UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 36

2013

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici ESCI-Born - Despatxos 19.517-19.527
Passeig Pujades, 1 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

What are the links in a web survey between response time, quality and auto-evaluation of the efforts done?

Melanie Revilla

RECSM, Universitat Pompeu Fabra

Carlos Ochoa

Netquest

Abstract:

Evaluating the quality of the data is a key preoccupation for researchers to be confident in their results. When web surveys are used, it seems even more crucial since the researchers have less control on the data collection process. However, they also have the possibility to collect some paradata that may help them evaluating the quality. Using this paradata, it was noticed that some respondents of web panels are spending much less time than expected to complete the surveys. This creates worries about the quality of the data obtained. Nevertheless, not much is known about the link between response times (RT) and quality. Therefore, the goal of this study is to look at the link between the RT of the respondents in an online survey and other more usual indicators of quality often used in the literature: absence of straight-lining, properly following an Instructional Manipulation Check (IMC), coherence and precision of answers, etc. Besides, we are also interested in the link of both RT and the "usual" quality indicators with the auto-evaluation of the respondents about the efforts they did to answer the survey. Using a SEM approach which allows separating the structural and the measurement models and controlling for potential spurious effects, we find a significant relationship between RT and quality in the three countries studied. We also find a

significant, but lower, relationship between RT and autoevaluation. However, we do not find a significant link between autoevaluation and quality.

Keywords:

Web surveys, Quality, Paradata, Response Time, Auto-evaluation, Efforts, Structural Equation Modeling

Acknowledgements:

We are very grateful to Netquest for providing us with the necessary data for this paper and to Willem Saris for his very useful comments on previous draft of this paper.

I) Introduction

Web surveys offer both new challenges and new opportunities. On the one hand, the absence of interviewers makes the control of the situation and of what the respondents are doing more difficult. Respondents can answer for another person, they can do several tasks at the same time, they can complete the survey in any place, at any time, with or without the presence of others persons, they can randomly answer the questions without even reading them. There is nobody to see it and stop it.

On the other hand, web surveys offer rich possibilities for collection of paradata, meaning data generated automatically by the data collection process that can be used as additional information to describe or evaluate this process (Couper, 2000). Even if nobody is there to observe what the respondents are doing, there are possibilities to record many kinds of information that can help the researchers identifying the respondents' behaviors during the survey completion. For instance, it is possible to track the movements of respondents' eyes, in order to see if respondents looked at all the information on one page. It is also possible to record the movements of the mouse, which can indicate hesitation or direct answering, but also to record all the clicks the respondents did, which can allow seeing if respondents change their answer, or go back in the survey. Moreover, it is possible to detect if the respondents went out of the survey to open another Internet window, which indicates if the respondents did some other tasks on Internet.

We could go on with the list, but we want to focus on one specific kind of paradata: the response times (RT) or response latencies. Response times are defined as the times between the moment when the page starts loading and the moment when respondents click on the "Next" button.

Indeed, even if there is a lot of paradata available, a lot of them are difficult to analyze in a systematic way that could be used to learn more about the response process. RT seems to be one of the most accessible ones.

Moreover, in practice, it was noticed that some respondents of web panels are spending much less time than expected completing the surveys. This fast responding, also referred to as speeding, creates worries about the quality of the data collected through these web panels. Of course, by answering regularly surveys, panelists get practice, so we can expect them to be quicker than non-panelists. However, part of them are spending such little time completing the surveys that it is nearly impossible that they have read the questions, even more that they have thought carefully about the answers before selecting them. Consequently, some survey institutes commonly apply as a quality rule the exclusion from the final dataset of the panelists that did the survey in less than *x*% of the expected time of completion.

The problem is that different respondents need different times to properly answer the questions. Some may be much quicker in reading, making up their mind and finding the adequate answers than others. Yan and Tourangeau (2008) find that age and education, as well as Internet skills, are affecting RT.

It is clear when RT are extremely short that it is simply impossible that the respondents read the questions before answering. However, when RT are quick but not so extreme, we do not know anymore if this is because the respondents are highly skilled and therefore they are able to answer quickly while going through all the necessary steps or if the respondents are just rushing through the questionnaire without reading and thinking properly.

Even if the idea that speeding is an indicator of low quality is quite common, there is little research about how RT and quality relate with each other. Malhotra (2008) looks

at the link between completion time and primacy effects (i.e. bias toward selecting earlier response choices). He finds that for low educated respondents, shorter RT are associated with higher levels of primacy effects. Zhang (2013) considers the link between RT and another quite often used indicator of quality: straight-lining, also referred to as non-differentiation. Straight-lining consists in selecting the same answer for all the items in a battery. The author finds that the respondents that speed more also tend to straight-line on more grid questions, suggesting that the tendency to speed is indeed related to the quality of answers.

Following this line of research, we want to go further in the investigation of the link between RT and quality, by considering more indicators of quality and adding also the auto-evaluation of the respondents about the efforts they made to answer the survey.

More precisely we want to investigate:

1) How strong is the link between RT and quality?
2) How strong is the link between quality and auto-evaluation of efforts done?
3) How strong is the link between auto-evaluation of efforts and RT?

We hope the answers to these three questions will give us some indication to start answering a very general and practical question: can we use paradata to improve the overall quality of online surveys and how? More specifically, we focus on one kind of paradata, RT, so we can wonder: can we use RT as a proxy for quality? Said differently, can we use RT to decide which respondents we should exclude from our analyses? Besides paradata, can we use auto-evaluation of the respondents for the same goals?

The next section gives some details about the data used to study these questions. Then, the structural and measurement models are presented, and combined to propose a complete Structural Equation Model (SEM). After that, we explain how the analyses

were done and how we corrected the initial model in order to get to the results, which are finally presented and discussed.

II) The data: a survey from Netquest in Spain, Mexico and Colombia

We used data from the web panel Netquest (www.netquest.com), which is accredited with the ISO 26362 quality standard, specific for online access panels. Netquest uses a database of users of many websites that agreed to receive emails from one of these websites. From this database, they invite people with the profile they need to participate in their panel. For each survey completed, panelists get points that they can exchange for gifts. The number of points is proportional to the expected length of the survey, but not the actual completion time of each respondent. Speeding can therefore appear as advantageous to the panelists since it allows them to get the same amount of points in less time: if respondents participation is mainly drawn by the incentives, "the purely rational approach is to satisfice" (Malhotra, 2008), meaning to "minimize effort in responding to surveys and simply provide the appearance of compliance (Krosnick and Alwin 1987; Krosnick 1991)."

The survey selected has been proposed to panelists in Spain, Mexico and Colombia between the 14th of May and the 18th of June 2013. One advantage is that it includes questions that can be affected by different kinds of satisficing behaviors. Another is that it includes an auto-evaluation of the efforts made and of the difficulty of the questions.

The survey was expected to take around 25 to 30 minutes to be completed (around 125 questions) and was about various topics: mainly drinks and food consumption, brands of cars and media use.

Quotas for age and gender were used to get samples' distributions similar on these variables to the population distributions. In each country, around 1000 respondents

completed the survey. However, an experimental design was used such that each sample was randomly split-up into three groups. For this study, we focus only on the first split-ballot group of each country, in order to have similar questionnaires for which we can compute the same quality indicators. At the end, we have 345 respondents in Spain, 305 in Mexico and 336 in Colombia.

III) The structural model

Our main interest is investigating what are the links between the RT, the quality of the answers and the auto-evaluation of the efforts made. Our main hypotheses are:

H1: a worse quality of the answers of the respondents is directly related with shorter RT (i.e. more speeding)

H2: a worse quality of the answers of the respondents is directly related with less reported efforts

H3: more reported efforts go together with longer RT.

In order to estimate these relationships properly, we need to introduce as control variables the ones that we expect will create spurious relationships between our main variables of interest. These variables are first age and education: indeed, we expect them to affect RT (e.g. following the results of Yan and Tourangeau, 2008) and to affect the quality of answers (e.g. following the results of Alwin and Krosnick, 1991). Then, we also include an auto-evaluation of the difficulty of the questions as control variable: if the questions are easier, the quality should be better and the RT shorter.

IV) The measurement model

RT, quality and auto-evaluation are theoretical concepts that are not directly observed. The indicators used to operationalize them are presented in this section.

Because there are errors, the indicators are never perfect measures of the concepts of interest, but by using several indicators for each latent concept, we can correct for measurement errors. This is the approach followed in this paper.

1) Indicators of RT

RT are paradata. As explained for instance by Yan and Tourangeau (2008, pp.53), they "can be collected or from the server or from the respondents' computers. Server-side response times show the elapsed time from the moment the server delivers a survey question to a respondent's computer to the moment when it receives an answer from the respondent. By comparison, client-side response times include the elapsed time from when a survey question is fully displayed on respondent's computer to when an answer is sent." In our study, we use server-side response times (the only ones we could get access to), which therefore include the downloading time. By consequence, if some respondents use quicker Internet connections than others, we may observe differences in RT that do not reflect really differences in the time spent to process the question, decide and select an answer.

Besides the downloading time, another issue of the measure of RT is that these times correspond to the time spent on a page. However, if a respondent let the survey on and start doing other activities (e.g. speak with another person, go to the kitchen to check the diner, go to the bathroom, etc), the time spent on the page will not be a good measure of the time spent to answer the question. It will be much longer. Detecting the multitasking behaviors is quite difficult. Nevertheless, in the case of very long RT for one page, we can be pretty sure that respondents have interrupted the survey process to do another task and came back later.

Therefore, in order to compute the RT of each respondent, for each page, we substitute the times of the 1% respondents with the highest time[i] (considered as the ones that clearly were multitasking) by the average time spent by the other 99% to answer to the question(s) on that same page. We substitute by the average time and not the maximum time of the other 99% because we do believe that the very long times do not indicate extremely slow respondents but respondents that interrupted the survey[ii]. Then, there is no reason to expect these respondents to spend a long time on the page once they come back to it.

In order to correct for measurement errors in our final model, instead of using only one indicator or RT, we use three. Each one corresponds to the average RT for a set of questions from the surveys. Together, the three sets of questions constitute the complete survey.

So our three indicators of RT (corrected for very long times) are, for $i = 1,2,3$:

$RT_i = RT_{\text{i-th set of questions}} = \text{Total Time}_{\text{i-th set of questions}} / \text{number of pages}_{\text{i-th set of questions}}$

In Spain, the mean RT are respectively: .23, .17 and .18 minutes. In Mexico, there are respectively: .26, .20 and .20 minutes, and in Colombia, .29, .22, and .23 minutes. Part of the differences could come from differences in average connections' speed across countries.

We have to note that in general, there is only one question per page, but when questions are part of a battery, there are several questions on the same page. Therefore, when there are batteries of questions, we can expect the RT to be higher.

2) Indicators of quality

In order to measure the quality of the responses, we use different traditional indicators and some a bit more specific but all are directly derived from the answers to

the questionnaire ("direct" data and not paradata anymore). We will now go through each of them.

2.1 Instructional Manipulation Check

First, we use an Instructional Manipulation Check (IMC), which "consists of a question embedded within the experimental materials (…) that asks participants (…) to provide a confirmation that they have read the instruction" and is supposed to measure "whether or not participants are reading the instructions and thus provides an indirect measure of satisficing" (Oppenheimer, Meyvis and Davidenko 2009, pp.867).

In our study, the IMC was included within a grid about media where respondents had to select the five most important options out of a list of 18, and this in three different cases. The IMC consisted in one additional row in this grid asking the respondents, if they were reading, to mark all three buttons on this row, besides the five most important options. This additional row was placed in the middle of the substantive alternatives, always on the same position.

The variable summarizing the results is called "passIMC". It is a dummy variable which takes the value 1 if the respondents correctly achieve the manipulation asked (i.e. checked the 3 boxes) and the value 0 otherwise.

In our sample, only 21.67% of the respondents did it correctly: 30.0% in Spain, 21.9% in Mexico and 14.1% in Colombia. There are clear differences across countries but overall, these are very low percentages which can make us worry about the quality of the data. But we should notice that this IMC was part of a really complex and badly designed grid. This grid was actually used in this survey because it was such a badly designed one, so it was part of an experiment aiming to improve it. Therefore, even respondents that are not by nature "bad" can be tempted in this context to take short-

cuts. It is also possible that some respondents did not understand exactly what they had to do in a context of such a complex grid.

2.2 Number of selected items in the media grid

The same grid was used also to compute a second quality indicator, which is the selected number of items in the three questions of this grid. As mentioned earlier, respondents should choose the five most important options in each of the three questions, apart from the IMC boxes.

In such cases, web surveys allows to use an automatic check to be sure that respondents comply with the requirement of choosing five items in each case, and not less nor more. However, since the question was part of an experiment to improve it, the automatic check was not applied in that case, in order to see how many respondents will not at first be able to carry out the task according to the instructions.

The information is summarized in a variable called “select5”. For each respondent, it scores from 0 (did not select five options in any of the three questions) to 3 (selected five options in the three questions). Table 1 gives the distribution of this variable in the different countries.

Table 1 here

Table 1 shows quite some variations across countries but overall, the percentage of respondents that really respected the instructions is only between 32 and 37%.

2.3 Straight-lining

The next indicator is straight-lining or non-differentiation between items. It occurs usually in grids with many items. This is one of the most used indicators for satisficing in web surveys.

The survey included two grids of 16 items about the frequency of consumption of several drinks (questions B1-B16 and HB1-HB16) and one of 12 items about opinions about brands of cars (questions HC1-HC12).

For each grid, we look if the respondents selected the same answer category for all items. If they do, then they are considered as pure straight-liner for this grid. Combining the scores for the three grids, we create the variable "nbstraight", which counts in how many of the grids the respondents are pure straight-liners.

Table 2 here

Table 2 shows again quite some differences across countries, with Colombia performing better. Overall, pure straight-lining is not as present as the previous undesirable behaviors. However, the questions for the two grids about drink consumptions were not very demanding since they were about central behaviors. Also, we only counted the pure straight-liners and not the ones that selected for instance 15 times the same answer category but changed for one of them. Finally, answering randomly instead of linearly to a grid of questions does not require really more efforts so straight-lining is not the only possible strategy to satisfice when answering a grid.

2.4 Incoherence of responses between repeated or opposite items

Then, we consider as an indicator of quality the incoherence across responses for repeated questions or opposite items. Indeed, the two grids about drink consumptions mentioned before are similar except that the scale is reversed. One is asked at the beginning of the questionnaire, the other at the end. The drink consumption of respondents cannot have changed in between the two sets of questions, so differences can be interpreted as incoherence in answers. Moreover, we have three pairs of opposite items (with one item worded positively and one negatively), such that if respondents

agree with one, they should disagree with the other one if they are coherent in their answers. For example: "[name] is a trustworthy brand" and "[name] is a brand in which I have no trust".

For each of these questions, we compute how incoherent the responses are in the following way: if the respondents selected twice the same category, it is not incoherent at all, so they got a score of 0 for the corresponding question. If the second answer is one category next to the expected answer according to the first question, there is a small incoherence and they get a score of .01 for this question. If the second answer is *n* categories next to the expected answer according to the first question, there is an incoherence of level *n* and they get a score of .01**n* for this question. Finally, we sum up all these scores in the variable "incoherence" (from 0 to 1.08).

The percentage of respondents that answered in a perfectly coherent way (i.e. "incoherence"=0) is only 3.6% overall, with big differences again between countries: in Spain it is much higher (9.9%) than in Mexico and Colombia (both 0.3%).

The percentages of respondents whose answers varied in average no more than one category (i.e. "incoherence"≤.19) are 86.6% overall, 91.9% in Spain, 87.2% in Mexico and 80.6% in Colombia. Spain is still performing better but now Mexico is doing better than Colombia.

2.5 Precision of answers in open narrative questions

In the case of open narrative questions, others indicators of quality are available. First, the precision of the answers can be estimated via the number of characters the respondents wrote. Respondents that do not want to make efforts will indeed tend to write less.

In our study, two open narrative questions are present. The first one asks respondents what they would do to improve the survey and the second one asks them to indicate all the topics they can remember from the survey.

The number of characters written in both questions is summed up to create the variable "charac-narrative". All countries together, the number of characters written varies from 2 to 621, with an average of 120. In Spain, it goes from 2 to 613, with an average of 111. In Mexico, it goes from 6 to 621, with an average of 125. Finally, in Colombia, it goes from 8 to 502, with an average of 125.

2.6 Non-sense in open questions

In open narrative questions, in order to reduce the efforts, besides writing short answers, respondents can also write "easy" answers: just a few letters or signs that have no sense, some unrelated text, that does not answer the question or simply "don't know".

We create a variable "nonsense" that counts how many of these undesirable answers we get (0, 1 or 2 since we have two narrative questions). Table 3 gives the percentages.

Table 3 here

Table 3 shows that the percentages of non-sense are relatively low. Overall, almost 94% of the respondents did not write any. Still, a small percentage did and when they did, it indicates a bad quality of the answers.

3) Indicators of the auto-evaluation of the efforts done

The previous indicators of quality are based on the idea that if respondents are doing the necessary efforts, the quality will increase, whereas if they satisfice, the quality will decrease. Therefore, the efforts done seem to be a key variable. But who knows better

the efforts they did than the respondents themselves? Why couldn't we ask directly the respondents if they did or not the efforts needed to answer properly?

In our study, such a question was asked: "how much efforts did you put in answering this survey?", on a scale from "0 - Minimum effort possible" to "10 – Maximum effort possible".

Even in an online survey, this question is susceptible of social desirability bias, i.e. over-reporting of socially desirable behaviors (here having performed the maximum efforts) and under-reporting of the undesirable ones (not having done any effort). The mean for this variable is 6.6 in Spain and Colombia and 6.8 in Mexico. This is quite high, which tend to confirm that indeed there is some social desirability bias.

Besides, some respondents may consider that they do not need to make effort to answer properly because answering questions is an easy task for them. Also, if respondents really answer without reading the questions carefully, it will apply to this question too, such that their observed answers might not be a good measure of the real efforts.

However, this is what we want to find out by including it in our model. If the efforts reported are highly correlated with the quality measured by the indicators specified in the previous section, then the easiest way of assessing response quality would be to ask directly to the respondents.

In order to correct for measurement errors, since we have only one available indicator in this case for our latent concept, we need to get an estimate of the quality of the question from an outside source (Saris and Gallhofer, 2007). We use the program SQP 2.0 (Saris et al, 2011) to predict the quality of this question (available for free at http://www.sqp.nl/) and get a value of .538. We use the same value in all three countries since they share the same language.

4) Indicators for the variables of control

For age, we use the answers to a direct question asking the age of the respondents.

For education, different answers categories are used in the three countries. But Netquest also provides a harmonized variable that is ordered from no or low education to high education (six categories). We use this harmonized variable.

For the last control variable, "easy", we use the question: "How do you feel about the questions of this survey?", on a scale from "0 – extremely difficult to answer" to "10 – extremely easy to answer".

Since we have only one indicator for each one, we need estimates of quality of the questions in order to correct for measurement errors. For the variable "easy", using the program SQP 2.0, we obtain a quality estimate of .613.

In the case of age and education, since these are background variables, we cannot use SQP. Instead, we use the estimates provided by Alwin (2007, p157): .997 for age and .884 for education. These estimates have limits, since there are not based on data from the countries of interest and since the education variable is not measured in the same way in our study than in the ones of Alwin. However, this is the best we can get for these variables. Besides, for age, the size of the errors is so small that in any case it will not change the results. But for education, it is much more realistic to use Alwin's estimates than assume that there are not measurement errors.

V) General Structural Equation Model

1) Initial model

A path diagram of the complete initial model can be found in Figure 1. It is a combination of the structural model discussed in section III and the measurement model discussed in section IV.

Figure 1 here

The control variables are treated as observed ones in the model, in order to simplify it. Then, in order to correct them also for measurement errors, we use the reduction of variance technique (putting the quality on the diagonal of the correlation matrix) in the case of these three variables. For quality and RT, the correction is done by using several indicators. For the auto-evaluation of the efforts, the loading between the latent variable and the observed one is fixed to the quality coefficient, which is the square-root of the quality prediction from SQP mentioned before, i.e. √.538=.734. The error variance is fixed for this indicator to 1-.538=.462.

2) Analyses and corrections of the model

The Maximum Likelihood estimation of LISREL (Jöreskog and Sörbom, 1991) for multiple group analyses is used to get the results. Each country constitutes a different group. We first specify all parameters invariant across countries. Then, the ones that are misspecified are allowed to be free.

The testing is done using both global fit measures (Chi-squared, RMSEA and CFI) and local fit measures (looking at the Expected Parameter Changes, Modification Indices and Power) using the program JRule (Van der Veld, Saris and Satorra, 2009) based on the procedure developed by Saris, Satorra and Van der Veld (2009).

First, we have to add correlations between some error terms (in one or more countries): between “select5” and “passIMC”, between “nonsense” and “charac-narrative”, between “incoherence” and “nbstraight”, between “passIMC” and RT_1 and

between "charac-narrative" and RT_2. The first three can be expected because pairs of indicators are based on the same questions, so it is logical that they correlate particularly. The two last also make sense because the IMC is only included in the set of questions used to compute RT_1 and both open questions are part of the set of questions used to compute RT_2.

Second, cross-countries differences are suggested for some correlations between control variables and some spurious effects. It seems acceptable to think that indeed education and age, for instance, correlate differently in the different countries. It seems also reasonable that these control variables can affect differently RT or quality. For example, similar levels of education may in practice correspond to different levels of knowledge in the different countries, which can explain different sizes of the effects. Therefore, we free in the adequate countries the spurious effects and correlations across control variables that were indicated as misspecified by JRule.

3) Results

By correcting the model as just indicated, we get an acceptable fit, both according to the global fit measures ($Chi^2(220) = 302.37$; RMSEA=.032; CFI=.96) and to local fit indicators (JRule did not suggest any big misspecification anymore). The final LISREL input is provided in Appendix. Table 4 gives the completely standardized estimates for the general structural equation model, as well as the unstandardized ones and the corresponding t-values.

Table 4 here

Table 4 contains a lot of information. But we focus on the main results. About the measurement part of the model first, we can see that the estimates are similar across the three countries. Even if they are all significantly different from zero, none of the

indicators of Quality are very good, since all loadings are quite low. This may be because here the error terms include not only the measurement errors but also unique components.

Nevertheless, if we would have to select fewer indicators of quality, the straight-lining seem to be the strongest one (standardized loadings between .47 and .55), followed by "incoherence" (standardized loadings between .41 and .48). On the contrary, the IMC, that is used in quite some companies as the unique measure of quality (respondents that fail this IMC are often immediately excluded from the final sample), has loadings of around -.24 to -.20 only. Using the IMC to exclude "bad respondents" seems therefore not very appropriate. However, we should mention that this may be due to the specific IMC used (in a very complex grid, such that most respondents failed it). Another choice of IMC may have lead to a higher loading.

Now, about the structural part of the model, we can notice that the three main relationships of interest are similar across countries. The standardized correlation between quality and RT is between -.22 and -.27 depending on the country (significant). The one of quality and auto-evaluation of the amount of efforts done is around -.05 (not significant) and the one of auto-evaluation and RT is around .15 (significant).

What is changing is the size and sometimes even the signs of the spurious effects. But in fact most of the effects of the control variables are not significantly different from zero. They are less spurious effects than we expected. The non-significance can be linked to the relatively small sample size (around 330 in each country). But overall the results suggest that the spurious effects created by "easy" and education are relatively limited. For age, more significant effects are found.

4) Discussion

Our main interest was to investigate the links between RT, quality and auto-evaluation. By developing a complete SEM model with a measurement and a structural part, we were able to estimate the size of the different relationships, controlling for possible spurious effects and correcting for measurement errors. Based on the results, we can answer our three introductory questions.

First, is there a link between RT and quality? Yes, there is. How strong is this link? The size is around -.25. This means that our first hypothesis is confirmed: a worse quality of the answers of the respondents is directly related with shorter RT, i.e. in more speeding.

Second, what about the link between quality and auto-evaluation? The analyses suggest that there is no significant relationship between quality and auto-evaluation of the efforts. A worse quality does not relate directly with less reported efforts. We do not find support for our second hypothesis. This can be due to social desirability bias, which pushed even bad respondents not to report the little efforts they did.

Third, how strong is the link between auto-evaluation and RT? Table 4 gives a standardized effect of around .15, statistically significant at the 5% level. This supports our third hypothesis: more reported efforts go together with longer RT. It may be because more reported efforts imply more real efforts (even if the relationship is not perfect) and more efforts lead to the respondents taking more time for answering. It can also be because if respondents take more time, they will give better quality answer. However, the size of the effect is small.

All in all, which practical advices can we derive from these results? Can we use paradata, and in particular RT, to improve the overall quality of online surveys? Can we use auto-evaluation of the respondents for the same goal?

The relationships with auto-evaluation are really small or even not significant. Whatever the reasons, it does not seem to be possible to use the auto-evaluation of efforts as a proxy of quality and therefore also not to decide which respondents did a "serious job" and which not based on this variable. More research would be needed to check if by dealing with some of the limits our analyses are facing it is possible to get stronger relationships, but based on Table 4, we have to conclude that we should not use the auto-evaluation of the efforts to decide which respondents to exclude.

Concerning RT, our results suggest there is a stronger relationship. Nevertheless, the standardized coefficients of around -.25 are not enough to conclude that we can use RT as a proxy of quality. RT can be used, as it is already done in practice by many companies, to exclude some clear speeders. But RT is very imperfectly linked with quality, so we should not give them too much importance by themselves. However, they may be used, in combination with a few other indicators (for instance incoherence and straight-lining, that appear to be the indicators with the highest loadings) in order to identify a set of respondents that overall give bad quality answers. Further research in that direction would be needed.

Biosketches:

Melanie Revilla (contact author: melanie.revilla@upf.edu) is a postdoctoral researcher at the Research and Expertise Centre for Survey Methodology (RECSM) and an associate professor at Universitat Pompeu Fabra (UPF, Barcelona, Spain).

Carlos Ochoa is Director of Global Operations at Netquest. He has an engineer degree in telecommunications and has been related to departments of technology services at companies like Sony and Auna.

References

Alwin, D.F. and J.A: Krosnick (1991). The Reliability of Survey Attitude Measurement. *Sociological Methods and Research* 20(1):139-81.

Couper, M.P. (2000). "Web surveys: A review of issues and approaches". *Public Opinion Quarterly*, 64: 464–494. http://www.jstor.org/stable/3078739

Jöreskog, K.G. and D. Sörbom (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.

Krosnick, J.A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys". *Applied Cognitive Psychology* 5:213-36. doi: 10.1002/acp.2350050305

Krosnick, J. A., and D.F. Alwin (1987). An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51:201-219.

Oppenheimer, D.M., Meyvis, T. and N. Davidenko (2009). "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of Experimental Social Psychology* (2009) 45:867–872. doi:10.1016/j.jesp.2009.03.009

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly, 72*(5), 914-934.

Saris, W.E. and I. Gallhofer (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York:Wiley

Saris, W.E, Satorra, A. and W.M. Van der Veld (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural equation modeling: A multidisciplinary Journal,* 16(4):561-582

Saris W.E, D.Oberski, M.Revilla, D.Zavalla, L.Lilleoja, I.Gallhofer and T.Grüner (2011). "The development of the Program SQP 2.0 for the prediction of the quality of survey questions". *RECSM Working paper 24.* Available at: http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf

Van der Veld, W.M, Saris, W.E and A. Satorra (2009). *Judgement Rule Aid software*. Jrule 2.0: User manual (Unpublished Manuscript, Internal Report). Radboud University Nijmegen, the Netherlands.

Yan, T., and R. Tourangeau (2008). "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times". *Applied Cognitive Psychology*, 22: 51–68, DOI: 10.1002/acp.1331

Zhang (2013). Satisficing in web surveys: implications for data quality and strategies for reduction. Doctoral dissertation University of Michigan 2013. Available at: http://deepblue.lib.umich.edu/handle/2027.42/97990

Tables and figures

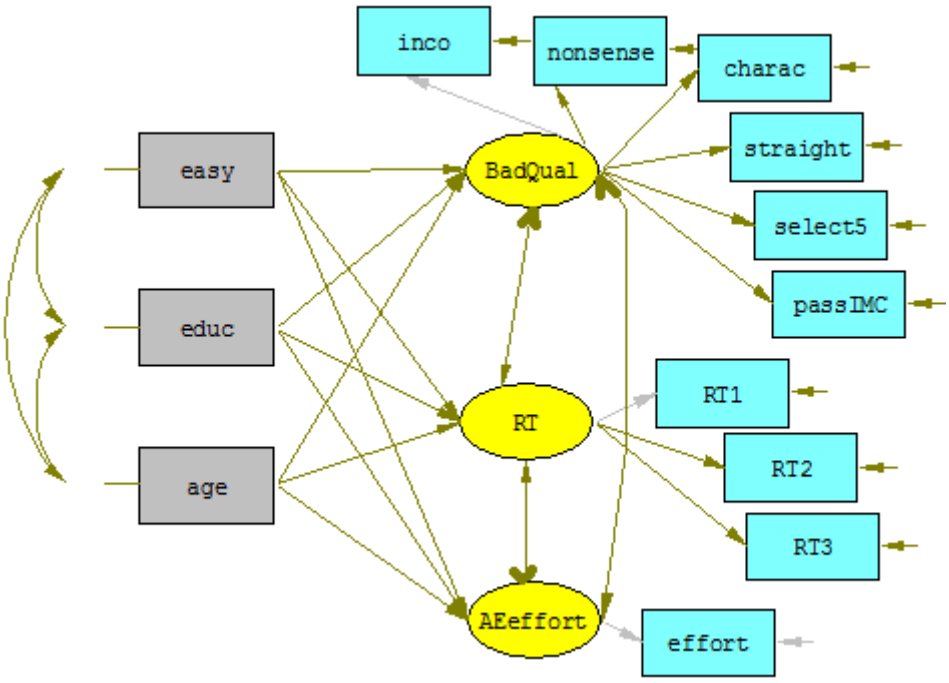

Figure 1: The complete SEM model

Select5	Spain	Mexico	Colombia	Overall
0 time	48.7%	40.2%	31.7%	40.2%
1 time	7.8%	12.4%	18.2%	12.8%
2 times	11.3%	12.4%	13.5%	12.4%
3 times	32.2%	35.0%	36.7%	34.6%

Table 1: Proportions of respondents that selected five options as requested

Pure straight-liner	Spain	Mexico	Colombia	Overall
0 time	81.4%	89.5%	94.1%	88.3%
1 time	17.1%	8.5%	5.3%	10.4%
2 times	0.3%	0.3%	0.6%	0.4%
3 times	1.2%	1.6%	0%	0.9 %

Table 2: Proportions of respondents that are pure straight-liners in 0 to 3 grids

Non-sense	Spain	Mexico	Colombia	Overall
0 time	91.9%	93.8%	95.3%	93.7%
1 time	7.0%	4.6%	4.4%	5.3%
2 times	1.2%	1.6%	0.3%	1.0%

Table 3: Proportions of respondents that wrote a non sense 0, 1 or 2 times in the open questions

Estimates		**Compl. Stand.**			**Unstand. with t-value in parenthesis**					
		Spain	**Mexico**	**Colombia**	**Spain**		**Mexico**		**Colombia**	
Structural model	Bad with RT	-.27	-.22	-.26			-.09*	(-4.54)		
	Bad with AE	-.05	-.04	-.05			-.02	(-.80)		
	AE with RT	.15	.15	.14			.12*	(3.22)		
	Easy on Bad	-.17	-.56	-.17	-.07*	(-2.75)	-.28*	(-5.98)	-.07*	(-2.75)
	Easy on RT	.04	.03	.04	.03	(.90)	.02	(.42)	.03	(.90)
	Easy on AE	.00	.00	.00			.00	(-.05)		
	Educ on Bad	-.06	-.05	-.06			-.02	(-1.16)		
	Educ on RT	-.14	-.04	-.14	-.11*	(-3.48)	-.04	(-.74)	-.11*	(-3.48)
	Educ on AE	.04	.04	-.11	.04	(.82)	.04	(.82)	-.11	(-1.43)
	Age on Bad	.05	-.14	-.17	.02	(.57)	-.07*	(2.60)	.02	(.57)
	Age on RT	.26	.27	.27	.21*	(7.52)	.21*	(7.521)	.21*	(7.52)
	Age on AE	-.01	-.01	-.20	-.01	(-.15)	-.01	(-.15)	-.20*	(-2.62)
Measurement model	Bad by Inco	.41	.48	.42			1			
	Bad by Nonsens	.19	.23	.20			.48*	(3.76)		
	Bad by Charac	-.26	-.31	-.26			-.63*	(-4.56)		
	Bad by Straight	.47	.55	.48			1.15*	(6.57)		
	Bad by Select5	-.15	-.18	-.15			-.37*	(-3.20)		
	Bad by PassIMC	-.20	-.24	-.20			-.49*	(-3.93)		
	RT by RT_1	.81	.80				1			
	RT by RT_2	.79		.78			.96*	(25.53)		
	RT by RT_3	.82					1.01*	(25.65)		
Global Fit		$Chi^2(df) = 302.37\ (220)$; RMSEA =.032 ; CFI=.96								

Note: when the estimates are similar for all three countries, they are presented only once. There is a star next to the estimates when the t-value indicates that the coefficient is significantly different from 0 when a significance level of 5% is used.

Table 4: Standardized and unstandardized estimates in the three countries

Appendix

```
!LISREL input final model, the first group is Spain

data ng=3 ni=13 no=345 ma=km
cm
*
1.0000
0.7093  1.0000
0.7260  0.6974  1.0000
0.1209  0.0890  0.1264  1.0000
0.0515  0.0657  0.0569 -0.0580  .613
-0.0574 -0.0236 -0.0814 -0.0224 -0.0211  1.0000
0.0199 -0.1054  0.0169  0.0045  0.0428 -0.0377  1.0000
0.1645  0.3489  0.1319  0.0397  0.0284 -0.0112  -0.1921  1.0000
-0.1680 -0.1993 -0.1342 -0.0145 -0.1442 -0.0294  0.1659  -.2041 1.0000
0.2563  0.1851  0.2765  0.0551 -0.0288 -0.0681 0.0222 0.1239 -.0642 1.0000
0.2568 0.1106  0.0393 0.0162 0.0763 -0.0986 0.0141 0.1794 -.1059 0.3749 1.0000
-0.1662 -0.1173 -0.1206 0.0403 -0.0402 -0.0589 0.0312 0.0135 0.0364 0.0509
0.0292  .884
0.1914 0.1791 0.1793 -0.0031 0.1578 0.0553 0.1558 0.0508 -0.02390.0414 -0.0038
-0.1280 .997

labels
RT1 RT2 RT3 effort easy inco nonsense nbcharac straight select5 passIMC educ
age

select
inco nonsense nbcharac straight select5 passIMC RT1 RT2 RT3 effort easy educ
age/

model ny=10 nx=3 ne=3 be=fu,fi ps=sy,fi ph=sy,fr ly=fu,fi te=fu,fi  ga=fu,fi

le
BadQual RT effort

!easy, education, age on quality and RT and AE
fr ga 1 1 ga 2 1 ga 3 1
fr ga 1 2 ga 2 2 ga 3 2
fr ga 1 3 ga 2 3 ga 3 3

! fix scale quality, RT
va 1 ly 1 1
va 1 ly 7 2


!other indicators quality and RT
fr ly 2 1 ly 3 1 ly 4 1 ly 5 1 ly 6 1
fr ly 8 2 ly 9 2

!AE: fix loading to quality coefficient for
va .734 ly 10 3

!fr variance latent variables
fr ps 1 1
fr ps 2 2 ps 3 3

!links between RT, quality and AE
fr ps 2 1 ps 3 1 ps 2 3

!error terms free expect for effort where it is fixed to .462
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
fr te 7 7 te 8 8 te 9 9
va .462 te 10 10


!extra corrections
```

```
!nbcharac with RT2 because the 2 open questions are in part 2
fr te 8 3
!because passIMC is in part 1
fr te 6 7
!because passIMC and select5 are based on same grid
fr te 5 6
!because straight and incoherence are mainly based on the same grids
fr te 4 1

out mi ad=off it=2000 sc ec rs


group2 Mexico
data ni=13 no=305 ma=km
cm
*
1.0000
0.5445  1.0000
0.5791  0.5975  1.0000
0.0973  0.0855  0.1185  1.0000
-0.0262  0.0002  0.0804  0.0496  .613
-0.0159 -0.0881 -0.0798  0.0614  -0.1784   1.0000
-0.0022 -0.0382 -0.0584  -0.0564  -0.1357  0.2496  1.0000
0.1061  0.4158  0.1686  0.0771   0.0455  -0.1830  -0.2457  1.0000
-0.0889 -0.0638 -0.1077  0.0239  -0.3287  0.2567  0.2256  -0.1407   1.0000
 0.1695 0.0437 0.1563 0.0503 -0.0014 -0.1472 -0.0170 0.1007 -0.0435  1.0000
0.1432 -0.0017 -0.0682 0.0133 0.1049 -0.1503 -0.0310 0.0810 -0.0907 0.1705 1.0
0.0891  0.0152  -0.0006   0.0177   0.0317  -0.0963  -0.0814  0.0354  -0.0505
-0.0302  0.0378   .884
0.2550  0.1538   0.1721  -0.0117   0.0419  -0.1390  -0.1118   0.0630  -0.1272
-0.0186  0.0162   0.3119   .997

labels
RT1 RT2 RT3 effort easy inco nonsense nbcharac straight select5 passIMC educ
age

select
inco nonsense nbcharac straight select5 passIMC RT1 RT2 RT3 effort easy educ
age/

model ny=10 nx=3 ne=3 be=in ps=in ph=in ly=in te=in ga=in

!corrections
fr ga 2 2, ga 1 1, ga 2 1
fr te 6 5 te 4 1
fr ph 3 2, ph 3 1
fr te 3 2
fr ga 1 3

out mi ad=off it=2000 sc ec rs

group 3 Colombia
data ni=13 no=336 ma=km
cm
*
1.0000
0.6618  1.0000
0.6190  0.6525   1.0000
0.0237  0.0327  -0.0112   1.0000
0.0830  0.0685   0.0471  -0.0250   .613
-0.1443 -0.1700  -0.1972  -0.0119  -0.0475   1.0000
-0.0263 -0.0468   0.0326  -0.0746  -0.0066   0.0799   1.0000
0.1064  0.3484   0.0677   0.0301  -0.0041  -0.2130  -0.1153  1.0000
-0.0477 -0.0475  -0.0243  -0.0095  -0.0638   0.2837  -0.0514  -0.0547   1.0000
0.0520  0.0472 0.1353 -0.0373 -0.0040 -0.0426 0.0422 0.0923  -0.0460   1.0000
0.1639 0.0492 -0.0273 -0.0112 0.0793 -0.0902 -0.0098 0.0652 -0.0351 0.0254 1.0
-0.0609 -0.0168 -0.0282 -0.1188 0.0249 -0.0174 0.0443 0.0870 -0.0788 -0.0002
0.0558   .884
```

```
0.1981   0.2771 0.2710 -0.1678 0.2124 -0.1313 0.0930 0.1126 -0.0784 0.0755
0.0228   0.2841   .997

labels
RT1 RT2 RT3 effort easy inco nonsense nbcharac straight select5 passIMC educ
age

select
inco nonsense nbcharac straight select5 passIMC RT1 RT2 RT3 effort easy educ
age/

model ny=10 nx=3 ne=3 be=in ps=in ph=in ly=in te=in ga=in

!corrections
fi te 6 5
fr ga 3 2 ga 3 3
fr ph 3 2

eq te 2 4 1 te 4 1
eq ga 2 1 3 ga 1 3

pd
out mi ad=off it=2000 sc ec rs
```

[i] Except for the pages before the survey really starts (with introduction and filter for gender and age) and for A6 and C12 for which we excluded the highest 5%. Otherwise, we still had some clearly impossible times. We also tried other computations, using different thresholds than 1 or 5%, but the overall results were similar.

[ii] We checked if the same respondents were often part of the highest 1% which could indicate more slow respondents but found that this was not the case. Very few respondents are more than one or two times part of the highest 1%.