Article

International Journal
of Market Research

International Journal of
Market Research
2018, Vol. 60(4) 352–365
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1470785318765537
journals.sagepub.com/home/mre
SAGE

Alternative methods for selecting web survey samples

Melanie Revilla
RECSM-Universitat Pompeu Fabra, Spain

Carlos Ochoa
Netquest, Spain

Abstract

Probability-based sampling is the gold standard for general population surveys. However, when interested in more specific populations (e.g., consumers of a particular brand), a lot of research uses data from non-probability-based online panels. This article investigates different ways to select a sample in an opt-in panel: without previous information, using profiling information, or using passive data from a tracker installed on the panelists' devices. Moreover, it investigates the effect of sending the survey closer to the "moment-of-truth," which is expected to reduce memory limitations in recall questions. Using additional information (profiling or passive) to select the sample leads to clear improvements in terms of levels of participation and fieldwork efficiency, but not in terms of data quality (measured by the proportion of don't know answers and the length of answers to open narrative questions) or accuracy (measured by comparing the answers to 14 questions to an external source of information). Doing the survey closer to the "moment-of-truth" further improves the fieldwork efficiency; however, there are still many challenges to implement true "in-the-moment" surveys. We also observed differences across the different samples in respondents' socio-demographic characteristics and in the survey evaluation.

Keywords

participation, passive data, profiling information, quality, survey evaluation, web surveys

Introduction

Probability-based sampling is usually considered as the gold standard to survey general population, because the probability to select any unit is known and sampling errors can be calculated

Corresponding author:
Melanie Revilla, Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra,
08003 Barcelona, Spain.
Email: melanie.revilla@hotmail.fr

(Cochran, 1977). However, using probability-based sampling is often difficult, in particular, for web surveys (Couper, 2000). In addition, a lot of research focuses on specific populations (Revilla, 2017). All in all, "probability sampling generally is not used in recruiting online panels" (Fulgoni, 2014, p. 133), but the majority of online research is based on non-probability panels (AAPOR Standards Committee, 2010).

To select respondents corresponding to the target population, filter questions are included at the beginning of the web surveys. Quotas on socio-demographic variables are often used to guarantee that the final sample has similar distributions on these variables than the target population or the general population.

This common way to select respondents for a given survey within online opt-in panels has several drawbacks. On the side of the respondents, being filtered out may be an unpleasant experience: they accepted to participate in the survey, started already answering questions and then are filtered out, and do not get the incentives that a complete participation would provide. This can be frustrating and discourage panelists to participate in follow-up surveys. If they are excluded several times, this might push them to drop out of the panel. Alternatively, respondents may lie on filter questions to get into the survey.

On the side of the panel organization, a lot of panelists may need to be invited to the survey, to get a relatively small final sample size, when a large part of the invited panelists does not match the target population. This implies costs (both time and money).

Therefore, this article investigates, in the frame of non-probability-based online panels, alternative ways to select a sample: using profiling information or using passive data from a tracker installed on the devices of the panelists.

Profiling information consists of information previously collected and stored by the panel organization, about different aspects of the panelists' life (e.g., health care, media consumption, food preferences, etc.). It usually does not correspond exactly to the definition of the target population, but is used as a proxy to increase the chances of selecting respondents with the required profile. This is already a common practice in some opt-in online panels, but still not in so many. One reason is that in many opt-in panels, panelists change very quickly.

Passive data from a tracker consist of information about all the URLs visited by the panelists on the devices (PCs, tablets, or smartphones) in which they have installed the tracker application (called "meter"), as well as the moment of the visit and the time spent on each URL. This information is used to decide who to invite to a given survey. For instance, if the target population is "people who like sport," we can select panelists who regularly visit sport-related websites.

This new idea, to our knowledge, has not been studied yet. Previous research mainly compared self-reported results from surveys with other passive data collection sources (e.g., Boase & Ling, 2013; Junco, 2013; Otten, Littenberg, & Harvey-Berino, 2010; Revilla, Ochoa, Voorend, & Loewe, 2015) and argue that to get better insights, there is a need for combining different data collection sources. Revilla, Ochoa, and Loewe (2017) suggest to reduce errors linked to the use of filter questions and to use passive data to select the sample for a survey. However, they do not provide evidence of how this combination of passive and active data performs in practice.

This article is a first step to handle this key question. It evaluates the impact on different aspects of using metered data to select the sample of a web survey, compared with a classic survey design and with a selection based on profiling information.

Moreover, the article takes into account the moment in which the survey is completed: classic surveys done at a given point in time, or surveys "in-the-moment of truth." Indeed, using passive data also allows proposing the survey to the panelists just after a specific event occurred (e.g., buying a flight ticket or seeing an advertisement). By contacting the respondents closer to the "moment-of-truth," we expect them to remember better the event and report about it more accurately. This could be a way to reduce the memory issues observed when asking panelists about their online behaviors in a classic web survey setting (Revilla et al., 2017).

The next section presents the research design used in this study. The following one provides the main results. The last one concludes.

Research design for the study

Experiment

The target population for this experiment includes all people who have visited (through PC, tablet, or smartphone) and/or purchased a flight on the website of at least one of the most common airline companies in Spain in the past 2 months: Vueling, Ryanair, Iberia, AirFrance, Norwegian, KLM, Easyjet, and AirEuropa. As any buyer is also a visitor, we work with a definition of the target population based on the less restrictive target, that is, visitors. The full questionnaire is available at http://ww2.netquest.com/respondent/glmkt/estudio_vuelos (in Spanish, as seen by the respondents) or as a Supplementary File (translation in English, word document).

It counts with a maximum of 112 questions, structured in a series of blocks:

(A) General questions about the airline companies;

(F) Main filter questions to assess which respondents have to answer Blocks B to D and questions used to select only one airline company for Blocks B to D;

(B) Questions about the last visit to the website of the airline company selected (only apply to panelists who visited at least one of the websites);

(C and D) Questions about the last flight purchased on the website of the selected airline company (only apply to panelists who bought a flight on at least one of the websites);

(E) Background variables and questions about the survey context/evaluation.

We compare the following groups:

- *Group 1. Classic survey using only filters*: Sample selected from a declarative opt-in online panel through filter questions on self-reported visits and buying to one of the airline companies' website. No previous information about the panelists is used.
- *Group 2. Classic survey using filters*[1] *and profiling information*: Sample selected using profiling information.
- *Group 3. Classic survey using filters and passive (metered) data*: Sample selected according to the information about their online behaviors collected through a meter.
- *Group 4. "In-the-next-48 hr" survey using filters and passive (metered) data*: Sample selected as in the previous group, but respondents are contacted in a maximum of 48 hr after the visit or purchase occurred.

Sending the survey in a maximum of 48 hr might already be too long to observe the benefits expected by in the "moment-of-truth" research. However, contacting panelists really "in-the-moment" by using pop-up to invite them to the survey just when they accessed the confirmation of purchase on one of the websites, only 18 persons completed the survey. In addition, there are still technological issues to be dealt with before this kind of approach can be implemented properly: for instance, the same panelist could see twice the pop-up if he or she would buy twice on one of the website during the fieldwork period. Thus, even if this might be an interesting option to develop in the future, we will not consider it in this article.

Data collection

The data collection took place in Spain, from February 23 to April 3, 2016. It was implemented by the Netquest online fieldwork company (www.netquest.com). For Group 1, any panelist could be selected. For Group 2, the profiling information used to select the respondents was based on the question, "Which companies have you flown with during the last 6 months?" We selected people who declared they flew with one of the airline companies of interest. For Groups 3 and 4, we considered panelists who have accepted to install and keep active the Wakoopa (www.wakoopa.com) tracker on at least one of their devices. For Group 3, we selected panelists who visited one of the websites of interest in the 2 months before the start of the survey. For Group 4, we reduced the time to 48 hr.

The information about the number of respondents, participation rates, and so on, is provided in the results section, as this is one of the main aspects that we wanted to study.

Research question and hypotheses

Using the experimental design and data just described, we study to what extent the four samples differ at several levels. Our hypotheses for each of these aspects are presented below:

> *Participation and fieldwork efficiency*: Using profiling or metered data in selecting the sample will increase the participation rate (*H1a*) and improve the fieldwork efficiency (*H2a*), by reducing the number of panelists contacted to achieve the number of complete interviews required. On the contrary, we do not expect that inviting respondents in the next 48 hr after their visit of the website will affect participation (*H1b*) and efficiency (*H2b*).
>
> *Respondents' profile*: The different ways to select the samples will not change the respondents' profile (*H3*). We consider both socio-demographic variables and variables related to the brand recognition. We also compared the declared prices for respondents who booked a flight. We should note that we will not be able to test which sample is the closest to the true target population but only to compare the different samples with each other.
>
> *Data quality*: Using metered data to select the sample will improve data quality (*H4a*), that is, it will reduce the proportion of "don't know" (DK) answers and increase the length of answers to open narrative questions. Indeed, by pre-selecting respondents with the profile of interest, we expect to reduce the proportion of respondents who passed the filter questions but do not correspond to the target ("fraudulent" respondents who lie or do not remember well). Moreover, by reducing the time which separates the survey from the event, we expect less memory issues, so sending the survey closer to "the moment-of-truth" will improve the quality even further (*H4b*).
>
> *Data accuracy*: The data accuracy (measured by the distance between survey answers and an objective source of information) will be improved in Groups 2 and 3 compared with Group 1 because of the reduction in "fraudulent" respondents (*H5a*). It will be improved even more in Group 4 (*H5b*) because of the smaller time gap between the visit and the questions about this visit.

Survey evaluation: No differences are expected between the first three groups (*H6a*). However, we expect the survey evaluation to improve when the survey is sent in the next 48 hr (*H6b*) because the questions will relate to a recent event, making it more relevant and easier to answer.

Main results

Survey participation

From the respondents invited, some decide to participate and start the survey. Because in the online panel used for this study (Netquest) the invitations to participate in a new survey do not inform about the survey characteristics (e.g., topic or expected length), the decision to start the survey can be considered as an indicator of the panelists' engagement.

There are mainly two types of invalid participations: screened out and quota full. Screened-out participations correspond to respondents excluded from the survey because they did not match the profile (e.g., they did not visit any of the eight airline companies' websites in the past 2 months). Quota-full participations occur when the target goal for a particular quota used in the project is exceeded. In this study, the different sample selection methods do not generate differences in the amount of quota full, so quota-full participations are not considered in our analyses. If respondents are not screened out or excluded because of quota full, they can finish the survey or voluntary abandon it at some point.

The number of respondents who finished the survey in anyway divided by the ones invited to participate to the survey is called Participation Rate:

$$Participation\ Rate = \frac{Finished\ anyway}{Invited}$$

where

$$\begin{aligned} Finished\ anyway = {} & Finished\ in\ target + \\ & Excluded\, screened\ out + Excluded\ quota\ full \end{aligned}$$

The number of respondents who voluntarily abandon the survey divided by the number of respondents who started it is called Dropout Rate. Usually, the Dropout Rate is affected by the survey characteristics. In this experiment, the survey being the same for all groups, this cannot play a role. Nevertheless, it can be related to the panelists' engagement: loyal panelists might resist better hard-to-complete surveys.

Because Netquest invites its panelists to install the meter only once they have completed at least 10 surveys, we expect the panelists included in the different sample groups to differ in terms of panel engagement. Metered panelists are, on average, more experienced and more prone to participate than non-metered panelists. Therefore, we expect higher Participation Rate for metered panelists, independent of the sample selection method used.

Thus, to really assess the effect of the sample selection method, we compare the Participation Rate with the Expected Participation Rate, an internal measure developed by Netquest to estimate the likelihood that a group of panelists participate in their next survey invitation. The Expected Participation Rate is calculated by Netquest as the percentage of panelists in the sample that

Table 1. Participation by group.

	Group 1 (No info)	Group 2 (Profiling)	Group 3 (Meter)	Group 4 (Next 48 hr)
Number of invited	1,450	888	861	287
Number of started	740	685	799	231
Number of finished in anyway	695	666	792	228
Participation Rate (%)	47.9	75	92	79.4
Expected Participation Rate (%)	37.1	52	55.1	65.9
Variation (Participation Rate/ Expected Participation Rate – 1) (%)	+29.1	+44.2	+66.9	+20.5
Dropout Rate (%)	6.1	2.8	0.9	1.3

participated in the last survey they were invited to, slightly reduced (around −3%) to take into account the average panel attrition. Practice shows that this is a good predictor of the willingness to participate in the next survey of a group of panelists. Given that the last survey is different for each panelist, this takes into account that some surveys are harder to complete than others.

If the Participation Rate of a particular survey matches the Expected Participation Rate, this survey is similar to the average survey that the panelists are used to complete. If the Participation Rate is higher than the Expected Participation Rate, the survey has a higher participation than average. On the contrary, if it is lower, the survey has a lower participation than average.

For each of the four groups, the total number of panelists at different steps of the process, as well as the participation rates, the variation between them (Participation Rate/Expected Participation Rate – 1) and the Dropout Rate, are reported in Table 1.

First, the Participation Rate is larger than the Expected Participation Rate for all four groups. If we consider Group 1 (no previous information), the +29.1% variation between both participation rates can only be caused by the survey features: short survey length (median time around 10 min), simplicity of the questions formats, and longer than average fieldwork period.

Second, Group 2 (profiling information) and Group 3 (passive data) are both composed by people who are expected in advance to match the target population. Both groups show higher differences between participation rates than Group 1, the difference being even much higher for Group 3 (+66.9% vs +44.2%). Using profiling information improves significantly the survey participation, and using passive data even more. In addition, using profiling or behavioral data produces lower dropout rates.

Finally, for Group 4 (passive data; next 48 hr), the variation between Participation Rate and Expected Participation Rate is the lowest. The Dropout Rate is quite low, but not lower than in Group 3. This may suggest that inviting people close to the "moment-of-truth" is not very well perceived by respondents. It might also require longer fieldwork periods.

Fieldwork efficiency

Each access panel has a limited capacity in terms of producing participations in surveys. This capacity over a period of time is the number of real respondents of the panel, times the maximum number of survey invitations that the panel is willing to send per panelist, knowing that over-invitation may have consequences on data quality. Thus, getting as many valid participations as possible from the same number of survey invitations is a key aspect in developing market research projects through online access panels.

Table 2. Number of respondents who finished in different ways and Incidence per group.

	Group 1 (No info)	Group 2 (Profiling)	Group 3 (Meter)	Group 4 (Next 48 hr)
Number of finished in anyway	695	666	792	228
Number excluded for quota full	117	412	114	1
Number screened out (i.e., excluded because they did not visit any website)	243	42	122	25
Number of finished in target (i.e., who have visited at least one website)	335	212	556	202
Incidence (%)	58.0	83.5	82.0	89.0

Number of finished anyway is the sum of the three others.

For this study, we define the Fieldwork Efficiency of the online data collection as the number of valid participations (respondents matching the target population who finished the survey) divided by the number of respondents who finished the survey in anyway:

$$Fieldwork\ Efficiency = \frac{Finished\ in\ \ target}{Finished\ anyway}$$

As quota-full participations are not playing a role in this study, the Fieldwork Efficiency can be considered as equivalent to the percentage of participants that match the profile of interest, also called Incidence:

$$Incidence = \frac{Finished\ in\ \ target}{(Finished\ anyway - Excluded\ quota\ full)}$$

The number of valid and invalid participations, together with the Incidence, is presented in Table 2.

As we expected, the Incidence for Group 1 is the lowest (58.0%), whereas Groups 2 and 3 get very similar levels of Incidence (83.5% and 82.0%). When sending the survey invitation in the next 48 hr after the visit of the website (Group 4), the Incidence is even higher (89.0%). However, it does not reach 100%, which can be due to both errors in the declared visits (e.g., a panelist has visited the website, but she or he reports that she or he did not because she or he does not remember) and errors in the passive data (e.g., because of shared devices, we think that a panelist has visited a website when in reality another person did so using the device on which the meter is installed). The higher Incidence for Group 4 may be then due to panelists remembering better if they visited the website and thus reporting about it more accurately in the survey, when the time gap between the website visit and the survey is reduced.

Overall, the results indicate that the Incidence can be improved by using additional sources of information and getting closer to the moment-of-truth. In addition, these results have been obtained for a quite common online activity. But for low Incidence studies, in which a really rare online activity must be studied, the use of behavioral data has potential to improve even more the Incidence. The same applies for activities that even the participants ignore they have made, but are easy to check through their navigation (e.g., exposition to an online ad or visiting a temporal beta version of a new website).

Respondents' profile

Even if different sample selections are used, we do not expect large differences in the respondents' profile, once focusing on the respondents who, after the filter questions, are selected to continue answering the survey. To check whether this is indeed the case, we consider socio-demographic and attitudinal variables (Table 3). For most variables, we present only the proportions for a given category (for instance, for education, only tertiary level, or for life satisfaction, only those who answered 4 or 5 on a 5-point scale from *totally unsatisfied* to *totally satisfied*). However, for the chi-square tests of independence between the groups and each of the other variables (Age, . . ., Life Satisfaction), we take into account all categories available for these other variables. The last column of Table 3 reports the *p*-values associated with each of these tests.

According to the chi-square test, there is a significant relationship between the group and age, gender, education, and household size. Looking at the proportions, we observe large differences: for instance, the proportion of men varies from 35.8% (Group 2) to 50.7% (Group 1). Thus, the way to select the sample seems to affect also the respondents' socio-demographic profiles, even if there is no clear pattern of one group being always more different from the others.

To control for these differences in respondents' socio-demographic profile, we could use quotas to make sure to get similar proportions on key background variables, or we could use post-stratification weighting procedures. In fact, this is a typical problem for opt-in online panels, which normally use quotas and/or post-stratification weights at least for basic socio-demographic variables.

For political orientation and life satisfaction, which are more attitudinal variables, the differences are minimal. The chi-square tests suggest there is no relationship with the group. On the contrary, six out of eight variables about brand recognition show a significant relationship with the group. The median of the declared prices is higher in the meter group compared with the three others. However, the *p*-value of a Kruskal–Wallis test between the price and the groups is higher than .05.

Data quality

We consider two indicators of data quality: the proportions of "DK" answers (higher proportions suggest lower quality) and the number of characters written in open narrative questions (higher numbers suggest higher quality).

However, previous research suggests that data quality can also be affected by the device used to answer the survey, finding, for instance, higher missing data rates (Lugtig & Toepoel, 2016; Struminskaya, Weyandt, & Bosnjak, 2015) and shorter answers to open questions for mobile than for PC respondents (Mavletova, 2013; Revilla & Ochoa, 2016). Because there are significantly more PC respondents in Groups 3 and 4 compared with Groups 1 and 2 (see Appendix 1), we analyzed the results for all devices together but also separately for PC and for smartphone respondents. We did not consider "only tablet respondents" because of the small number of observations (N<30 in all groups).

DK answers. For these analyses, we consider all the questions after the filters (after Block F) where a DK option is available. For these 20 questions, we compare the percentages of respondents who received the questions and selected DK, by groups, altogether and separating PC and smartphone respondents. Some significant differences are observed across groups, but there is not a clear tendency for one group to present a lower or higher proportion of DK.

To test further whether the sample selection method affects the level of DK answers, we run a linear regression (Table 4) where the percentage of DK is explained by the experimental group, the device *(*PC/ Smartphone) and the block in which the question is asked (B, C, or D).

Table 3. Respondents' profile: proportions of respondents with different characteristics.

	Group 1 (No info)	Group 2 (Profiling)	Group 3 (Meter)	Group 4 (Next 48 hr)	*P*-value (χ^2 test for all except the price: Kruskal–Wallis test)
Age (years)					
<24	28.7	18.9	17.6	18.3	<.01
25–34	20.0	35.4	24.5	28.2	
35–44	17.9	22.2	21.4	22.3	
45–54	14.6	16.0	15.5	16.3	
55+	18.8	7.5	21.0	14.8	
Gender					
Male	50.7	35.8	50.5	43.6	<.01
Education					
Tertiary	51.0	65.6	64.6	58.9	<.01
In couple					
Yes	74.0	71.2	78.4	73.8	.15
Level of urbanization					
Big city	31.0	37.3	36.0	40.1	.82
Household size					
1–2 persons	36.5	53.7	42.2	40.6	<.01
Political orientation					
Right	8.4	8.5	10.4	12.4	.07
Life satisfaction					
4 or 5	63.9	65.1	65.5	64.4	.08
Knowledge of brands					
Vueling: Yes	92.8	96.2	96.4	95.5	.09
Ryanair: Yes	96.1	99.1	98.6	99.0	.02
Iberia: Yes	98.8	100.0	99.5	99.5	.35
Easyjet: Yes	73.1	89.1	86.9	88.1	<.01
AirEuropa: Yes	87.5	92.0	93.0	92.6	.03
AirFrance: Yes	78.2	85.8	86.1	87.6	<.01
KLM: Yes	43.3	56.6	62.9	61.9	<.01
Norwegian Airline: Yes	35.8	54.2	52.5	55.4	<.01
Price spent for the booking (median)	110	110	140	116	.26

The number of observations is 335 for Group 1, 212 for Group 2, 556 for Group 3, and 202 for Group 4, except for the price paid (where it is, respectively, 85, 89, 263, and 98). Number of categories taken into account for the chi-square test: 6 for age, 2 for gender, 9 for education, 2 for in couple, 5 for level of urbanization, 6 for household size, 5 for political orientation and for happy, and 2 for the brands known.

Only the block in which the question was asked has a significant effect on the proportions of DK. None of the groups' coefficient, on the contrary, is significantly different from zero effect. Thus, there is no support for our hypotheses H4a/b: there is no significant reduction in the proportion of DK in the groups using additional information to select the samples.

Table 4. Regression of DK on the groups, device, and block in which the questions appear.

DK	Coefficient	*p*-value
Group 2 (Profiling)	–0.382	.88
Group 3 (Meter)	–2.747	.29
Group 4 (Next 48 hr)	–2.865	.27
Smartphone	1.335	.47
Block C	9.679	<.01
Block D	31.887	<.01
Constant	2.927	.33
Adjusted R^2	.543	

DK: don't know.
Reference categories are Group 1 (No info), PC, and Block B.
If we change the reference category, the results are similar.

Table 5. Length of answers to open questions (median number of characters typed).

Questions	All					PCs					Smartphones				
	G1	G2	G3	G4	*p*	G1	G2	G3	G4	*p*	G1	G2	G3	G4	*p*
B3	21	18	24	22	<.01	23	20	25	22	.56	18	17	21	22	.04
C2	105	104	127	122	.03	112	124	142	143	.14	84	90	96	91	.60
C5	44	50	42	40	.43	50	67	44	45	.03	38	41	38	25	.23
C6	19	20	23	18	.49	24	30	25	20	.59	15	14	20	16	.12

G1–G4 stands for Group 1 to Group 4. B3–C6 are the questions names. *p* corresponds to the *p*-value of the Kruskal–Wallis test of significant differences between Groups 1 and 4.

Length of answers to open narrative questions

Next, we consider the length of answers to all four open questions available after the main filters. Table 5 presents the median number of characters for each question and group and the *p*-values corresponding to non-parametrical Kruskal–Wallis tests.

In line with previous research, we can see that overall PC respondents type more characters. However, there is little evidence that one experimental group provides, in general, longer answers than others. The Kruskal–Wallis tests indicate significant differences for questions B3 and C2 but not for C5 and C6. Looking at the medians, we observe longer responses for Groups 3 and 4 for question C2, which is also the question generating the longer answers.[2] Thus, maybe the sample selection method plays a role only for really demanding open questions, but not for the ones requiring shorter answers. Further research would be needed to investigate this idea.

Data accuracy

Studying accuracy of answers is usually done by comparing survey answers to a more objective source of information (see, for example, de Nicola and Giné, 2014; Pursey, Burrows, Stanwell, & Collins, 2014). Thus, to study the accuracy, we considered 14 questions about the last flight purchase (Block D) for which we could check the true value ourselves by looking at the airline companies' websites, so we could validate the answers. These questions asked to the respondents if different options were offered to them during their purchase on the given website (e.g., free traveling for babies up to 3 years old, renting a car, or buying extra large seats).

Table 6. Mean number of accurate answers out of 14.

	All				PCs				Smartphones			
	G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4
Mean no. of accurate answers	5.6	6.3	5.9	6.2	5.6	6.5	6.0	6.2	5.6	5.8	6.1	6.4
t-test: *p* < .05	Only G1 vs G2				Only G1 vs G2				None			
N	151	125	338	123	89	66	249	96	45	48	67	21

Table 7. Survey evaluation by group and chi-square test.

		Proportions resp. answering 4 or 5				χ^2 test
		Group 1 (No info)	Group 2 (Profiling)	Group 3 (Meter)	Group 4 (Next 48 hr)	*p*-value
Easy	All	79.7	82.1	90.5	90.1	<.01
	PC	81.5	82.1	91.6	90.4	<.01
	Smartphone	78.8	80.7	87.6	89.2	.12
Liked	All	60.6	53.8	80.0	76.2	<.01
	PC	61.0	62.5	80.1	77.7	<.01
	Smartphone	61.1	41.0	81.0	67.6	<.01

We counted the total number of questions for which a respondent answer was in line with what we observed on the website: this is our measure of accuracy.

The means per group of the total number of accurate answers (out of 14) are reported in Table 6, together with which *t*-tests had a *p*-value lower than .05.

Overall, the accuracy of the answers is very similar in all groups. We only found one significant difference between Groups 1 and 2 (both when considering all respondents together or only the PC respondents), Group 2 showing a higher data accuracy compared with Group 1. However, we did not find any significant differences for Group 4, even if this is where we mainly expected them.

We can think about two possible reasons for this result. First, 48 hr is already too long for respondents to remember well their experience, and we should ask them even closer to the "moment-of-truth" to improve the data accuracy. Second, people may not have paid attention to which options were available if they were not interested in them: for instance, if they are not planning to travel with children, they may not realize that there is something related to that on the webpage, mainly if they are not forced to take any action concerning it (e.g., not forced to say "0 children" explicitly).

Survey evaluation

Finally, we studied how respondents evaluated the survey across groups, by comparing the answers to the two following questions:

- "***Easy***": Please indicate how easy or difficult you found it to answer to this survey. (5-point scale from *extremely difficult* to *extremely easy*)
- "***Liked***": And to what extent did you like to answer this survey? (5-point scale from *didn't like it at all* to *totally liked it*)

The proportions of respondents choosing the answer categories 4 or 5 for each of these questions are reported in Table 7. Groups 3 and 4 present higher proportions than Groups 1 and 2, for both questions, and considering all respondents together, only PC, or only smartphone respondents.

A chi-square test is used to see whether there is a statistically significant relationship between the group and each of these two questions measured on the 5-point scales (Table 7 too). A statistically significant relationship between the group and "***Easy***" is found when considering all devices, or only PCs, but not when considering only smartphones. A statistically significant relationship between the group and "***Liked***" is found in the three cases.

Thus, these results indicate that the survey evaluation is linked to the group. Nevertheless, the differences seem to be between groups selected using the passive data and the others. This might be linked to the selection of the panelists to be part of the metered panel (only experienced panelists are invited, which could explain that they found it easier to answer). Besides, metered panelists are in the panel for a longer time, suggesting that they like answering surveys (more than panelists dropping-out of the panel after a few participations only). The fact that the survey is completed closer to the event of interest, on the contrary, does not improve the survey evaluation. Again, this might be because 48 hr is already too long.

Discussion/Conclusion

This article used data from the Netquest panels in Spain to study the effect, at different levels, of using additional sources of information to select the sample to which a web survey was sent. This is done by comparing a classic sample selection with no additional information (Group 1) with a sample selected using profiling information (Group 2), and two samples selected using passive data from a meter, one based on the activities of the last two months (Group 3) and the other of the last 48 hr (Group 4). These four types of samples are compared in terms of participation to a short survey about airline companies, fieldwork efficiency, respondents' profile, data quality, data accuracy, and survey evaluation.

Main results

Participation: In comparison with Group 1, the variation between the Participation Rate and the Expected Participation Rate is increased and Dropout Rates are reduced when using profiling data and even more when using passive data to select the sample (H1a supported). However, doing the survey in the next 48 hr does not improve further (H1b not supported).

Fieldwork efficiency: The Incidence improved by over 25% when using profiling or passive data for the sample selection compared with classic surveying (H2a supported). Doing the survey in the next 48 hr allows improving it even further (H2b supported).

Respondents' profile: There are differences across all four groups in terms of the main sociodemographic variables (H3 not supported), but not in terms of attitudinal variables. Using quotas or some weighting techniques to control for these differences might help. But if there are unobserved variables related to the topic of interest, this may be a problem for the representativeness. Also, we do not know which group is closer to the population of interest. To study this, we would need information about the population of users of at least one of the eight airline companies studied in this article.

Data quality and accuracy: There is little support of improvements in data quality or accuracy (i.e., little support for H4a/b and H5a/b), as measured in this study. However, the new sample selection methods do not harm data quality/accuracy either (again, the way they are measured here).

Survey evaluation: Both groups selected using passive data show a more positive survey evaluation (H6a only partially supported). This suggests that metered panelists have in general a more positive attitude toward surveys, more than it suggests support for the mechanism behind our hypothesis H6b.

Limits

One limit of this study stands in the indicators available to measured data quality (DK and open questions answers length). Being able to use other measures (more precise ones, for example, reliability and validity estimates) could lead to different results. The same is true for data accuracy. It would have been really interesting to be able to check the information about the price of the purchase or the destination and time of the flights. Further research could investigate whether by using alternative indicators of data quality and accuracy improvements are achieved, in particular for in-the-moment surveys.

In addition, in this study, the visitors of an airline company website were a relatively easy-to-reach target population. This could partly explain why we do not observe as much improvements as expected. In the case of hard-to-reach populations, the improvements might be clearer.

Implications for practice

Overall, using additional information from profiling or passive data seems recommendable as it improves the participation and fieldwork efficiency without hurting the data quality (measured by the proportion of DK answers and the length of answers to open narrative questions) or accuracy (measured by comparing the answers to 14 questions to an external source of information). However, they might be some problems in terms of representativeness which need to be studied further. Moreover, more research would be needed to test what happens for different indicators in particular of quality/accuracy and to check the robustness of the results in different contexts (different target populations, different countries, etc.).

Trying to contact the panelists in the next 48 hr after the event of interest, on the contrary, does not seem really worth it. It only improves the Incidence, but it also makes the fieldwork longer and more complex to handle for a survey agency. Nevertheless, maybe by doing the research really in-the-moment, the improvements would be larger.

As mentioned, we tried this, but only managed to get 18 responses, which is the reason why we did not analyze these results in the article. This suggests that doing the research really in-the-moment is still really challenging and requires some additional efforts. Nevertheless, this could be the key to improve the data quality and accuracy.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The filter questions are always asked, but in addition we use the profiling or passive data.
2. This question asked respondents to describe the flight they bought. They were instructed to give as much details as possible (from where to where, when, kind of luggage . . .). We should notice that many respondents in this question wrote the maximum number of authorized characters, suggesting that we used a limit that was too low.

Supplementary material

Supplementary material for this article is available Online.

References

AAPOR Standards Committee. (2010) AAPOR report on online panels (AAPOR executive council). Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOROnlinePanelsTFReportFinalRevised1.pdf

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication*, *18*, 508–519.

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Couper, M. P. (2000). Web surveys: A review of issues and approach. *Public Opinion Quarterly*, *64*, 464–494.

de Nicola, F., & Giné, X. (2016). How accurate are recall data? Evidence from coastal India. *Journal of Development Economics*, *106*, 52–65.

Fulgoni, G. (2014). Uses and misuses of online-survey panels in digital research. *Journal of Advertising Research*, *54*, 133–137.

Junco, R. (2013). Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior*, *29*, 626–631.

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, *34*, 78–94.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*, 725–743.

Otten, J. J., Littenberg, B., & Harvey-Berino, J. R. (2010). Relationship between self-report and an objective measure of television-viewing time in adults. *Obesity*, *18*, 1273–1275.

Pursey, K., Burrows, T. L., Stanwell, P., & Collins, C. E. (2014). How accurate is web-based self-reported height, weight, and body mass index in young adults? *Journal of Medical Internet Research*, *16*, e4.

Revilla, M. (2017). Analyzing the survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *Methods, Data, Analyses*, *11*, 1–28. doi:10.12758/mda.2017.xx

Revilla, M., & Ochoa, C. (2016). Open narrative questions in PC and smartphones: Is the device playing a role? *Quality & Quantity*, *50*, 2495–2513.

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. *Social Science Computer Review*, *35*, 521–536. doi:10.1177/0894439316638457

Revilla, M., Ochoa, C., Voorend, R., & Loewe, G. (2015). When should we ask, when should we measure? In *Proceedings of the ESOMAR world Dublin congress–Revelations*. Retrieved from: https://www.esomar.org/knowledge-center/library?conference=174

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality: Evidence from a probability-based general population panel. *Methods, Data, Analyses*, *9*, 261–292.

Appendix 1. Proportions (%) of respondents who answered through different devices.

	Group 1 (No info)	Group 2 (Profiling)	Group 3 (Meter)	Group 4 (Next 48 hr)
PCs	58.2	52.8	72.5	77.7
Tablets	6.6	7.1	4.7	3.5
Smartphones	33.7	39.2	21.8	18.3
Others	1.5	0.9	1.1	0.5
Number total of observations per group	335	212	556	202