An exploratory study of the participation in the online surveys using access panels

Which are the profiles of the non-participants? What are the influences of the revival and the time to respond?

*Philippe **JOURDAN***
Assistant professor
IUT Evry
26 rue Richer
75009 Paris
Email. : philippe.jourdan5@wanadoo.fr

*Valérie **JOURDAN***
Chief Executive Officer
Panel On The Web SA
36 rue Vivienne
75002 Paris
Email. : valerie.jourdan@panelontheweb.com

1)- INTRODUCTION

During the last ten years, the evolution of ad hoc studies has been marked by the rise of the surveys carried out on access panels. Initially confined to longitudinal studies requiring some samples of significant sizes of individuals questioned at regular frequencies on their purchase or spending patterns, the studies on access panels have today extended to brand image tracking, to the tracking of advertising campaigns, to the lifestyles and values surveys and even to the opinions, attitudes or behaviors inquiries. Apart from the longitudinal studies or from those confined to the U&A inquiries, the access panels are also used to conduct a widespread range of surveys from opinion polls to marketing-mix optimization surveys: screening of concepts, concepts tests, product tests, etc. The access panels are also used in both the FMCG (B to C) and the durable (B to B) sectors.

By definition, an access panel is a permanent sample of individuals recruited on the initiative of a market research firm and who volunteer to take part in ad hoc surveys, generally conducted by telephone, in face-to-face or online, etc. A fraction of the total panel can be easily sampled and questioned on a specific issue: thus the recourse to an access panel makes it possible to have a pre-qualified and a pre-recruited population, easy to mobilize in a short time; moreover, the voluntary nature of the registration in addition to the incentive associated with the participation in the surveys ensure a higher rate of complete questionnaires than the one usually obtained in samples recruited for ad hoc purposes.

The development of the Internet caused the appearance of many online access panels: by simplifying the recruitment and the surveys' data gathering (online surveys are self-administered by nature), the online access panels dispose of great advantages such as a budget optimization thanks to the reduction of the data-gathering costs; an increase in the size of the survey samples (some access panels gathering several hundreds of thousands of Net surfers); a high qualification of the samples allowing the questioning of rare targets and a better geographical dispersion of the respondents.

However, these advantages come with new risks of distortion of surveyed data. Firstly, online access panels have been criticized for their low representativeness of a large population and this because of the weak rate of penetration of the media in the households and the unequal sharing of the access to the Web with regards to the dwellings' areas, the agglomeration size, the gender, the age or the socio-professional categories. Although these criticisms were legitimate in the recent past, they should logically grow softer because of the acceleration of the Internet equipment in French households: in the 4th quarter 2003, 6.9 million French households which constitute 27.7% of the households and 65% of the microcomputer-equipped households have an access to the Web from their home (*Médiamétrie, 2003*). In the meantime, another factor of distortion is often neglected: the distortion caused by the non-respondents, which because of the voluntary step of the registration to an access panel possesses specific characteristics.

The objective of the research is to explore which are the variables that are correlated with the total non-response and to measure the effectiveness of the methods used to increase the response rate in the case of an online B to C access panel. First of all we will present the determinants and the consequences of the non-response in the surveys by questionnaire. We will then specify the methodology of the research and present the exploratory results obtained.

2)- THE NON-RESPONSE IN THE SURVEYS BY QUESTIONNAIRE

A)- THE VARIOUS FORMS OF THE NON-RESPONSE

As Lebart (*Lebart, 2001*) very rightfully underlines it, non-response must be regarded by the statistician as a meta-information insofar as it is advisable to distinguish the total non-response related to the impossibility of contacting the respondent, from that, partial or total, that accounts for refusals, the incomprehension or the lack of knowledge of the answer ("does not know") or even the fact that the respondent is not concerned with the question ("not applicable").

In the general case, it is thus advisable to distinguish several origins of the non-response as it is underlined in the following non-exhaustive list:

- The technical failure: (wrong telephone number, incorrect postal addresses or wrong electronic address) or physical impossibility to join the contacted person (absence of home, refusal to pick-up the phone, little-used electronic mailbox, etc.).
- The refusal to answer which results in a total or partial non-response (abortion of the questionnaire, unanswered questions in self-administered questionnaires, questions considered to be too intrusive or disturbing, etc).
- The absence of response which is accounted for by a wrong comprehension of the question or an ignorance of the response by the questioned person in a self-administered questionnaire. This non-response is usually punched as a “does not know" answer).
- The involuntary omission of the response due to an awkwardness of the interviewer (omitted question, answer unreported, incomplete observation, etc.) or due to the respondent himself in the case of a self-administered questionnaire. The respect of the instructions, the quality of the training prior to the launch of the field work or even the clarity and the legibility of the questionnaire limit -without eliminating it completely- the consequences of this type of non-response.
- The deletion of the answer during the examination of the data: it is generally an involuntary or accidental deletion resulting from a bad transcription, a wrong coding or an error in the data-punching.
- The deliberate non-response because the question isn’t relevant for a particular subsample (e.g.: a filtered question addressing men only).

Only the first five sources of non-responses are likely to introduce bias into the results of the survey, especially in the case of randomly-generated samples. Indeed, the random selection of a sample within a finite population rests on the principle of the assignment for each individual of a non-null probability of being questioned (*Kalton, 1993*). Now, it is unlikely that the incapacity of joining the person, the refusal to answer the questions, the incomprehension of certain questions, the involuntary omission of the response, and the deletion of the answer be randomly spread among the sample; consequently the equiprobability of the observations within the datafile is not guaranteed. Thus, the non-response has an impact on the reliability of the estimates.

In the particular case of the surveys carried on a Net surfers' access panel, two characteristics of the data collection process exert an influence on the expected non-response rate:

- Firstly, as in all access panels, the respondents are pre-recruited and consequently wilful to take part in studies: the rate of participation in the subsequent investigations is thus usually high (between 60% and 80%). However, this rate of participation should not be confused with a rate of answer. Indeed, to compare the rate of response of an access panel survey with that of a telephone survey, it is advisable to take account of the rate of refusal during the recruitment of the respondents. This information is hardly conveyed by the majority of the institutes. In practice, it consists in taking into account those who accept (or refuse) the principle of adhesion out of 100 people randomly selected to take part in a panel. The actual non-response rate (R_{nr}) to an access panel survey is thus calculated from the rate of adhesion (R_{adh}) and the rate of participation (R_{part}) according to the following formula:

(1) $$R_{nr} = [1 - (R_{adh} \text{ x } R_{part})]$$

Provided the recruitment of the access panel be carried out starting from a random selection of individuals belonging to a finite population of reference and if it is admitted that a third of the selected persons agree to adhere to the panel and that they will be 7 out of 10 to take part in the first investigation, the actual rate of answer only amounts to: 33% x 70% $\cong$ 23%!

- Secondly, certain technical reasons of aborts are obviously specific to the media. In the majority of the online access panels, the Net surfer is informed of an ongoing survey by an e-mail sent to his last registered electronic address. However the Net surfers are frequently brought to change electronic addresses for various reasons: a change of Internet provider, a desire for several addresses dedicated to distinct uses (business, personal, e-shopping addresses, etc), a change of status implying a change of email address (students, members of associations, professionals, etc), a manifest will to escape advertising harassment ("Spam"), etc. Although few statistics are published on the subject, the renewal of email addresses would be approximately 10% per year in France. It is thus advisable to distinguish two major causes within the rate of non-participation: one is a technical cause, with the impossibility to join the respondent whatever the reason is (deactivated or invalid addresses, badly typed adresses or little or never consulted mail boxes, etc.), and the other is a real refusal to answer. Consequently, the formula (1) can be detailed as follows:

(2) $$R_{nr} = [1 - (R_{adh} \times (1 - R_{inv}) \times (1 - R_r))]$$

With : R_{nr} = rate of non-reponse
R_{adh} = rate of panel adhesion
R_{inv} = rate of invalid addresses
R_r= rate of refusal to answer

B)- THE DETERMINANTS OF THE NON-RESPONSE

What are the principal origins of the non-response? There are many works on the subject that are advisable to the interested reader (*Madow and Al, 1983; Schafer, 1997; Little and Rubin, 2002*). Of all the causes of the non-response which we have identified, the most worrisome to the researcher is naturally the refusal to answer that Cochran (1977) qualifies as "refusal to cooperate", be it for any reason: lack of time, intimate or private character of the questions, fear to reveal one's opinion, etc. The proportion of people refusing to answer a questionnaire is all the more critical as it is often correlated with the object of the study (*Wiseman and McDonald, 1980*), with the length of the questionnaire, its complexity, its structuring, the mode of data-gathering, the granted remuneration (*Brennan, Hoek and Astridge, 1992*) and the recruiting skills of the interviewers (*O' Muircheartaigh and Campanelli, 1999)*, these causes being potentially likely to depreciate the internal validity of the data-gathering. Lastly, many researches have proven for a long time that the refusal to answer is not equally distributed among the population: the gender, the age, the education level, the income are correlated with the rate of non-response (*Chen, 1996, Marjorie, 1960, Ferber, 1948, etc.*). These researchs however relate to postal or telephone surveys, as there are few recent articles dealing with the determinants of the refusal to answer in online surveys (*cf. Bosnjak and Tuten, 2001*).

In many cases, the market research institutes do not apply any particular procedure for the handling of the non-response: either the refusals to answer are deemed marginal, or the implicit assumption is that the distribution of the answers is identical among the respondents and non-respondents alike (be it a partial or total non-response). In practice, it is unlikely that this assertion be checked. Litlle and Rubin (1987) suggested a classification of the non-responses into 3 families: in the first case, the non-response is distributed in a purely random way, namely that its distribution does not depend on any other studied variable; in the second case, the non-response is dependent on the observed values; finally in the third case, the non-response is dependent on the missing or not-observed values. In practice this third case is the most difficult to deal with from a statistical point of view since the estimate of the non-response can precisely be neared only by using missing variables (e.g.: the upper classes are generally reticent to take part in studies on income, investment or saving).

C)- THE CONSEQUENCES OF THE NON-RESPONSE

What are the consequences of the non-response on the reliability of the results of the studies? Let us suppose that we seek to estimate the value of the parameter μ in the population: the mean $\overline{\mu}$ in the population is the sum of the two means $\overline{\mu}_r$ (the mean observed from the respondents) and $\overline{\mu}_{nr}$ (the mean estimated for the non-respondents), all that being balanced by the relative weight of each group ω_r and ω_{nr} (cf. formula 3). A simple transformation makes it possible to evaluate the estimation error related to the exclusive accounting of the responses to the survey, that is to say ($\overline{\mu}_r$ - $\overline{\mu}$) according to the other parameters (cf. formula 4).

$$(3) \qquad \overline{\mu} = \omega_r \,.\, \overline{\mu}_r + \omega_{nr} \,.\, \overline{\mu}_{nr} \qquad\qquad (4) \qquad \overline{\mu_r} - \overline{\mu} = \omega_{nr} \,.\, \left(\mu_r - \overline{\mu}_{nr}\right)$$

The estimation error is thus a function of the proportion of non-respondents in the survey and of the variation of the answers between respondents and non-respondents (*Kalton, 1983*). It is thus illusory to think that this difference is negligible or even that it is equally distributed in all the segments of the population.

In all the cases, the first precaution consists in minimizing the non-response rate. In the case of a telephone study a common practice consists in multiplying the call-back at various hours during the course of the day or on various days of the week or even using different data-gathering methods one after another (telephone, mailing, face-to-face, etc.); this last way of proceeding however emphasizes the intricate problem of the distortion of the answers related to the variation of the methods of collection (*Frankel and Frankel, 1977*).

In the case of partial non-responses, certain statistical methodologies make it possible to estimate the missing values from the observed values. When the non-responses concern a restricted number of respondents or only a few questions, the statistical softwares propose several methods for estimating the missing values: these methods are based on the knowledge of the values taken by other filled-in variables of the questionnaire which are strongly correlated with a value of replacement[1] which can take the form of the barycentre of a class of observations for all the non-respondents belonging to this same class. Among the methods most usually used in the presence of missing values dependent on actual values (MAR or "Missing At Random"), we can quote algorithm EM ("Expectancy-Maximization") of Dempster, Laird and Rubin (1977) or even the methods of Monte Carlo simulation by chains of Markov (*Robert 1996*). Lastly, regarding the models operating variable by variable, the mostly used are the linear regression, the analysis of the variance and the covariance, the logistic regression, the discriminating analysis, the decision trees and more recently the neuronal methods

[1] The non-response to a specific variable is actually more dependent on the true unknown value of this variable rather than on the values of the observed variables which are correlated to it. Thus the fact that the individual refuses to declare his income is a function of the amount of this last; for obvious reasons, only the strongly correlated and answered variables are used to consider the missing values (for examples, the age and the profession of the individual interviewed in the case of the income estimation). Like Lebart (2001) points it out, it is "*an undeniable limitation but much less fraught with consequences than a model of total independence*".

which generalize the preceding methods. For more details, see the work of Schafer (1997) and the articles of Schafer and Graham (2002), Collins et al. (2001), etc.

3)- METHODOLOGY OF THE RESEARCH

A)- THE NON-RESPONSE IN AN ACCESS PANEL

In an online access panel, the invitation to answer the questionnaire is addressed to the selected panelists by e-mail. The URL towards the questionnaire is generally included in the body of the mail. The panelists can also be revived at the semi-period of the data collection: the revival is thus an inexpensive mean used to decrease the non-response rate. Lastly, the non-response rate is also a function of the delay the panelists are granted to answer. This delay is generally fixed in an empirical way (according to the distribution of the answers within the period assigned for answering) and by also taking account of the imperative deadline for delivering the results to the client.

The access panel thus has two levers easy to implement to decrease the non-response rate: to grant a reasonable time to the panelists to answer (the distribution of the answers in time is an increasing and asymptotic function of the granted time) and to revive in an inciting and selective way the non-respondents. As far as online studies are concerned, the budgetary impact of these two corrective actions is lower than in the case of an offline study (e.g.: telephone or face-to-face) whose field cost is proportional to the daily rates of questionnaires carried out.

This is why the undertaken research gives itself three objectives:

- [1]- To identify the profiles of the non-respondents;
- [2]- To explore the influence of the time of response on the answer;
- [3]- To explore the effect of the revival on the answer.

B)- PRESENTATION OF THE DATA ACQUISITION

The studied panelists are the Web surfers regularly registered in the panel of Panel On The Web; they are invited to answer an online questionnaire during one of the three following months, June, July or August 2003. To neutralize the impact of the subject or the duration of the questionnaire, we retained the same barometric study undertaken each month near a sample of individuals who are representative of the French Net surfers aged 15 or more. Each panelist was thus retained for only one of the three waves. The fact of having three successive waves enables us to have an aggregate sample of big size (4.467 observations) and to neutralize, if necessary, the effect of an atypical month on the rate of participation.

The neutrality of the subject (a study of the awareness for a category of sites targeted to a large public) and the characteristics of the self-administered questionnaire (a number of questions limited to 10 of which a majority are closed questions, a duration of 5 minutes on average, the exlusion of all questions dealing with private or intimate life) bring a reasonable certainty that the subject and the questionnaire do not interfere on the studied phenomenon, the participation of the panelist in the study. For the same reasons, a contractual remuneration of an amount of 1 € was allotted to each respondent.

C)- THE THEORETICAL MODEL OF THE RESEARCH

To comply with the three aims of the study, the first of the explained variables selected is the participation (PART) of the panelist in the wave of the study for which he is requested. For the reasons called upon in the paragraph 2)-A), we rather employ the terms of "participation" (and "rate of participation") rather than those of "response" (and "rate of answer") to the survey. Two other variables are also studied: the time to respond (DREP) measured among the respondents only and the effect of the revival (RELA).The explained variables selected are divided into three groups according to whether they concern the three facets of the respondents as an individual, as a Web surfer or as a panelist:

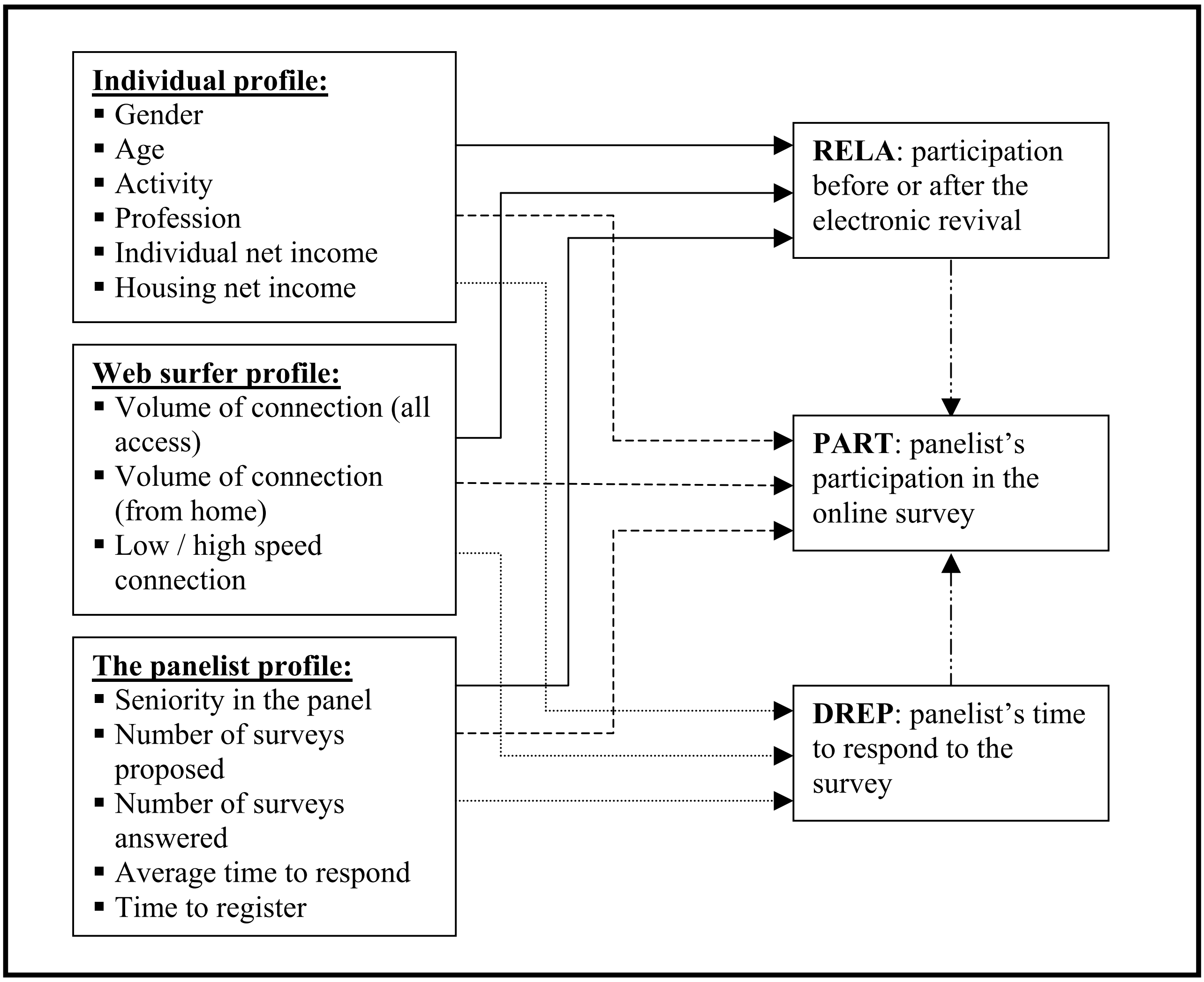
- [1]– The profile of the individual: the gender, the age, the activity, the profession, the education level, the individual income and the housing income. These variables inquired in the recruitment questionnaire (and updated annually) were retained because they are used to target samples to be surveyed. Moreover, they are also generally strongly correlated with the marketing phenomena (brand image opinion, attitudes and values, purchase habits, consumption frequencies, etc).

- [2]- The profile of the Net surfer: volume of connection on the Net whatever the access, volume of connection on the Net from home only. These two variables were retained because of the distortion usually granted to the online access panels: a study of Hoppe and Lamp (2001) quoted by Jolibert and Jolibert (2003) shows that the intensity of the use of the Internet is a factor influencing the results of an online study. We add to it a variable concerning the speed of connection (low versus high speed rate) because the speed of the access to the Net and the generalization of an unlimited use for an all-in-price in the high-speed subscription exert a notable impact on the uses of the Web : thus the Net surfers who have a high-speed connection are heavy consumers of the medium who clearly exceed the other Net surfers on the whole of the uses: the audio-visual, the games, the uploads and the publication of personal home pages more strongly attract these users and more generally all the practices which mobilize significant resources (source: *Médiamétrie*, "study on the uses of the high-speed connection", 2001).

- [3]– The profile of the panelist: the seniority in the panel, the numbers of surveys proposed the number of surveys answered, the average time of response to previous surveys, and the time to register. In the access panel of Panel On The Web, the panelist has one month to fill in the registering form and the attached questionnaire. It is thus interesting to insert the criterion of the time put to register in order to pinpoint in all the variables describing the former behavior of the panelists the ones which are predictive of his future behavior. In the same way, it is advisable to question to what degree the former behavior of the panelist (in particular his assiduity to take part in the studies proposed to him) is a good predictor of his future behavior.

If we take into account these variables altogether, the theoretical model of the research can be summarized by the following diagram (cf graph 1). At this stage of our work, we do not claim to propose (and test) a complete causal model. Our ambition is rather to explore the antecedents of the three explained variables selected: the participation in the survey (PART), the fact of taking part to the survey before or after being revived (RELA) and finally the average time to respond to the survey (DREP).

GRAPH 1: MODEL OF THE RESEARCH


Individual profile:
▪ Gender
▪ Age
▪ Activity
▪ Profession
▪ Individual net income
▪ Housing net income
Web surfer profile:
▪ Volume of connection (all access)
▪ Volume of connection (from home)
▪ Low / high speed connection
The panelist profile:
▪ Seniority in the panel
▪ Number of surveys proposed
▪ Number of surveys answered
▪ Average time to respond
▪ Time to register
RELA: participation before or after the electronic revival
PART: panelist's participation in the online survey
DREP: panelist's time to respond to the survey

4)- THE RESULTS OF THE RESEARCH

The results of research detail the existing correlations between the three groups of the explanatory variables selected (the profile of the individual, the profile of the Net surfer and the profile of the panelist) and the three explained variables: the participation in the survey (PART); the participation before or after the revival (RELA) and finally the time put by the panelist to answer the survey (DREP). We deliberately chose to study two other variables (RELA and DREP) in addition to the participation (PART) in order to highlight which are the effects of the revival on the rate of participation (and on its distribution within the sample) or what is the minimum time necessary to ensure a heterogeneous distribution of the respondents within the final sample. These two questions are of a great interest for the market research practitioners who must fix the optimum time for his data collection and choose (or give up) to revive the participants in the survey.

A this stage, we present a set of first exploratory results intended to be the subject of a modeling by means of the structural equations in order to validate the theoretical diagram of research presented above (cf graph.1).

A)- VARIABLES CORRELATED WITH THE PARTICIPATION

The periods of the data collections (June, July and August 2003) explain an average participation rate weaker than the one usually measured on the studies carried out with the help of an online access panel: 57% of the invited panelists took part in the survey against 65% on average (source : Panel On The Web). The rates of participation are also significantly different from one month to another: the highest participation rate is raised for June (64%) against 56% for July and 51% only for August (χ^2: 50.2; $p = 0.00$), a discrepancy which is explained by more panelists being on leave (involving an absence of their residence) in August and July than in June. Thereafter, the three waves were aggregated although one cannot completely draw aside the assumption that the "exceptional" timing of the period can exert a certain influence on the reported results (only a replication of the research during another quarter would make it possible to cancel this assumption).

Table 1 shows the variables correlated (or not) with the participation in the survey (SHARE). The variables within the three groups are classified by decreasing order of statistical significance; we also remind the variables tested which are not correlated with the participation at a 5% level of statistical significance. The participation is accounted for since the respondent answered the questionnaire within the time limit (21 days) and possibly after the only revival. The tools developed by Panel On The Web make it possible to each panelist to declare periods of unavailability during which he cannot (or does not wish) to be questioned.

TABLE 1 – VARIABLES CORRELATED (OR NOT) WITH THE PARTICIPATION IN THE SURVEY (SHARE)

Variables	χ^2	p
The individual profile (demographics)		
▪ Age	64.69	0.000 (*)
▪ Gender	11.63	0.000 (*)
▪ Activity	75.19	0.000 (*)
▪ Individual net income	57.88	0.000 (*)
▪ Education level	11.60	0.114 (n.s)
▪ Profession	13.83	0.129 (n.s)
▪ Housing net income	14.30	0.216 (n.s)
The Web surfer profile [2]		
▪ Volume of Internet connections (from home)	16.17	0.027 (*)
▪ Volume of Internet connections (all access)	7.47	0.382 (n.s)
The panelist profile		
▪ Seniority in the panel	124,29	0.000 (*)
▪ Number of surveys proposed	97.60	0.000 (*)
▪ Number of surveys answered	563.26	0.000 (*)
▪ Average previous time of response	30.20	0.000 (*)
▪ Time to register	20.45	0.004 (*)

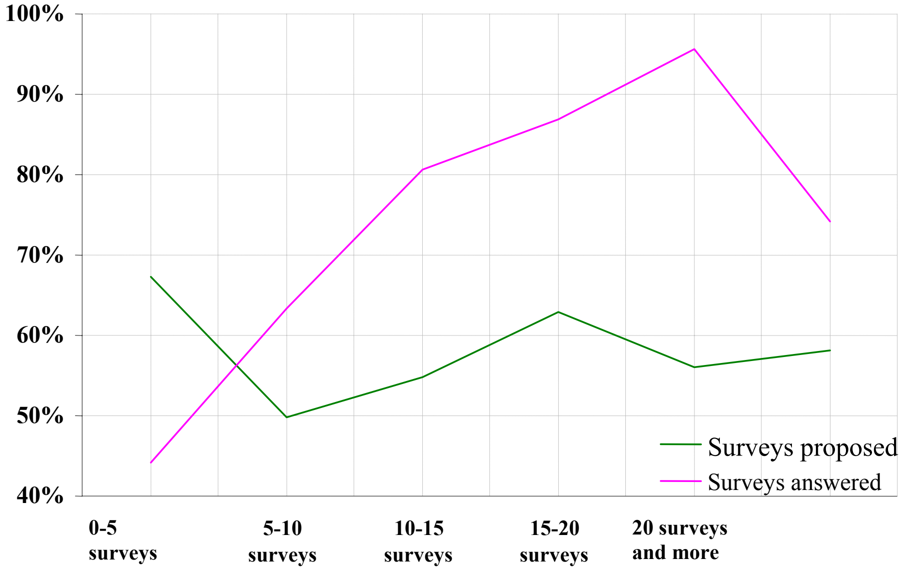
(*) Significant at a 5% level; (n.s) Non significant at a 5% level.

[2] The speed of the Internet connection (high versus low) is an information qualified in the survey and it is thus available only for the respondents to the study (cf. paragraph B and C).However considering the strong correlation of the speed of the connection with the volume of connection to the Net from home, we can make the assumption that the speed of the Internet connection is also correlated with the participation, an assertion which needs to be checked in a forthcoming replication of the study.

The first teaching of research: the participation in an online survey is not equally distributed according to the profile of the individuals. The gender, the age, the activity and the income of the respondent influence significantly the rate of participation with the three surveys proposed: the men have a higher propensity to answer (59%) than the women (54%); the Net surfers between 35 and 54 years respond more to the survey (63%) than those who are 15 to 24 years old (48%); it results from this that the wage earners have also a stronger participation (62%) than the students (26%) and that those who state to have no income give less often their opinion online (52%) than the others (the highest rate of participation is observed among the panelists declaring a monthly net income between 2400 and 3000 €uros: 62%). Conversely, the profession, the education level and the housing net income do not influence on the fact of taking part or not to the study. These results militate in favor of looking for the highest rate of answer (enhanced partly by the use of a revival and the granting of a reasonable time to answer); if not, the sample might be distorted insofar as the propensity to answer is influenced by the variables describing the individual profiles.

The second teaching: the profile of the panelist also explains his propensity to take part in the survey. The most recent registered panelists take part more to the survey, evidence of a natural mortality or attrition rate of the panelists (71% of participation of those who registered since less than one year against 51% only for those who registered since more than one year but less than two years). The rate of participation increases again beside the panelists registered for more than one year. The number of surveys proposed and the number of surveys answered also exert an influence on the participation as attested by the graph 1 below.The rate of participation culminates near those who answered 20 surveys and more (96%) posing the delicate problem of the maturation effect (or of the learning effect); it is also higher beside the more recent panelists (1 to 5 surveys proposed) and more mature (between 15 and 20 surveys proposed), but it decreases beyond.

GRAPH 1 - RATE OF PARTICIPATION ACCORDING TO THE NUMBER OF SURVEYS



Finally and logically, the participation in the survey depends on the profile of the Net surfer and more particularly of the volume of connection to the media (since the residence only). The highest rate of participation in the survey is observed among the panelists who connect themselves to the Web on average between 5 and 10 hours per week (61% against 57% among the other sub-segments).

B)- VARIABLES CORRELATED WITH THE REVIVAL EFFECT

The current practice in the access panels consists in reviving by electronic mail and in a selective way only the panelists who have not answered the survey at approximately halfway through the period of the data collection. The low cost of the revival by electronic mail explains the generalization of this practice, although few studies exist on the impact of the revival on the rate of participation and on a possible distortion of the sample which the revival might induce.

Table 2 shows the variables correlated (or not) with the effect of the revival (RELA) or more precisely with the fact of answering before or after being revived. The platform developed by Panel One The Web for the automated management of the surveys makes it possible to revive by mail only the panelists who have not started the questionnaire or those which have not completed it entirely. To investigate the selective effect of the revival on the participation in the surveys according to the profile of the individual, the one of the Net surfer or of the panelist, we divide the sample of the respondents into two: the individuals who answered the questionnaire without being revived and the others. Again, the variables within the three groups are sorted by descending order of statistical significance.

TABLE 2 – VARIABLES CORRELATED (OR NOT) WITH THE EFFECT OF THE REVIVAL (RELA)

Variables	χ^2	p
The individual profile (demographics)		
▪ Gender	3.80	0.051 (n.s)
▪ Age	11.05	0.086 (n.s)
▪ Profession	14.29	0.111 (n.s)
▪ Individual net income	11.83	0.756 (n.s)
▪ Activity	3.88	0.794 (n.s)
▪ Level of education	3.75	0.810 (n.s)
▪ Housing net income	6.32	0.852 (n.s)
The Web surfer profile		
▪ Low-speed versus high-speed connection	6.90	0.032 (*)
▪ Volume of Internet connections (all access)	3.59	0.827 (n.s)
▪ Volume of Internet connections (from home)	2.10	0.953 (n.s)
The panelist profile		
▪ Average previous time of response	539.99	0.000 (*)
▪ Seniority in the panel	9.88	0.019 (*)
▪ Time to register	14.71	0.039 (*)
▪ Number of surveys proposed	5.38	0.371 (n.s)
▪ Number of surveys answered	5.04	0.411 (n.s)

(*) Significant at a 5% level; (n.s) Non significant at a 5% level.

The table 2 reveals that contrary to the conclusion drawn for the participation variable, no criterion which describes the profile of the individual is correlated with the fact of answering before or after the revival. The men and the women, the youngest or the oldest, the more or less wealthy or educated are equally numerous to answer either before or after the revival. This result tends to show that if the revival increases the rate of participation in the survey, it does not distort the composition of the sample on the demographic criteria.

The same conclusion does not apply when we consider the variables which characterize the respondent as a Net surfer and as a panelist. The velocity of the Internet access (either low-speed or high-speed connection) influences the effect of the revival as the individuals who are connected in low-speed are in proportion more likely to answer after they have received the revival (in comparison with those who are using a high-speed connection):27% against 21%. On the opposite, the volume of connection to the Net from one's home as well as from any other place does not influence the fact of answering before or after the revival, a surprising result if it is brought closer to the fact that the volume of connection is correlated with the average time for response to the study (see below).

Lastly, the revival has a selective effect on the participation according to the profile of the panelist. Depending on the delay for registering, on the time for response to the former studies and on the seniority in the panel, the propensity to answer before or after the revival is unequally distributed: 80% of those who discharged their inscription in one day (in only one session generally), when they answer the study, do it before revival against 76% for the others; 90% of those who answer their former surveys in three days at the most did not await the revival to answer the study which is the subject of the experimentation. (on average 21% of the respondents answered the online questionnaire after they were revived); finally the respondents most recently registered in the panel are also more reactive: only 18%of them await the revival to answer against 24% of those who have between one year and two years of seniority.

In conclusion, the effect of the revival is mitigated: if the participation before or after revival is weakly correlated with the profile of the individual in terms of demographics, it proves unequally distributed according to the speed of Internet connection (the revival more strongly increases the participation of those who are connected at low speed); the effect of the revival is moreover dependent on the historical profile of the panelist: the most recent panelists are less often revived by mail (because undoubtedly their motivation for answering their very first surveys is more pronounced, although no effect of the number of surveys proposed or answered is statistically checked); the revival is all the more necessary as the panelist reveals a larger inertia by his last behavior (delay for registering, average time for response to the previous surveys, etc.). These results are confirmed by the study of the variables correlated with the time for response.

C)- VARIABLES CORRELATED WITH THE TIME OF RESPONSE

To answer one of the three waves of the survey, the panelists had three weeks as of the first notification by electronic mail. Each respondent thus had the same time to answer the study. At halfway, a revival by mail was made to the non-respondents or to the panelists who had started the questionnaire without completing it (insofar as the questionnaire was short, the latter were marginal). Three weeks of data acquisition is a time longer than that usually retained for online studies; in our case, this choice was justified by the fact that the investigation proceeded for the summer period.

The graph 2 plots the distribution of the answers throughout the period assigned for answering; the examination of the three curves confirm the exponential pace and quickly asymptotic shape of the distribution of the answers to a survey conducted with the recourse to an online access panel: on average 50% of the respondents fill in the questionnaire in less than 24 hours and approximately 70% in less than two days. Although the three curves show a similar shape, the distribution of the answers for the August wave is more "slack" than that of July and June 2003. Let us recall however that the people having declared their unavailability for the period of the survey were withdrawn from the sample, which accounts for the three curves to preserve a comparable pace altogether.

GRAPH 2 – DISTRIBUTION OF THE ANSWERS WITHIN THE PERIOD ASSIGNED FOR ANSWERING.

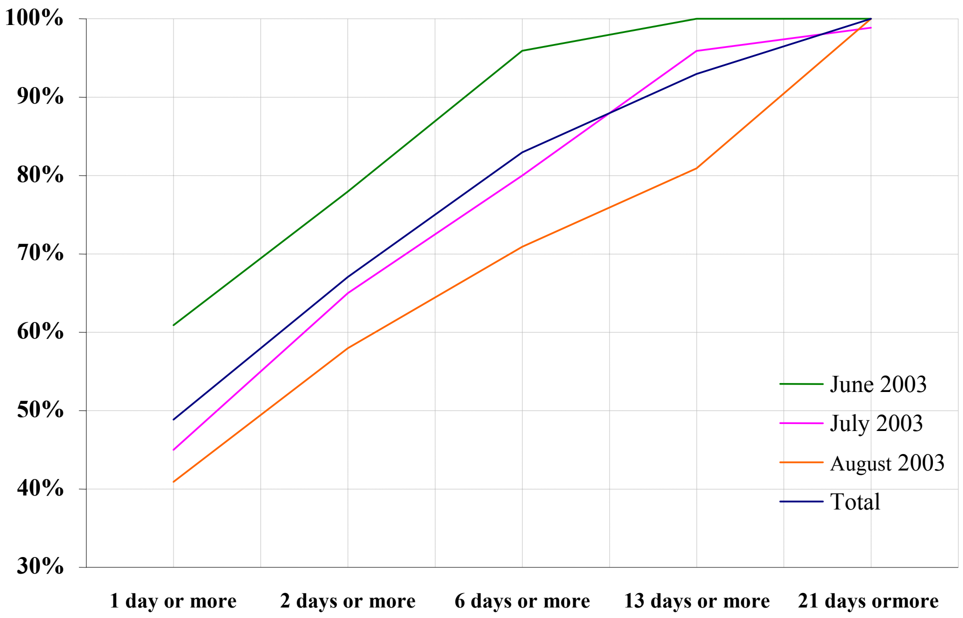


Table 3 shows the variables correlated (or not) with the time for response to the survey. The time for response to the survey is an ordinal variable including five modalities: less than 24 hours; between 1 and 2 days; between 3 and 6 days; between 6 and 13 days; between 13 and 21 days. As for the two preceding tables, the variables within the three families are classified by decreasing order of statistical significance.

TABLE 3 – VARIABLES CORRELATED (OR NOT) WITH THE TIME FOR RESPONSE (DREP)

Variables	χ^2	p
Individual profile (demographics)		
▪ Gender	8.60	0.071 (n.s)
▪ Activity	32.92	0.238 (n.s)
▪ Net housing income	48.78	0.287 (n.s)
▪ Net individual income	69.61	0.294 (n.s)
▪ Age	26.14	0.346 (n.s)
▪ Profession	35.43	0.495 (n.s)
▪ Education level	26.46	0.548 (n.s)
Web surfer profile		
▪ Volume of Internet connections (all access)	44.11	0.003 (*)
▪ Volume of Internet connections (from home)	46.08	0.017 (*)
▪ Low-speed versus high-speed connection	16.57	0.035 (*)
Panelist profile		
▪ Average previous time of response	992.19	0.000 (*)
▪ Time to register	41.56	0.047 (*)
▪ Seniority in the panel	17.24	0.140 (n.s)
▪ Number of surveys proposed	20.86	0.405 (n.s)
▪ Number of surveys answered	12.32	0.905 (n.s)

(*) Significant at a 5% level; (n.s) Non significant at a 5% level.

The examination of table 3 shows us that no variable which describes the profile of the individual is correlated with the time of response to the study; it is true that on average seven panelists out of ten answer the study in the first two days. The gender, the age, the activity, the profession, the education level or the income of both the individual and the housing do not exert any influence on the propensity to answer immediately or not to the study. This result pleads in favor of one of the assets of the recourse to the online access panels: the reactivity of the respondents which does not appear to induce a distortion on the demographics of the respondents.

However the same conclusion does not apply to the variables describing the use of the Internet. As it is logical, the volume of connection on Internet (from one's residence and whatever the access) and the low-speed versus high-speed connection are positively correlated with the time of response: those who are connected more than 10 hours per week are 55% to answer within 24 hours; finally, those who have a high-speed connection at home are 52% to answer in one day (against 44% of those who have a low-speed connection). These results indicate that it is currently necessary to grant a sufficient time when the subject of the study is likely to be influenced by the profile of the Net surfer (study on the use of the Internet, on the attended sites, on the online centers of interest, on the online purchases, etc.) under penalty of distorting the representativity of the sample and the quality of the estimates.

Lastly, concerning the previous behavior of the panelist, two variables only appear positively correlated to the time of response to the study: the average time of response to the previous surveys and the time put to complete one's registration to the panel. These two variables seem to indicate that there is some constancy in the behavior of the Net surfer with respect to online studies: those who answer the latest behave the same way for all the surveys and were also those who spent the longest time to achieve their registration. The implementation of a tool to score the reactivity of the panelist, based on these two criteria is for this reason under consideration at Panel On The Web. This tool would make it possible to lay out of a very reactive panel for “express” studies subject to the limitations indicated in the preceding paragraph. Another interesting lesson: the seniority in the panel (between 1 month and 4 years), the number of studies proposed and the number of surveys which the panelist answered do not influence the average time of response (whereas these same variables influence the participation in the study).

D)- HIERARCHY OF THE EXPLANATORY VARIABLES OF THE PARTICIPATION (PART)

Of the three explained variables, one of it deserves a thorough exploration: the rate of participation because of the greater number of explanatory variables which are correlated with it (10 out of the 14 studied) and of its managerial implications, the participation rate being one of the key variables of the management of an access panel. The χ^2 test enables us to establish the statistical relations between the two groups of variables, dependent and independent, without however establishing a hierarchy between them and without eliminating the variance shared between the explanatory variables.

To achieve this goal, we resort to a segmentation according to the method of the interactive decision tree (IDT) proposed by the software SPAD. This method is based on a discrimination using a binary tree, path opened by Morgan and Sonquist (1963) and Morgan and Messenger (1973) with a method known as AID for "*Automatic Interaction Detection* ". The segmentation by binary decision tree has the advantage of being little constrained by the nature of the data: one can indeed use as explanatory variables simultaneously continuous, ordinal and nominal variables, the same algorithm being implemented to analyze a nominal variable (discrimination) and a continuous variable (multiple regression). For more details, we emphasize the work of Celeux and Nakache (1994) and the illustrated article of Gueguen and Nakache (1988). There are various algorithms of segmentation: two more widespread are CHAID (*Kass, 1980*), method of induction of decision tree resting on a statistical criterion of discrimination, the measurement of χ^2 and C&RT, resulting from a monograph (*Breiman and Al, 1980*) which proposes a unified method to deal with the problems of discrimination and regression starting from the concept of "purity", i.e. of successive pruning of the tree so as to reduce the rate of false classification. We choose the algorithm of CHAID, more suitable when one wants to carry out the first exploration of the data (*Kass, 1980*) and in which the decision to segment a top depends on a χ^2 test of independence carried out on the table of contingency associated with the sheets which will be produced by the segmentation. If this test is negative, the top is not segmented and becomes a final top. The algorithm is applied successively to two samples: a first known as "of test" corresponds to 33% of the randomly drawn individuals; it makes it possible to determine the rules of segmentation which will be applied later on to the remaining fraction of the sample (called training sample).

We choose to withdraw from a first analysis the variable “number of surveys proposed” as this one proves correlated (and therefore redundant) with two other variables integrated in the segmentation: seniority in the panel (χ^2 = 4098; p = 0.000) and numbers of surveys which the panelist answered (χ^2 = 5396; p = 0.000). The matrix of confusion on the samples of test and training enables us to validate the quality of the tree of segmentation obtained: with respectively 75% and 76% of the individuals correctly classified, the segmentation analysis appears satisfactory.

The measurement of the impact of each attribute enables us to know the role of each variable in the construction of the tree. The values indicated to the table 4 labelled under the wording "impact" represent a weighted average of the impact of each attribute on all the segmentations candidates, knowing that less importance is conferred on the impacts measured on the lower parts (on the right) of the tree.

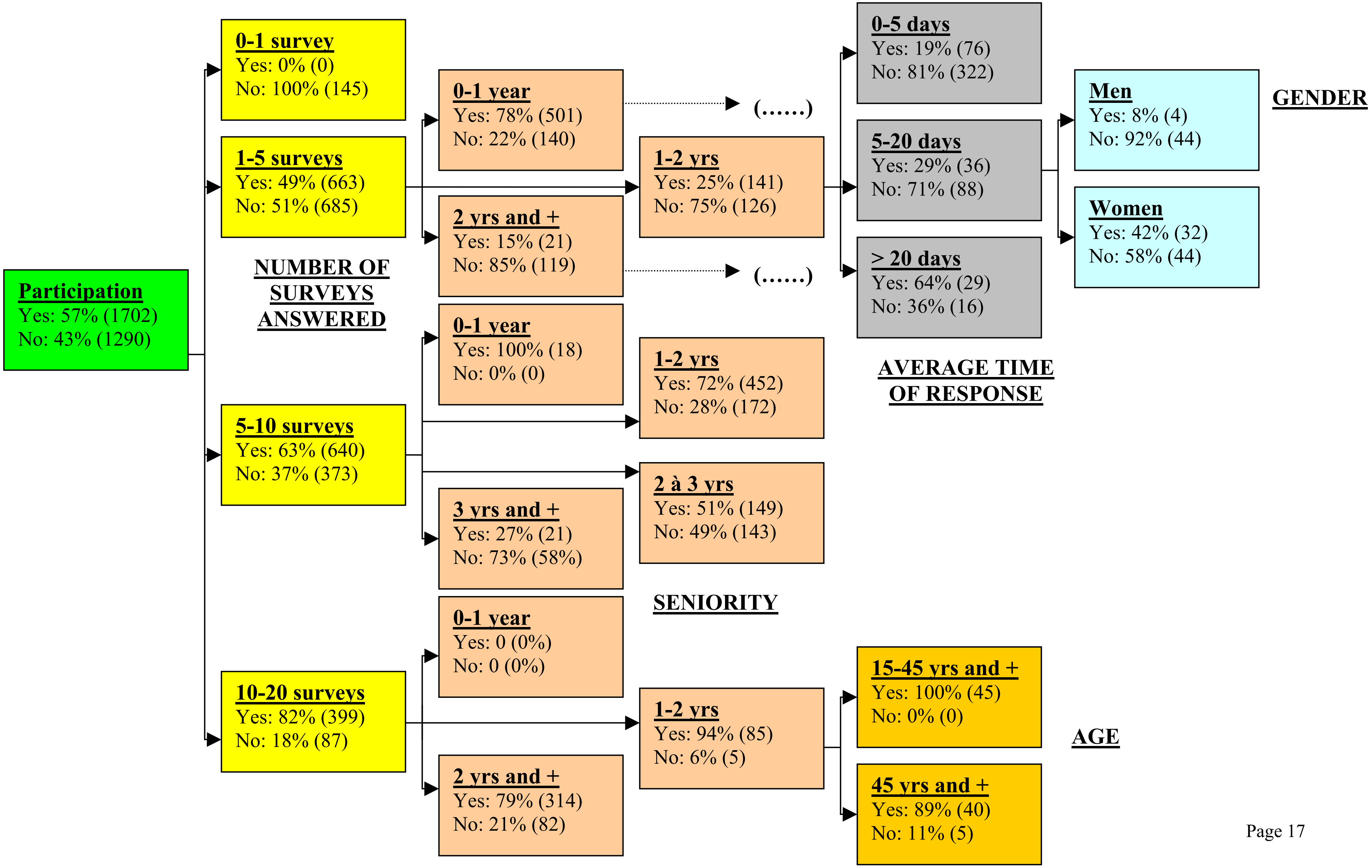
TABLE 4 – CHARACTERIZATION OF THE PARTICIPATION (PART) BY THE EXPLANATORY VARIABLES

Variables (*)	χ^2	Test value	P	Weighted impact
▪ Number of surveys answered	563.26	99.99	0.000	0.2020
▪ Average previous time of response	296.30	99.99	0.000	0.1458
▪ Seniority in the panel	124.29	10.65	0.000	0.1266
▪ Activity	76.52	7.23	0.000	0.1132
▪ Age	64.69	6.81	0.000	0.0787
▪ Net individual income	57.88	4.72	0.000	0.0680
▪ Gender	11.42	3.18	0.001	0.0448
▪ Time to register	21.79	2.55	0.005	0.0390
▪ Volume of Web connection from home	19.35	2.22	0.013	0.0000

(*) Variables are sorted by decreasing order of the probability value $p < 0.05$.

The analysis of the table 4 and of the segmentation tree whose illustration is presented at diagram 3 reveal us that the three variables most strongly correlated with (and explanatory of) the rate of participation are all related to the profile of the panelist, namely the average number of answered surveys, the time of response and the seniority in the panel. The demographics (gender, age, activity and individual income) have an impact less pronounced on the fact of taking part in the study, and then comes the delay for registering and finally the volume of connection to the web from one’s residence. The propensity to take part in a regular way to online surveys seems partly independent of the profile of the Net surfer and undoubtedly more related to a general attitude on the fact of giving one’s opinion or of conforming to the rules subscribed in the phase of recruitment of the panel; in other words, the probability of taking part in a subsequent survey is conditioned by one’s previous participation, one’s reactivity and one’s seniority in the panel. Let us underline finally that the highest rate of participation is observed among those who participated to at least five surveys (69% against 44%), who answer within three days on average (63% against 57%) or, in another register, who have less than one year of seniority in the panel (71% against 53%). Criteria considered altogether, the segment which possesses the highest rate of participation is that of the panelists aged from 15 to 45, having from 1 to 2 years of seniority and having already answered 10 surveys at least and 20 at the maximum.

GRAPH 3 – SIMPLIFIED INTERACTIVE DECISION TREE (IDT) FOR THE EXPLAINED VARIABLE “SHARE” (PARTICIPATION IN THE STUDY)



5)- CONCLUSIONS AND LIMITATIONS OF THE RESEARCH

The research which is the subject of this article is a first investigation of the data collected on the refusal to answer (more accurately named « non-participation »); it is the first stage of a more global study of the behavior of the interviewees inscribed to an online access panel. I t is thus logical at this stage that the first results, though interesting, call for reserves and limits which are advisable to point out.

One of the noticeable results of the study is that, even though the participation in a study undertaken on an online access panel is high, it is not homogeneously distributed among the individuals. The demographics of the respondents (the gender, the age, the activity and the income level) appear correlated with the fact of answering the study. These variables exert however a lower influence than the variables that describe the past behavior of the panelist, in particular the number of surveys answered, his seniority in the panel and the average time of response to the previous studies. There thus seems to exist a propensity to answer in a constant way to online studies, an attitude which is partly independent of the demographic characteristics of the individual. Which are the determinants of this attitude? Could it be explained by a willingness to give one's opinion in order to make one's desire better known? Could it even be enhanced by a sense of belonging to the Net surfers' community? It is difficult to answer this question which calls for other investigations.

The very first results on the rate of participation are usefully hightlighted by the study of two other variables: the fact of answering before or after a mail of revival and the time of response. If the effect of the revival does not depend on the demographics, it however depends on the profile of the panelist: the impact of the revival on the participation is more pronounced among the older panelists and among those who revealed a lesser reactivity in the past; finally the revival is more effective to the Net surfers using a high-peed connection rather than for those using low-speed. The same conclusions apply for the time of response: the seniority in the panel and the past observed reactivity are good predictors of the time of response; High-speed connected Net surfers respond more quickly than low-speed connected ones. The volume of Internet connection (whatever the access mode) is also a variable of influence.

Our research took for dependent variable the participation and two other variables which are attached to it: the time of response and the effect of the revival. Being the more general problems of the non-response, we drew aside a significant effect: the one that the non-participation (or the refusal to answer) exerts on the quality of the answers itself (that one can bond to the reliability of the estimates).The better knowledge of the variables correlated with the participation (together with the effect of the revival and the time of response) has a managerial interest all the more pronounced since they produce a distortion on the reliability of the end results; in other word the non-participation is all the more critical as there might exist (as we pointed out) a difference between the opinions of the respondents and the unknown opinion of the non-respondents, a subject which our research does not tackle and which forms one of its caveat. Lastly, the period of data collection also constitues a limitation to our research: it would be advisable to replicate this study apart from the period of summer to reinforce the external validity of the reported conclusions.

REFERENCES

Bosnjak Michael M., Tuten Tracey L. (2001), « *Classifying Response Behaviors in Web-based Surveys* », Journal of Computer-Mediated Communication, 6, 3, http://www.ascusc.org/jcmc/vol6/issue3/boznjak.html

Breiman L., Friedman J., Olshen R. A. et Stone J.C. (1984), *Classification and regression trees,* California : Wadsworth International, 358 p.

Brennan Mike, Hoek Janet et Astridge Craig (1991), « *The effect of monetary incentives on the response rate and cost-effectiveness of a mail survey* », Journal of the Royal Statistical Society, 33, 1, p. 229-241.

Celeux G., Nakache J.P (1994), « *Analyse Discriminante sur Variables Qualitatives* », Paris, Polytechnica, 270 p.

Chen Henry C. K. (1996), *« Direction, magnitude and implications of non-response bias in mail survey »,* Journal of the Market Research Society, 38, 3, p.267-276.

Collins L. M., Schafer Joseph L. et Kam C. M. (2001), « *A comparison of inclusive and restrictive missing-data strategies in modern missing data-procedures »,* Psychological Methods, 6, p. 330-351.

Dempster A. P., Laird N. M. et Rubin D. B. (1977), « *Maximum likelihood form incomplete data via the EM algorithm »,* Journal of the Royal Statistical Society, 39, 1, p. 1-38.

Ferber Robert (1948), *The problem of bias in mail return : a solution,* Public Opinion Quarterly, 39, 3, p.239-244.

Frankel M. R., Frankel L. R. (1977) « *Some recent developments in sample survey design »,* Journal of Marketing Research, 14, p. 280-293.

Gueguen A., Nakache J. P. (1988), « *Méthode de discrimination basée sur la construction d'un arbre de décision binaire »,* Revue de Statistique Appliquée, 36, 1, p.19-38.

Hoppe, Michael et Lamp, Rainer (2001), « *The quality of online panels - a methodological test* », dans Fellows, Deborah S. (Ed.): « *Worldwide Internet Conference and Exhibition Net Effects* », 4, Barcelone, Espagne, 11-13 février 2001, Amsterdam : Esomar, p. 243-262.

Jolibert Alain, Jolibert Bertrand (2003), *« La validité des panels d'internautes »,* p. 213-220, dans *Savoir gérer : mélanges en l'honneur de Jean-Claude Tarondeau,* coordonnée par Pierre Le Moal, Paris : Vuibert.

Kalton Graham (1983), *Introduction to Survey Sampling,* Quantitative Applications in the Social Sciences Series, Newbury Park, California : Sage Publications, 94 p.

Kass G. V. (1980), « *An exploratory technique for investigating large quantitites of categorical data* », Applied Statistics, 29, 2, p. 119-127.

Lebart Ludovic (2001), *« Introduction au prétraitement des fichiers d'enquêtes : redressement ; données manquantes, fusions / injections »,* dans *Traitement des fichiers d'enquêtes : redressement ; données manquantes, fusions / injection,* Michel Lejeune (éditeur), Presses Universitaires de Grenoble, p. 9-15.

Litlle Roederick J. A., Rubin Donald B., *Satistical analysis with missing data,* 2ème édition, Hoboken, New Jersey : Wiley, 381 p.

Madow William G., Nisselson Harold et Olkin Ingram (1983), *Incomplete data in sample surveys,* 2 vol., New York : Academic Press.

Marjorie Donald N. (1960), « *Implication of non-response for the interpretation of mail questionnaire data »,* Public Opinion Quarterly, 1, p. 99-114.

Morgan J. N., Messenger R. C. (1973), « *THAID : a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables* », Institute for Social Research, université du Michigan.

Morgan J. N., Sonquist J. A. (1963), « *Problems in the Analysis of Survey Data and a Proposal »,* Journal of the American Statistical Association, 58, p. 119-127.

O'Muircheartaigh Colm, Campanelli Pamela (1999), *« A Multilevel exploration of the role of interviewers in survey non-response »,* Journal of the Royal Statistical Society, series A, , 162, 3, p. 437-446.

Robert Christian (1996), *Méthodes de Monte Carlo par chaînes de Markov,* série Statistique Mathématique et Probabilité, Paris : Economica, 340 p.

Schafer Joseph L. (1997), *Analysis of Incomplete Multivariate Data,* Monographs on Statistics and Applied Probability Series, 72, London : Chapman & Hall, 448 p.

Schafer Joseph L., Graham J. W. (2002), « *Missing data : our view of the state of the art »,* Psychological Methods, 7, p. 147-177.

Wiseman Freederick, McDonald Philip (1980), *Towards the development of industry standards for response and non-response rate*, Cambridge, Mass. : Marketing Science Institute.