UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 32

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

Impact of formulating a careful introduction and ask respondents to commit themselves on the quality in a web panel survey

Melanie Revilla

Research and Expertise Centre for Survey Methodology

Universitat Pompeu Fabra

Abstract:

Web surveys are becoming every day more present in survey research. However, there are still problems with these surveys that have not been solved yet, in particular in the case of online panels using economic incentives. Some people can participate in order to get the reward and without any intention of answering properly and sincerely to the surveys. Speeding and low quality data may be expected in that case. One way to try to face these undesirable behaviors is to sensibilize the respondents to the importance of their answers for research and the necessity for the conclusions to be valuable that they are answering carefully. The sensibilization can be done through motivational messages. This paper reports the results of an experimental design in which respondents from the online panel Netquest in Spain were randomly assigned to three groups: one control group, one group getting a carefully formulated introduction that aimed to motivate respondents in making an effort to answer properly, and one group getting the same introduction but with an additional commitment stating that the respondents could accept or reject by clicking on a button. The results show no effect when showing only the introduction. When the introduction is combined with the commitment, a small effect is found on some behaviors but not all. It seems to come only from some specific respondents, who were answering to the survey by doing some effort but not much and can then be moved to make a higher level of effort. But respondents that behave badly continue behaving badly even when they commit to do their best. It suggests that more radical solutions than a simple sensibilization may be necessary to reduce the undesirable behaviors of these online panelists.

Keywords:

Online surveys, motivational message, commitment, speeding, quality, satisficing

Acknowledgement:

We are very grateful for the permission we got to use the questionnaire of the Reputation Institute in Spain (http://www.reputationinstitute.com) for our experiment. We also appreciate very much the careful way in which Netquest performed the data collection and all the help given to us by different persons of the team, in particular Carlos Ochoa and Salvador Masdeu. We finally want to thank Willem Saris for his comments on previous drafts of this paper.

Number of words (excluding abstract, tables, references, appendices): 6341

I. Introduction

Web surveys are becoming every day more present in survey research. However, there are still problems with these kinds of surveys. Coverage and non-response problems are some of them and make the representativity of many online surveys doubtful (Bethlehem 2010; Couper 2000; Dillman, Tortora and Bowker 1999).

In order to fight the non-response problem, many online surveys are using incentives to attract more respondents. Some use incentives like personal feedback, but most use economic incentives, in the form of lotteries, points that can be cumulated and exchanged for gifts, or money.

Quite some research has been done to test if the expected positive impact of incentives in web surveys on the response rates is found. Even if the results are not all going in the same direction (e.g. Cook, Heath and Thompson 2000, found no impact of the incentives on the response rates) most of the evidence indicates some increase in response rates when economic incentives are used (Bosnjak and Tuten 2003; Cobanoglu and Cobanoglu 2003; Couper 2000).

Nevertheless, the non-response problem is only part of the picture. An increase in response rate is not necessarily a good thing if it involves a decrease in the quality of the answers. If respondents are attracted only by the incentives, they may be unwilling to go carefully through all the cognitive steps (Tourangeau, Rips and Rasinski 2000) required to answer properly to the survey questions. Instead, they can just give random answers in order to finish the survey quickly and get the reward. This behavior of not putting the maximum effort into answering the questions is referred to as satisficing (Krosnick 1991).

Satisficing exist in all data collection modes but we expect it to be higher and stronger in most online surveys because of their characteristics. First, in online surveys, the researchers have little control over the real identity of the respondents (difficult to check if the person answering is the one we think). In addition, many online surveys are not based on random sampling but on volunteers. As a consequence, it is easy for someone who wants to make money to get in. Finally, an important part of online research is done via panels in which the participation is rewarded every time. So a respondent can cumulate rewards.

The main reason for satisficing is the lack of adequate motivation of the respondents. In order to try to limit satisficing, the researchers could therefore stop using economic incentives that make respondents participate for the wrong reasons. Nevertheless, researchers have also to take into account the representativeness of their sample and it is not so easy to make respondents participate without economic incentives. They can therefore try to solve the problem in different directions. They can improve the design of the survey for the respondent to be easier and more pleasant to answer: for that, researchers should formulate clear and parsimonious questions to measure their concepts of interest (Saris and Gallhofer 2007) and order the questions in a logical way, facilitating the work of the respondents.

Besides, online surveys offer many possibilities to improve the survey experience of the respondents. However, their impact on the quality of the answers is not always clear. Some elements have already been studied, for instance the impact of personalizing the email invitation to the survey on the response rates and the quality of answers (Heerwegh, Vanhove, Matthijs, Loosveldt 2005). Some work has also been done on the visual presentation of online surveys: use of plain versus fancy design (Dillman, Tortora, Conradt and Bowker 1998), use of colors, sound, videos, images (Couper, Tourangeau and Kenyon 2004). Some work has been done on the impact of providing interactive feedback to the respondents, in the form of a progression bar or of additional screens or pop-up windows (Conrad, Couper, Tourangeau and Galesic 2005; Couper, Traugott and Lamias 2001).

Additional screens or pop-up windows can be used to provide information on the respondents progression ("you are about one third done with the survey", Couper, Traugott and Lamias 2001, pp.233) and/or motivational messages that remind the respondents of the importance of their answers at key points of the survey ("please continue as your answers are important to us", Couper, Traugott and Lamias 2001, pp.233).

The use of motivational messages can be done in other ways than punctual screens. For instance, Kapelner and Chandler (2010, section 2.2) look at the impact of adding at the bottom of each screen of the survey an extra message in red text: "Please, answer accurately. Your responses will be used for research."

Following this line of research, this paper looks at the impact of a motivational message presented in the introduction to the survey. The idea is to sensibilize the respondents by underlining the importance of the research conducted and the necessity for the conclusions to be valid that respondents take enough time to read the questions and think about the answers. The introduction is also warning them that if they do not have time to answer properly at that moment, they should return later to complete the survey. Besides the impact of the motivational message, we look at the extra impact of asking respondents to commit themselves to fill the survey properly by checking a button before being able to start the survey. We look at the impact of these two elements (introduction, additional commitment) on the speed of answers as well as different indicators of data quality.

Next, we will present the hypotheses we want to test in more detail, the method used to do so and the data. Then, the results will be shown and discussed.

II. Hypotheses, method, data

For this experiment, we randomly assigned the respondents to three split-ballot groups. The first group is a control group. Respondents in that group got the usual introduction page that is always shown in the Netquest surveys and includes very little text. Therefore, respondents can identify at the first glance that this is the introduction they usually get. The second group got a developed introduction, insisting on the importance for research that respondents take the time to answer properly ("intro only" group). The third group got the same introduction but with an additional commitment box they had to check if they agreed to do their best to answer the survey (“intro+sign” group). The text of the introductions is presented in Appendix 1.

We look at the impact of the two treatments on the total time of completion of the survey and different indicators of the quality of the data.

Our hypotheses are:

H1a: The presence of an introduction including the motivational message will reduce speeding behaviors, understood as very short total time of completion.

Said differently, *H1a* means that the presence of the introduction will increase the average completion time of the survey - much beyond the increase due to the time spent on the introduction page itself.

H1b: The presence of the introduction will improve the data quality.

H2a: The additional fact that respondents have to commit themselves will reduce speeding even more.

H2b: The additional fact that respondents have to commit themselves will improve the quality even more.

Indeed, when reminded how crucial for research answering honestly and properly is, we expect that the respondents will feel bad satisficing strongly and will make more effort in answering. If they have to commit themselves, we expect they will respect their own commitment and the effort will be even higher.

The data is taken from a survey ordered by the Reputation Institute, a survey institute specialized in corporative reputation, and answered by 1621 panelists from the Netquest Panel in Spain between the 22nd of April and the 9th of May 2013. Netquest[1] is an online fieldwork company that started its first online panel in Spain in 2006. Since then, it has created many other online panels in Portugal and in central and Latin America. In order to recruit the panellists, Netquest asks at the end of a short satisfaction survey proposed by different websites if the respondents would be willing to take part in a panel. If yes, they ask their contact information. In a second step, they use this information to create a panel "by invitation only" selecting within the email addresses who they invite. Because of this recruitment procedure and the efforts made by the company to improve the data quality, Netquest is accredited with the ISO 26 362 quality standard.

The survey used for testing our hypotheses was about several insurance companies. The expected questionnaire length was 20 minutes on average. The sample was drawn using quotas for the different Spanish regions.

III. Control of the treatments

In the experimental design used, respondents were randomly assigned to three groups. Since the assignment is random, if we get significant differences between the control group and the treatment groups, we can conclude that the experimental manipulation is what is causing the difference.

[1] For more information: www.netquest.com

Nevertheless, the first treatment (a developed introduction) has the particularity that even if it is “given” to the respondents, it is up to them to decide if they really “take” it or not. Indeed, in a self-completed survey, we can make sure that the respondents get the introduction text but not that they read it. The same applies to the second treatment, except that, this time, the second part of this treatment (check the commitment box) is “compulsory”. If the respondents do not check any box (accept the commitment or refuse it) they cannot go on with the survey. In that case, we can control that they did commit, but still we cannot be sure that they did so knowing what they were committing to (they may have checked the box without reading the text).

In order to get a first idea of how many of the respondents really “took” the treatments, Table 1 looks at the time spent by the respondents of the different split ballot groups on the introduction page, and at the percentage of respondents that agree to commit in the second treatment group.

Table 1: Percentages by split ballot group of respondents spending different times on the introduction page

ALL	Control		Intro only		Intro+sign	
	%	Cumul %	%	Cumul %	%	Cumul %
≤ 5 seconds	34.32	(34.32)	7.52	(7.52)	0.19	(0.19)
6 to 10 seconds	31.73	(66.05)	19.45	(26.97)	5.77	(5.96)
11 to 15 seconds	11.50	(77.55)	15.05	(42.02)	12.66	(18.62)
16 to 20 seconds	4.45	(82.00)	15.96	(57.98)	14.90	(33.52)
21 to 30 seconds	4.08	(86.09)	16.15	(74.13)	22.91	(56.42)
31 to 45 seconds	3.53	(89.61)	8.99	(83.12)	18.62	(75.05)
46 seconds to 1 minute	1.86	(91.47)	4.22	(87.34)	8.01	(83.05)
1.01 to 2 minutes	2.97	(94.43)	5.32	(92.66)	6.89	(89.94)
2.01 to 5 minutes	2.79	(97.22)	2.01	(94.68)	2.42	(92.36)
More than 5 minutes	2.78	(100.0)	5.32	(100.0)	7.64	(100.0)
No. observations	539		545		537	
Agree to commit					99.25%	

It is clear from Table 1 that a part of the respondents is not really reading the introduction, not only in the control group (where we expect so) but also in both treatment groups. Indeed, in the group “intro only”, 42.02% of the respondents spent 15 seconds or less on the introduction page, which we consider too quickly[2] for reading all the text carefully. In the group “intro+sign”, this percentage is lower (18.62%) but the introduction is also a bit longer, as they have an additional sentence they have to agree with. We expect this extra task to take them at least five seconds if they do read before clicking (and read fast). Therefore, we can add to the respondents that answered between 15 and 20 seconds. Then, we have 33.52%, which is still significantly lower (almost 10 points less) than the percentage of respondents that spent less than 15 seconds on the introductory page in the “intro only” group.

This suggests that when the respondents have to check a box of commitment, a lower percentage tends to speed on the introduction page. More respondents somehow want to make sure what they are committing to. Still, even in the “intro+sign” group, there is a relatively high percentage of respondents that cannot have read the introduction properly. For this reason, we expect the effect of the experimental manipulations to be smaller than we first thought. If the respondents decide not to “take” the treatments, these treatments cannot have a significant impact on the speed and quality of the answers.

However, a majority of respondents spent a reasonable time on the introduction page, suggesting they probably read the content. For this majority, we expect our hypotheses

[2] We have to notice that reading on Internet is not the same as reading on paper: people are much more used to go through Internet pages by looking for keywords and reading “in diagonal” than by reading everything in details. Still, experimenting with a few subjects, we found that less than 15 seconds seems too short to read carefully all the text of the introduction page in the treatment groups.

to be true. Besides, almost all respondents in the “intro+sign” group (99.25%) checked the commitment box, suggesting that they were reading at least something. Otherwise, if it would be just by chance, we would expect more respondents to check the box refusing to commit. We therefore look next to the differences between control and treatment groups with respect to speed and quality.

IV. Analyses and results

III.1. Total time of completion

We first consider the speed of answers measured by the total time of completion. This indicator is difficult to compute because we can only get a direct measure of the time respondents spend on a given webpage but this is not necessarily the “real” time spent to answer the questions. Indeed, we cannot know if they really spent this time answering the question or if they did another task during this time: answer to the phone, chat on Facebook, let the survey open and go shopping or go to sleep and come back hours later. If they did another task, this is referred to as “multitasking”.

Therefore, in order to compute the total time of completion of each respondent, we substituted for each page the times of the 1% respondents with the highest time[3] (considered as the ones that clearly were multitasking) by the average time spent by the other 99% to answer to the questions on that same webpage.

[3] Except for the introduction page, Qareas, Q550, Q750, Q950 and Q606 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.

Table 2 presents for the three split-ballot groups the distribution of total time spent to answer the survey.

Table 2: Percentage of respondents for different total time of completion categories

ALL	Control		Intro only		Intro+sign*	
	%	Cumul%	%	Cumul%	%	Cumul%
7 min or less	1.11	(1.11)	1.83	(1.83)	**0.93**	**(0.93)**
7.01 min to 10 min	7.24	(8.35)	6.79	(8.62)	**4.28***	**(5.21*)**
10.01 min to 15 min	20.96	(29.31)	20.92	(29.54)	**17.32**	**(22.53*)**
15.01 min to 20 min	25.23	(54.55)	22.94	(52.48)	**22.53**	**(45.07*)**
20.01 min to 25 min	17.25	(71.80)	19.63	(72.11)	22.16*	(67.23)
25.01 min to 30 min	11.69	(83.49)	11.93	(84.04)	13.78	(81.01)
30.01 min to 45 min	13.73	(97.22)	13.94	(97.98)	15.27	(96.28)
45.01 min and more	2.78	(100.0)	2.02	(100.0)	3.72	(100.0)
Mean in minutes	21.2		21.0		22.7*	
No. observations	539		545		537	

Note: The total time of completion is computed as the sum of the time spent on each webpage substituting the times for the highest 1% for each page by the mean of the other 99%. The stars next to the numbers indicate that the differences in proportions (percentages or cumulative percentages) or means between the "intro+sign" and the control group are statistically significant at the 95% level. The stars next to the name of the group indicate that the distributions are significantly different (Kolmogorov Smirnov test) for the corresponding treatment group compared to the control group.

Table 2 shows that, compared to the control group, there is no reduction of the percentage of very quick total times of completion for the "intro only" group. On the contrary, in the "intro+sign" group, there is a significantly lower percentage of respondents answering in less than 20 minutes, which was the expected mean time of completion for the survey. The mean time of completion is also significantly higher in the group that got the commitment treatment compared to the control group.

This first result suggests that the commitment has some positive effect on speeding behaviors (support for *H2a*), whereas the introduction in itself does not have the expected effect (no support for *H1a*).

However, respondents answer the same questions for one, two or three brands, so if in the different split ballot groups there are different proportions of respondents answering one brand versus two versus three, this could cause differences that are not a direct effect of the introduction.

On the one hand, since the split-ballot groups are randomly drawn, we do not expect that to happen. On the other hand, the number of brands answered by each respondent is determined by the answers to a filter question. If they report they know one brand, they will get the rest of the questions for one brand. If they report they know two, they will get them for two brands. If they report they know three or more, they will get them for three brands. It is possible that the groups with treatments took this filter question more seriously and, on average, reported more known brands. This would then have an indirect impact on the total time of completion.

In order to check that this is not the cause of the differences, Table 3 presents the times of completion for each split-ballot group focusing only on the respondents that were answering the questions for three different brands.

Table 3: Percentage of respondents for different total time of completion categories

ONLY IF 3 brands	Control		Intro only		Intro+sign	
	%	Cumul%	%	Cumul%	%	Cumul%
7 min or less	0.81	(0.81)	0.81	(0.81)	**0.39**	**(0.39)**
7.01 to 10	2.02	(2.83)	2.02	(2.83)	**1.16**	**(1.54)**
10.01 to 15	9.72	(12.55)	10.53	(13.36)	**6.95**	**(8.49)**
15.01 to 20	21.46	(34.01)	22.67	(36.03)	**15.44**	**(23.94*)**
20.01 to 25	22.67	(56.68)	21.46	(57.49)	27.80	(51.74)
25.01 to 30	16.19	(72.87)	16.60	(74.09)	16.99	(68.73)
30.01 to 45	21.86	(94.74)	21.86	(95.95)	25.10	(93.82)
45.01 and more	5.26	(100.0)	4.05	(100.0)	6.18	(100.0)
Mean in minutes	25.4		24.8		26.6	
No. observations	247		247		259	
% of previous sample	45.8%		45.3%		48.2%	

Note: The stars indicate statistically significant differences between control and treatment groups at the 95% level.

First, Table 3 shows that 48.2% of the respondents of the group “intro+sign” are answering to the questions for three brands. This is a bit higher than the percentages for the control group (45.8%) and the “intro only” group (45.3%). However, the differences in proportions between the control and the treatment groups are not significant at the

95% level. The treatments do not lead to more respondents reporting they know three or more brands.

Again, we observe that the percentage of respondents answering in less than 20 minutes is significantly lower in the “intro+sign” group compared with the control group. On the contrary, we do not observe this trend for the “intro only” group.

We can notice also that in all split-ballot groups, still 1.54% to 2.83% of respondents answer to the questions for three brands in less than 10 minutes, which seems practically impossible if they are reading the questions properly. This suggests that we do not get rid of all the “speeders”.

Finally, the difference in mean time of completion is not significant anymore for the group “intro+sign”, but still the direction remains the same.

III.2. Satisficing

The first indicator of data quality studied is the level of strong satisficing, which is measured in two different ways: looking at the results of an instructional manipulation check and at the non-differentiation over items.

III.2.1 Instructional manipulation check (IMC)

An instructional manipulation check “measures whether or not participants are reading the instructions and thus provides an indirect measure of satisficing. It consists of a question embedded within the experimental materials (…) the IMC asks participants (…) to provide a confirmation that they have read the instruction.” (Oppenheimer, Meyvis and Davidenko 2009, pp.867).

In the survey studied here one instructional manipulation check was introduced. It was proposed as one more item within the battery Q550 with the goal of detecting respondents that were not answering carefully to the questionnaire. In this battery of 26 items, was therefore added one row which instead of proposing a statement on which the respondents had to give their opinion, was giving the following instruction: “Select, to show that you are reading, the option ‘describe very well’ in the scale”. The common scale for all 27 items of the battery including the IMC was from “1-don’t describe well” to “7-describe very well”. One “don’t know” category was also available for all the items.

The percentage of respondents that failed selecting the answer category “describe very well” is 22.55% in the control group, 23.85% in the “intro only” group and 19.14% in the “intro+sign” group.

First, the percentage of respondents failing the IMC is huge: more than one out of five respondents failed to select the option they were asked to. This gives reasons to worry about the quality of the answers of more than one fifth of the sample. Nevertheless, the IMC was introduced within a battery that was in a relatively advanced position within the survey and that was particularly demanding to answer: the number of items was huge, the number of answer categories was quite big and the items were asking about things difficult to know for most respondents. In these conditions, even respondents that usually are making effort may decide to satisfice. Moreover, the respondents that failed the IMC were excluded from the sample in the substantive survey done by the Reputation Institute.

With respect to the impact of the treatments, proposing simply an introduction has no positive impact: the percentage of failure for the group "intro only" is even a bit higher but the difference in proportions is not significant at the 95% level[4]. Asking to the respondents to commit themselves leads to a slightly lower percentage of respondents failing the IMC. The difference however is small and not statistically significant.

Overall, no support is found for *H1b* and *H2b* when the quality of the data is measured by the failure to the IMC.

III.2.2 Non-differentiation between items

The second indicator of strong satisficing considered is the non-differentiation between items. We especially expect this behavior to happen when several items are presented in a battery. Then, respondents that do not want to make much effort may have the tendency to select the same answer category for all the items. This behavior consisting in selecting the same answer for all the items in a battery is also called "straight-lining".

The questionnaire includes several batteries but some are really short. Since all the items are in the same direction and sometimes very similar, choosing always the same answer in such batteries is not an indicator of satisficing. It seems even normal to answer for instance three times "I agree" to three items measuring the same concept. Therefore, we test the hypotheses only on the batteries that include more than five items. There are two: one battery of 26 or 27 items[5] (repeated up to three times if the respondents answered the questions for three brands) and one of six items (possibly also repeated up to three times).

[4] The tests of significance were all done using Stata 10.

[5] The first battery included one IMC in addition to the 26 items; in the repetitions for brands 2 and 3, the IMC was not repeated.

III.2.2.1 Straight-liners in the long battery

The battery Q550 is a very interesting candidate to study straight-lining because it is really a long battery, so we cannot expect people to have exactly the same opinion on all the items. Besides, it is a battery with a relatively large scale (from “1-don’t describe well” to “7-describe very well”, plus the “don’t know” category) offering the respondents more possibilities to nuance their opinion.

Table 4 reports how many respondents in each split ballot group selected 26 or 27 times out of 27 items the answer category “1”, the answer category “2”, etc. We call “pure straight-liners” the ones that select 27 times the same answer category. But we also considered the ones that selected 26 times the same answer category as straight-liners, for three reasons.

First, one of the 27 items was the instructional manipulation check. We may have respondents that are going too fast to process the different items carefully but still are able to detect the IMC and answer it properly. Second, the quickest way of answering such a battery on a computer is to use the keyboard and select with the tabulation key. If doing so, the first answer category still has to be clicked and can be whatever, but starting from the second item until the last, the second answer category will always be selected. If the respondents are doing this, we expect them to select 26 times the same answer (category “2”), but not necessarily 27 times. Third, even the straight-liners can make mistakes and just by chance select once another answer category.

Table 4: First brand Q550: 27 items (includes instructional manipulation check)

	Control (501 obs)		Intro only (499 obs)		Intro+sign (491 obs)		Total (1491 obs)		
	26 times	27 times	26 times	27 times	26 times	27 times	26 times	27 times	Total 26+27
Answer “1”	0	2	0	2	0	1	0	5	5
Answer “2”	0	0	0	1	0	0	0	1	1
Answer “3”	0	1	1	0	0	1	1	2	3
Answer “4”	2	1	4	5	1	5	7	11	18
Answer “5”	0	4	0	2	1	1	1	7	8
Answer “6”	1	1	0	1	0	1	1	3	4
Answer “7”	1	1	0	1	0	1	1	3	4
Answer “9” (DK)	11	14	10	19	11	11	32	44	76
Total in %	2.99%	4.79%	3.01%	6.21%	2.65%	4.28%	2.88%	5.10%	7.98%

Note: The differences in proportions between control and treatment groups are not significant.

In total, 5.10% respondents are pure straight-liners and 7.98% straight-line on either 26 or 27 of the items. This is quite large if we think that this means non-differentiation over 26 or 27 items, which is a very strong form of satisficing.

The introduction does not allow reducing the proportion of this problematic behavior. The percentage of pure straight-liners is even larger in the “into only” group: 6.21% versus 4.79% in the control group. But the difference is not statistically significant. When respondents have to commit themselves, the percentage of pure straight-liners goes down to 4.28% but again, the difference with the control group is not statistically significant. The percentage of respondents selecting 26 times the same answer is also the lowest (2.65%) for the “intro+sign” group but once more the reduction is not significant at the 95% level. Even when committing themselves to do their best, a non negligible part of the respondents are therefore satisficing strongly on this long battery.

Looking at Table 4 in more detail, we see that the straight-liners are not using the keyboard to go as fast as possible. Indeed, nobody, in any of the three split ballot groups, is selecting 26 times answer “2”.

The most chosen answers are the middle of the scale (“4”) for the substantive answers and the “don’t know” option. This last one is chosen in 26 or 27 of the items by up to 5.10% of the total sample.

Therefore, the straight-liners are choosing categories suggesting they have no real opinion or do not want to figure it out or tell it to the researcher. They go for the middle and for the “don’t know” option. This combines two sorts of satisficing: failing to differentiate among a set of diverse objects in ratings and saying “don't know” instead of reporting an opinion. Thus, it may seem that these respondents are doing an even worse “job”.

However, it is also possible that some respondents had no idea about many of the questions that required having a relatively good knowledge of the companies. As a consequence, they started to answer “don’t know” to all items without considering them anymore because they did not know the first ones and then assumed it would be the same for the following ones. This is to some extend confirmed by looking at one open question where around 8% of the respondents complained that some questions were too much oriented towards clients or workers of the firms and a “normal” respondent could not know what to answer.

The same battery is repeated for the respondents that knew more than one brand. We look at straight-lining in the other second and third repetitions. These batteries have one item less (no IMC) so the “pure straight-liners” are the ones selecting 26 times the same answer category.

The trends are similar to what we have seen: most of the straight-liners always select the “don’t know” option (6.36% of the total sample of respondents to brand 2, and 5.70% to the respondents to brand 3) and then the middle category; no evidence at all in favor of the use of the keyboard. Therefore, Table 5 only presents the percentages of respondents selecting 25 or 26 times the same option (whatever the option was).

Table 5: Repetition of the battery for brand 2 (Q750) and brand 3 (Q950): 26 items (no IMC)

	Control		Intro only		Intro+sign		
	25 times	26 times	25 times	26 times	25 times	26 times	Total 25+26
Brand 2 Q750	2.17%	8.94%	3.59%	8.97%	2.83%	9.51%	12.02%
Brand 3 Q950	4.54%	5.94%	4.06%	10.51%*	2.56%	7.03%	11.52%

Note: Because of filters, for the second brand (resp. third), we have 369 (286) respondents in the control group, 390 (295) in the "intro only" group and 389 (313) in the "intro+sign" group. The stars indicate significant differences in proportions between the control and the treatment groups.

For the second brand, we observe 9.15% of pure straight-liners and for the third brand 7.83%. For the second brand, 12.02% of the respondents choose 25 or 26 times the same answer and for the third brand 11.52%. The treatments do not help to reduce these behaviors at all. There is even a significant higher percentage of pure straight-liners in the group "intro only" compared to the control group.

III.2.2.2 Straight-liners in a battery of 6 items

The second battery used to look at straight-lining is much shorter. Thus, it does not give such strong indicators. Yet, there are six items that are quite different: e.g. one item is "I will speak positively about the company" whereas another item is "if I had a chance I would work for the company". Each item is measured on a scale from "1-Surely I won't do it" to "7-surely I will do it", and also includes a "don't know" option. Giving exactly the same answer six times seems unlikely if this is not due to satisficing.

The patterns observed are similar to the ones of the long battery. Therefore, Table 6 only presents the percentages of pure straight-liners for each repetition[6].

[6] In the first repetition, there are two different variables (Q450 and Q451) depending on which brand the respondents have to answer. Since the questions are the same for both variables, we combine them.

Table 6: Percentages of pure straight-liners

	Control		Intro only		Intro+sign		Total	
	%	No.obs	%	No.obs	%	No.obs	%	No.obs
Brand 1 Q450/451	11.58	(501)	13.43	(499)	10.79	(491)	11.94	(1491)
Brand 2 Q651	12.74	(369)	15.90	(390)	15.98	(388)	14.91	(1147)
Brand 3 Q851	15.38	(286)	20.68	(295)	11.82	(313)	15.89	(894)

Note: The differences in proportions between control and treatment groups are not significant.

Table 6 shows that the "intro only" group has a higher percentage of straight-liners in all three repetitions. So again, it seems that the introduction does not help limiting undesirable behaviors. The group "intro+sign" has a lower percentage of straight-liners than the control group in the first and third repetitions, but not in the second one. In any case, the differences are not statistically significant. Overall, similarly to the results of the long battery in the previous section, we find no impact of the treatments on straight-lining in the shorter battery.

III.3. Precision of the answers

Straight-lining is a behavior that occurs in batteries of questions. In other types of questions, there exist other ways of answering without providing the maximum cognitive efforts: in open narrative questions, one possibility is to give a short answer instead of developing a complete response. Therefore, Table 7 looks at the precision of the answers to all open narrative questions, measured by the average number of characters in each question. A summary measure of how much the respondents wrote in the survey is also considered. It is computed, for each respondent, as the sum of the numbers of characters over all the open narrative questions of the survey. The average over respondents of each split ballot groups of this total number of characters is presented in Table 7.

Table 7: Number of characters on average in all narrative open questions by split ballot group

	Control		Intro only		Intro+sign	
Name of question	Avg. char	No. obs	Avg. char	No. obs	Avg. char	No.obs
Q520	63.1	(488)	62.4	(482)	71.9*	(478)
Q580A	44.5	(150)	39.3	(129)	40.1	(127)
Q552	74.0	(380)	75.6	(367)	82.7*	(375)
Q407	63.8	(80)	53.2	(69)	77.0	(72)
Q406	56.0	(65)	47.6	(63)	50.0	(55)
Q452	66.8	(426)	65.5	(420)	70.4	(421)
Q720	58.3	(363)	53.7	(387)	64.8	(380)
Q780A	28.5	(104)	38.5 *	(93)	33.7	(101)
Q752	61.4	(240)	58.9	(253)	65.1	(265)
Q607	49.6	(55)	56.3	(53)	71.0*	(55)
Q606	44.9	(43)	37.0	(36)	34.5	(42)
Q652	56.5	(319)	53.5	(331)	63.1*	(316)
Q920	51.2	(281)	53.2	(294)	55.4	(311)
Q980A	38.6	(90)	36.9	(93)	32.0	(77)
Q952	57.1	(182)	55.6	(196)	62.3	(204)
Q807	51.6	(55)	65.2	(51)	75.7*	(44)
Q806	53.9	(35)	53.4	(21)	34.4*	(31)
Q852	57.0	(249)	50.4	(257)	56.8	(271)
ADO6	62.12	(535)	65.57	(545)	66.73	(537)
Mean total no. charac.	454.17	(539)	443.12	(545)	497.25*	(537)

Note: The stars indicate statistically significant differences between control and treatment groups at the 95% level.

Table 7 shows that the precision of the answers in the "intro only" group is not improved: in most of the questions, the average number of characters is even lower than in the control group, but the differences are not significant.

On the contrary, the "intro+sign" group seems to give answers that are a bit more precise: the average number of characters is higher in that group compared to the control group in 13 out of 19 questions, even if the increase is not always statistically significant. Considering the summary measure, the difference is of around 43 characters more in the "intro+sign" group compared to the control group, which is a significant difference in means.

III.4. Coherence of responses

For open questions, not only the length but also the content of the answers may be considered as an indicator of quality. Respondents that are writing something that has no sense (e.g., some numbers or letters that are not a word) are clearly a threat to the data quality (classified as "not an answer"). Respondents that are writing something that has a sense but is not answering the question (classified as "not answering question") are also problematic: they may not have understood the question and/or may not have read it carefully enough. Finally, respondents that do not want to do all the efforts needed to answer the survey may opt for easy answers like "I don't know".

We consider the open narrative question ADO6 that asked respondents "what would you do to improve this survey" and look at the percentages of respondents that give the undesirable answers just mentioned. Table 8 presents the results.

Table 8: Percentage of respondents giving different kinds of answers to the open question ADO6

ALL	Control		Intro only		Intro+sign	
	%	Cumul %	%	Cumul %	%	Cumul %
Not an answer	2.41	(2.41)	0.73*	(0.73*)	0.56*	(0.56*)
Not answering question	0.37	(2.78)	2.57*	(3.30)	1.49	(2.05)
Answer Don't know	10.39	(13.17)	8.44	(11.74)	6.70*	(8.75*)
No. observations	539		545		537	

Note: The stars indicate statistically significant differences between control and treatment groups at the 95% level.

Table 8 shows that the percentages of respondents writing an answer without sense and of respondents saying "don't know" are lower in both treatment groups compared to the control group (statistically significant for both groups with non-sense answers, but only for the "intro+sign" group with "don't know" responses). The percentage of respondents giving an answer that is not corresponding to the question asked is, on the contrary, a bit higher (statistically significant for "intro only").

Overall, the cumulative percentage of respondents for these three kinds of undesirable answers is 10.39% in the control group, whereas it is 11.74% in the “intro only” group (difference not significant) and 6.70% in the “intro+sign” group (difference significant). By adding the commitment, we achieve a small but significant reduction of undesirable answers in this open narrative question.

III.5. Total score of bad respondents

So far, we have considered several indicators of quality but each separately. However, respondents can make mistakes even if they put effort into answering (they are human beings). We can understand that there are moments in which they get distracted and satisfice. What we really want to detect are not the punctual satisficers but the respondents that satisfice repeatedly as a default behavior when answering surveys.

We expect the treatments to impact more on this repeated satisficing. To test this idea, we define a score of “bad respondents”: it is the sum of the different undesirable behaviors of a respondent that we were able to detect in the survey (see Appendix 2 for more details). This score has limits since some kinds of “bad” behaviors (e.g. random answers) could not be detected. However, by summarizing at least several of the previously studied indicators (failing the IMC, straight-lining, non-sense answers), it allows detecting respondents that repeatedly satisfice.

Table 9 gives the percentages of respondents with different scores in the scale of bad respondents. It focuses only on the respondents that answered three brands: they had much more opportunities of behaving in undesirable ways that we could detect.

Table 9: Score of bad respondents by split ballot group

ONLY if 3 brands	Control		Intro only		Intro+sign	
	%	Cumul %	%	Cumul %	%	Cumul %
Score ≤1	61.94	(61.94)	63.16	(63.16)	68.34	(68.34)
1<score ≤3	27.53	(89.47)	22.27	(85.43)	22.39	(90.73)
3<score ≤5	5.67	(95.14)	5.26	(90.69)	3.09	(93.82)
5<score	4.86	(100.0)	9.31	(100.0)	6.18	(100.0)
Mean score	2.24		2.54		2.16	
No. observations	247		247		259	

Note: Differences in means, proportions (percentages and cumulative percentages) and distributions (Kolmogorov Smirvov) between control and treatment groups are not significant. The highest score was 10.5 (maximum possible is 13.5).

Table 9 clearly shows that getting just an introduction has no positive impact on the respondents’ behaviors. This is what we saw for the separate indicators and this is what the summary variable also shows.

When the respondents have to commit themselves, even if the differences are not statistically significant, it seems there is a small positive impact of the treatment since around 7% more of the respondents almost do no behave in undesirable ways (score lower than 1). There is also a lower percentage of respondents behaving sometimes in undesirable ways but not so often (score between 1 and 5). However, there is a percentage a bit higher of respondents behaving repeatedly badly (score higher than 5).

Our interpretation is that commitment has a positive effect on the respondents that normally would answer in a not bad but not very well way. At least a part of these respondents, when they commit themselves, try to make some effort and behave better. Nonetheless, the ones that normally behave badly do not care about the introduction or the commitment and continue behaving badly.

III.6. Effort reported

Previous indicators were based on external observations of the answers. We can also look at internal auto-evaluation of respondents about the quality of their own answers measured by the amount of effort they reported: "How much effort did you put in answering this survey?", on a scale from "0-minimum effort I could" to "10-maximum effort I could". Table 10 reports the percentages of respondents for each split-ballot group that chose the different answer categories.

Table 10: Efforts reported by the respondents

ALL	Control		Intro only		Intro+sign	
Effort reported	%	Cumul %	%	Cumul %	%	Cumul %
0 minimum effort	3.15	(3.15)	2.02	(2.02)	2.61	(2.61)
1	1.67	(4.82)	2.57	(4.59)	2.23	(4.84)
2	2.78	(7.61)	2.94	(7.52)	2.05	(6.89)
3	2.60	(10.20)	3.49	(11.01)	4.28	(11.17)
4	3.34	(13.54)	3.49	(14.50)	4.66	(15.83)
5	11.87	(25.42)	8.44	(22.94)	8.75	(24.58)
6	11.69	(37.11)	9.91	(32.84)	12.48	(37.06)
7	18.18	(55.29)	21.47	(54.31)	13.41*	(50.47)
8	16.51	(71.80)	17.61	(71.93)	16.39	(66.85)
9	9.28	(81.08)	10.28	(82.20)	12.85	(79.70)
10 maximum effort	18.92	(100.0)	17.80	(100.0)	20.30	(100.0)
Mean score	6.90		6.96		7.0	
No. observations	539		545		537	

Note: Differences in means and distributions between control and treatment groups are not significant at the 95% level. The stars indicate when the differences in proportions are significant at the 95% level.

Table 10 shows that a few more respondents report the highest levels of effort (9 or 10) in the "intro+sign" group (significant at the 90% level). Nevertheless, as many respondents in this group as in the control group report a low effort made (score of 5 or less).

In the "intro only" group, a few percents less of the respondents report a score lower than five compared to the control group. But overall, respondents seem quite insensitive to the treatment as indicated by the tests of equality that cannot be rejected: a large majority of respondents does not try to make more effort when getting a treatment compared to when they did not. However, maybe a small number of respondents, who usually would have made an effort between 6 and 8, took the commitment seriously and made a higher effort, which would explain the small increase of scores 9 and 10 (significant at the 90% level).

V. Discussion

All the different analyses show that the impact of the treatment is small or even not existing. Getting only the introduction has no positive impact in general. The hypotheses *H1a* and *H1b* have to be rejected. Getting the introduction combined with the commitment has a little impact on some elements (e.g. total time of completion) but not on others (e.g. straight-lining). This means that *H1b* is supported but *H2b* only partially, depending on the indicator of quality considered. Besides, even when existing, the effects for *H2b* were small. In the end, even respondents that committed themselves did not do all the effort to answer the survey optimally.

Why does the treatments have so little or no effect at all? A first limit has been mentioned: it is the fact that even if the respondents are assigned to a treatment they are the ones deciding if they "take it" or not. According to the time spent on the introduction page, it seems that even if a majority of the respondents "took" the treatments, still, a substantial percentage did not.

Second, some results suggest that the introduction together with the commitment may affect a particular kind of respondents only: the ones that are usually answering not optimally but not badly. The sensitivity to the treatment would depend on the respondents' characteristics and usual motivation. The treatment would be effective only if usually the respondents already have a middle level of motivation.

Third, maybe the introduction used was not convincing enough. So by itself it had no positive effect. In combination with the commitment, it had some but little effect because it still did not make the respondents sufficiently realize the importance of their answers. Different texts for the introduction could be tried out in further research.

Another possible reason is that panel respondents are so used to answer surveys that they are losing all interest in what the researchers can say. Further research could look at the difference between new and experienced members of the panel to see if survey experience is moderating the effects of introduction and commitment. The fact that they are panel respondents also means that even if the "intro+sign" treatment had some effects once, if it would be repeated in each single survey that the panelists receive, the punctual effect would probably be even more limited.

In conclusion, more research is needed to clarify why the treatments had no or very little effects. But overall, it seems difficult with motivational introductory message to push online panelists to pay more attention to the questions and to retrieve and report properly their positions. Bad respondents, who most probably participate in the panel mainly in order to get the reward, continue being bad. Only some of the middle respondents may increase their effort and become very good ones, but this is not enough to guarantee an acceptable quality of the results. More radical strategies should therefore be thought of in order to face the speeding and satisficing problems.

References

Bethlehem, J. (2010). “Selection Bias in Web Surveys”, *International Statistical Review* (2010), 78, 2:161–188, doi:10.1111/j.1751-5823.2010.00112.x

Bosnjak, M. and T.L. Tuten (2003). “Prepaid and Promised Incentives in Web Surveys. An Experiment”. *Social Science Computer Review* (2003), 21, 2:208-217, doi: 10.1177/0894439303021002006

Cobanoglu, C. and N. Cobanoglu (2003). “The effect of incentives in web surveys: application and ethical considerations”, *International Journal of Market Research* (2003), 45(4): 475–488.

Conrad, F. G., Couper, M. P., Tourangeau, R., and M. Galesic (2005). “Interactive Feedback Can Improve the Quality of Responses in Web Surveys”. Paper presented at the conference of the American Association for Public Opinion Research, Miami Beach, Florida. AAPOR - ASA Section on Survey Research Methods. Available at: https://www.amstat.org/Sections/Srms/Proceedings/y2005/Files/JSM2005-000938.pdf

Cook, C., Heath, F., and R.L. Thompson (2000). “A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys”. *Educational and Psychological Measurement*, 60 (6):821–836. doi: 10.1177/00131640021970934

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

Couper, M. P., Tourangeau, R., and K. Kenyon (2004). “Picture This! Exploring Visual Effects in Web Surveys”. *Public Opinion Quarterly*, 68 (2): 255–266. doi: 10.1093/poq/nfh013

Couper, M. P., Traugott, M. W., and M.J. Lamias (2001). “Web Survey Design and Administration”. *Public Opinion Quarterly*, 65 (2): 230–253. doi: 10.1086/322199

Dillman, D.A., Tortora, R.D., and D. Bowker (1999). *Principles for Constructing Web Surveys*. Washington: SESRC. Available from http://www.isurveys.com.au/resources/ppr.pdf

Dillman, D. A., Tortora, R. D., Conradt, J., and D. Bowker (1998). “Influence of Plain vs. Fancy Design on Response Rates for Web Surveys”. In *Proceedings of the American Statistical Associations Survey Methods Research Section*. Washington, DC.

Heerwegh, D., Vanhove, T., Matthijs, K. and G. Loosveldt (2005). "The Effect of Personalization on Response Rates and Data Quality in Web Surveys*." International Journal of Social Research Methodology: Theory and Practice* 8(2):85-99. doi:10.1080/1364557042000203107

Kapelner, A., and D. Chandler (2010). “Preventing Satisficing in Online Surveys: A “Kapcha” to Ensure Higher Quality Data”. In *Proceedings of CrowdConf 2010*. October 4, 2010, San Francisco, CA. Available at: http://www.danachandler.com/files/kapcha.pdf

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

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

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

Tourangeau, R., Rips, L.J., and K. Rasinski (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Appendices

Appendix 1: text of the introduction

Group “intro only”

Important information

Please….

- We need your collaboration to be able to obtain useful and valuable results
- We need you to take your time to understand the questions of the survey
- We need you to answer the questions completely sincerely
- Your opinion will help us a lot to reach a high quality research
- If you do not have much time right now to be able to read carefully the questions and evaluate what is the best answer to each question, or if you are in a situation which makes it difficult to concentrate properly, please, just come back at a more adequate moment to complete the survey correctly
- Nicequest guarantees you that the reward you will get will always be proportional to your efforts. This is your right.

Click on the button >> in order to start

Group “intro+sign”

The respondents of this group got the same as the ones of the previous group with but one addition:

○ I have read carefully the text below and I understand the importance of mi answers for the survey, therefore I commit myself to answer the best I can
○ No, I do not want to commit myself

Appendix 2: the score of "bad respondent"

The score of "bad respondents" was computed as follow:

- 2 points if a respondent failed the IMC
- 2.5 points if a respondent straight-line on the 27 items of Q550
- 1.5 points if a respondent straight-line on 26 out of the 27 items of Q550
- 2 points if a respondent straight-line on the 26 items of Q750 or Q950
- 1 point if a respondent straight-line on 25 out of the 26 items of Q750 or Q950
- 2 points if a respondent answer a non-sense in the open narrative question ADO6
- 1 point if a respondent wrote something that was not answering the question ADO6
- 0.5 point if a respondent answer "don't know" or "nothing" in the question ADO6

The number of points added for each undesirable behavior we defined it in function of how bad we evaluated each of the behaviors. For instance, straight-lining on 26 items is worse than straight-lining on 6, so it is associated to more point. Or saying a non-sense is worse than saying something meaningful but not answering the question, so it is associated to more points.