intro	Occupation	Firm1	Firm2	Education	Contract - working hours	Wage1
GB	1706	21	159	222,5	254	348	446	605
DE	1645	46	166	268	296	374	455	544
NL	999	21	154	240	262	388	470	582
BR	198	38	250	348	386	524	653	789
fr_BE	179	28	184	268	288	408	489	633
nl_BE	178	22	142	202	228,5	309	381	530,5
PT	116	35	162	236,5	284	360	455	585,5
ES	79	26	133	189	222	327	411	522

* The presented N refers to the number of respondents at the beginning of the survey. The median is calculated among the respondents who have finished the corresponding page.

Table 2 shows the median overall time spent on the survey until finishing the corresponding web survey page (as it appears in the first row of the table). The average value is not the most suitable measure of central tendency for response times because there are cases with extremely large values. Response times are generally right skewed and the average value is sensitive to outliers. Therefore, as a measure of central tendency, I use the median value which is robust to extreme values.

Diagram 1 displays the dropout rates on occupation page and overall median time spent on the introduction questions for each WageIndicator survey I have included into the analysis. There is a positive relationship between time spent on introductory questions and dropout rate. The OLS model (R^2=0.355) shows that for every 10 additional seconds spent, dropout rate increases by 3.9%

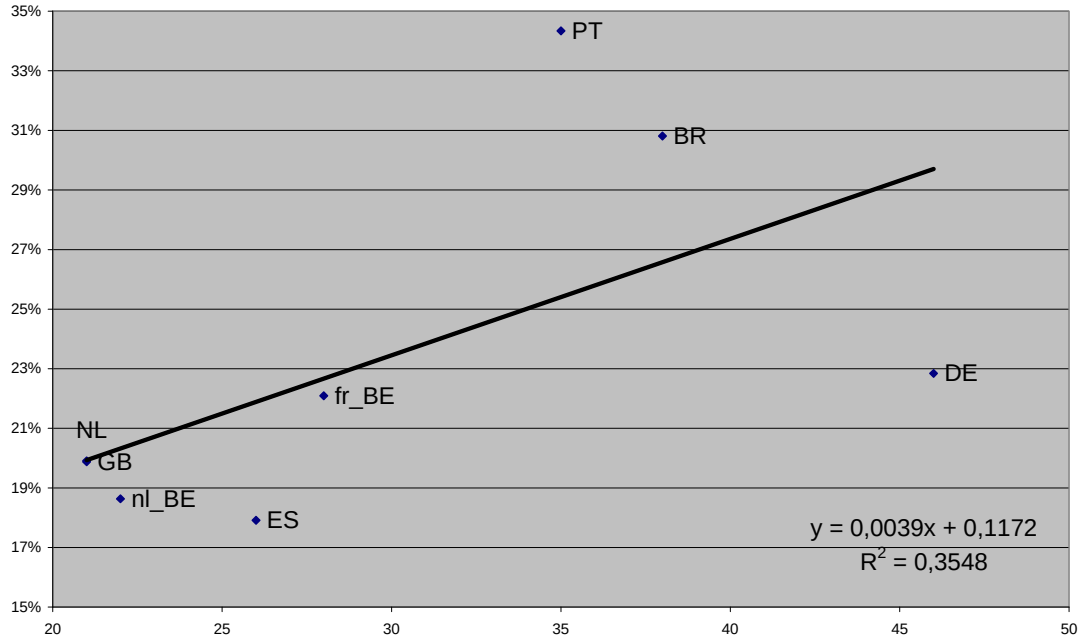

Diagram 1 Dropout rates on occupation page and overall median time spent on the introduction questions

Diagram 2 displays the dropout rates on firm1 page and overall median time spent until finishing the occupation page. Similar to Diagram 1, this diagram also shows a positive relationship between overall time spent the previous page and dropout rate on current page. The OLS model shows a stronger correlation than before (R^2=0.455).

Portugal is a significant outlier in both diagrams. Even after accounting for the overall time spent on the survey, the dropout rate of Portugal remains significantly larger than the expected value according to the least squares model. This means that we cannot attribute the significant higher dropout rates from the Portuguese survey only to response times. Of course, the Portuguese sample is rather small (116 valid cases) and the preliminary findings presented here should be verified with larger datasets. In any case, Portuguese survey should be tested more thoroughly with subsequent analyses.

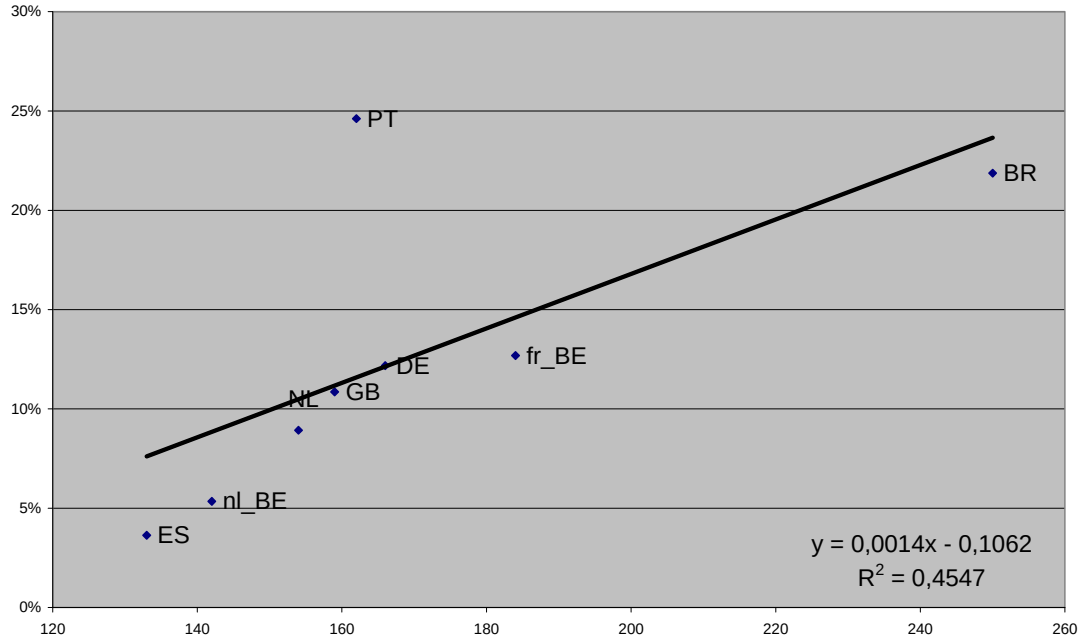

Diagram 2 Dropout rates on firm1 page and overall median time until the occupation page

Table 3 shows the linear regression coefficients and the values of R^2 of the models with dependent variable the dropout rate and independent variable the overall median time spent until the previous page. It becomes obvious that after the first two models that have been discussed above, the correlation becomes significantly weaker in the third model and after that it is diminishes to values very close to zero. Even the sign of the relationship in the last four models is not clear, as two of the models have negative and the other two models have positive beta coefficients.

Table 3 Linear Regression Coefficients and values of R^2

Regression for page:	Beta	R-squared
Occupation	0.00390	0.355
Firm1	0.00140	0.455
Firm2	0.00020	0.192
Education	-0.00007	0.031
Contract - working hours	0.00004	0.015
Wage1	-0.00010	0.014
Wage2	0.00020	0.069

Diagram 3 shows a comparison of time spent until the previous page between respondents who continue and respondent who dropout in GB WageIndicator survey. It seems that respondents who dropout either on the occupation or the firm1 pages have spent more time on the survey than those who continue. On the other hand, respondents who dropout near the end of the survey, have spent less time on the survey than those who continue.

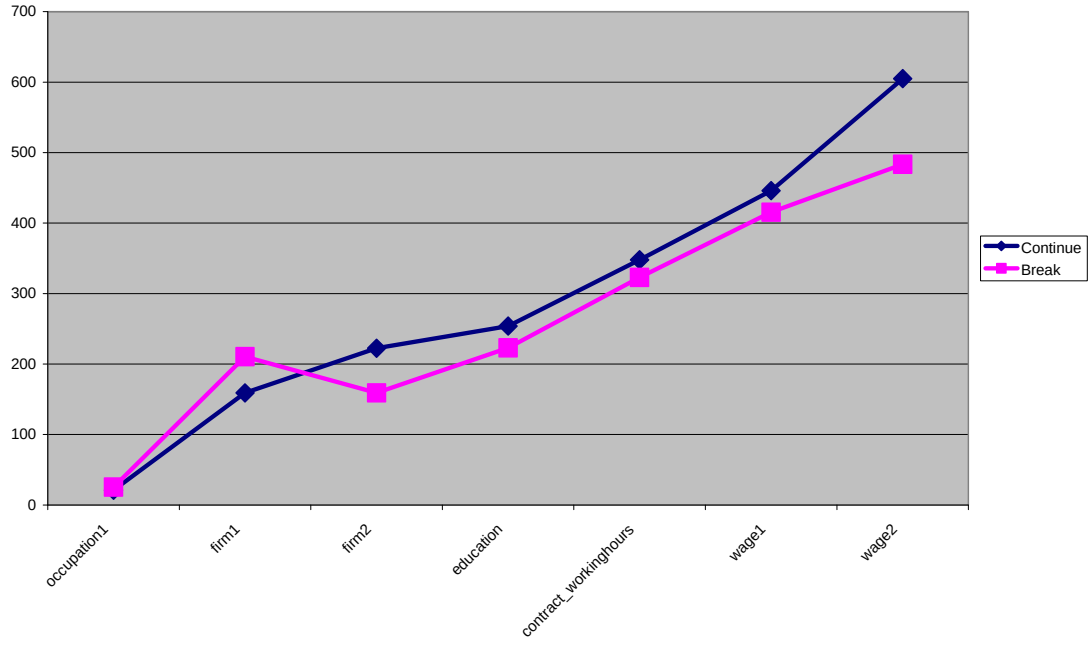

Diagram 3 Comparison of time spent until the previous page between respondents who continue and respondent who dropout in GB

References

Andreadis, I. (2012) To Clean or not to Clean? Improving the Quality of VAA Data *XXII IPSA World Congress*: http://www.polres.gr/en/sites/default/files/IPSA12.pdf

Couper, M.P. and Kreuter, F. (2013) Using paradata to explore item level response times in surveys *Journal of the Royal Statistical Society*: Series A (Statistics in Society) Volume 176, Issue 1, pages 271–286

Galesic, M. (2006). Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey. *Journal of Official Statistics*, 22 (2), 313–328.

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)

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