METHODOLOGICAL TESTS ON ONLINE RESEARCH: INCIDENCE OF FORMAL ASPECTS/QUESTIONNAIRE LAYOUT ON THE RESULTS AND RESPONDENTS' ATTITUDES CONCERNING ONLINE QUESTIONNAIRES

Bruno Schmutz and Olivier Lê Van Truoc, Ipsos MediaCT

1– General context:

Today, a large and increasing number of quantitative studies are making use of online interviewing and data collection.

This technology evolution is occurring rapidly, and research institutes have questions about it, notably concerning the representativeness of Internet users but also in terms of the quality of the responses obtained. A variety of studies have been carried out on the subject of sampling: How can we build a reliable sample base while optimizing response rates? Is it possible to obtain information that is pertinent for a larger population simply by interviewing Internet users (for example, for Internet users and non-users), and under what conditions? ... Other research has focused on comparing the results of offline research (face-to-face and phone) with those of online studies.

However, little real research has been focused on the **formal aspects of online questionnaires**: question presentation and positioning, formatting of scoring scales, questionnaire ergonomics, the more or less directive style of interviewee instructions. Research institutes are obviously trying to construct questionnaires that are attractive and flow smoothly, and through the studies they conduct they are gradually accumulating experience in terms of what is likely to work well or less well, but there is little written about evaluating the incidence of questionnaire layout on interviewee responses. Since we are talking about self-administered questionnaires, the layout is clearly very important, an aspect we might compare to **the influence of interviewers in directed interviews.**

In addition, it is quite likely that **respondent-questionnaire interactions and interviewee perceptions** are very different on line from what they are in self-administered printed questionnaires. What we know from the past is certainly outdated and will become increasingly obsolete.

In fact, with the expansion of new information technologies and widespread use of online forms, snapshot polls… numerous interviewees now have an "expert" or "confirmed user" status concerning research institutes' questionnaires.

These observations apply to all quantitative studies and are probably even truer in the context of press and press readership evaluations: these are established studies, standardized, that are based on a series of questions (filters, frequency and date of most recent readership) that appear to be simple but that are often viewed by respondents as being difficult or artificial, and certainly highly repetitive. The influence of the interviewers, of their probes and prompts, the rhythm of conducting the questionnaire and the nature of the visual stimuli are all determining factors (this has been shown widely during previous symposiums) that can have an impact on the readership levels that are recorded.

With a CAWI interview (Computer Assisted Web Interview), in the absence of an interviewer, we have very little control over the interview conditions. Question layout and instructions for filling out the questionnaires thus become crucial aspects of the methodology: They should contribute to the questioning process being attractive and fluid, but should also ensure that interviewees remain attentive and take sufficient time to think about their answers. The measurement tool also has to treat all publications equally, regardless of their nature, and, undoubtedly more difficult, ensure that the conditions for filling out the questionnaire be as uniform as possible for all respondents.

2 – Methodological tests on questionnaire layout and attitudes of online respondents:

Since 2007, Ipsos MediaCT France has carried out a research program covering the layout of online questionnaires and respondent attitudes concerning these questionnaires. As this is a wide-ranging and ambitious area of research, we can't pretend to offer exhaustive answers but merely to offer a number of contributions.

Our test phases combine two approaches:

A) Quantitative tests within the large framework of the "Ipsos Profiling" online survey[1]

The large size of this survey's sample population (30 to 50,000 interviews among Internet users aged 15 and +) enables us to test the impact of various formal aspects of online questionnaires on large, matched sub-populations (more than 1000 interview at a time), corrected posteriorly for a combination of criteria (socio-demographic, types of Internet connections and use frequency) in order to make them as comparable as possible.

We have covered subjects that concern all types of quantitative studies: tests on the effects of order / fatigue, rotation of where items are presented in the questionnaire, variance of question blocks (4 to 6 questions per block), impact of the obligation to answer (for 2 successive questions, we alternated between making responses mandatory and leaving interviewees free to respond or not), evaluation of the formal aspects of questionnaires (checking answer boxes on the left vs. the right-hand side of the page), layout and presentation of scoring scales…

The initial test results (see paragraph 3) showed the importance of questionnaire layout for interviewee responses, and notably the risks of bias associated with poor presentation conception and instruction ergonomics.
We felt that Press readership would likely be measured (at least partially) online in the future, and decided starting in 2009 to implement an additional wave of tests dedicated to press readership measurement.
.

B) A more qualitative phase of re-interviewing of 1400 respondents

This was done to gain a better understanding of interviewee perceptions of online questionnaires, their motivations for answering, their difficulties, how they filled out the questionnaires… and some of their preferences in terms of presentation.

3 – Selected general results[2]:

Before talking specifically about our readership research in 2009, here are a few illustrations that concern all types of studies and that show, in our opinion, the need to continue these investigations.

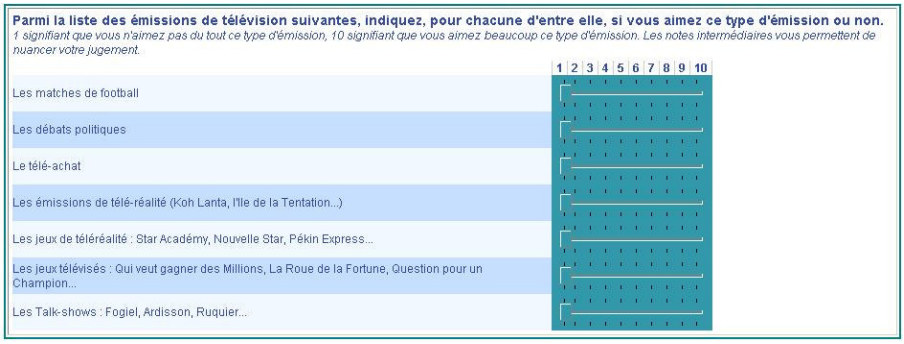
a) Scoring scales

We asked the same question covering 10 items and asked for answers in the form of scores on a scale of 1 to 10 to five perfectly matched sub-populations (1600 interviews each): only the presentation of the scoring scale varied.

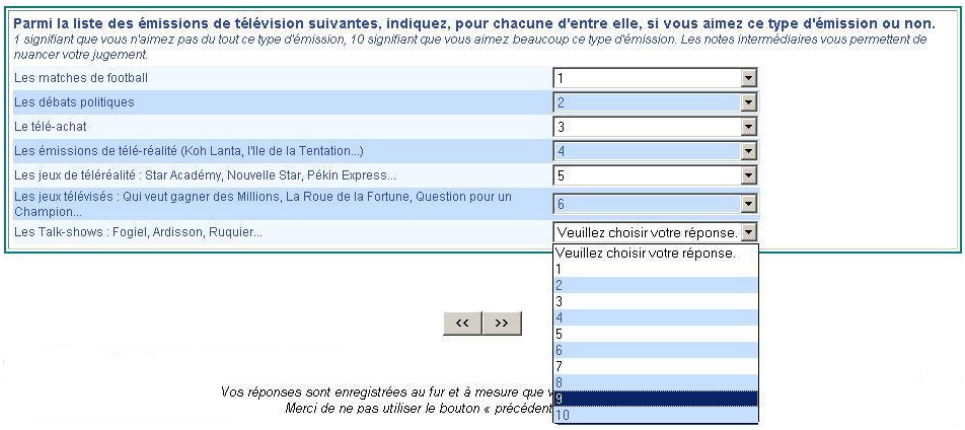
Our tests clearly show that the layout of the scales has an incidence on answer distribution and scoring averages. This incidence can be linked to aspects that are purely practical such as the "effort" needed to answer. Presenting the scores in the form of a drop-down menu or across a horizontal cursor field generate fewer high scores (particularly scores of 9 and 10 on a scale of 10) because both presentations require interviewees to move their cursors to the bottom or far right of the lists. The scale with horizontal cursor generated twice as many grades “2” compared to the average, and almost 30% less often scores 1: the cursor being initially positioned on grade 1, some interviewees get the impression that they need to move the cursor so as to answer the question. The first low grade thus becomes 2 instead of 1 for them.

[1] The Profiling survey concerns the profiles of visitors to more than 200 major French Websites that carry ads plus a dozen Web portals. The Profiling survey was launched in 1998 and is conducted twice yearly. Between 50,000 and 80,000 interviews are carried out each year with Internet users aged 15 and over who live in France and have visited the survey Websites in the past 30 days. Interviewees are recruited by screen banners appearing on more than 80 partner Websites. This survey produces profile information for each site (socio-demographic and professional characteristics, Internet equipment and use, lifestyle and consumption habits, media behavior …). The very large sample population enables us to build matched sub-populations for our tests.

[2] A part of our first observations was published in “Méthodes de Sondages – Dunod 2008” by the “ Société Française de Statistique”.

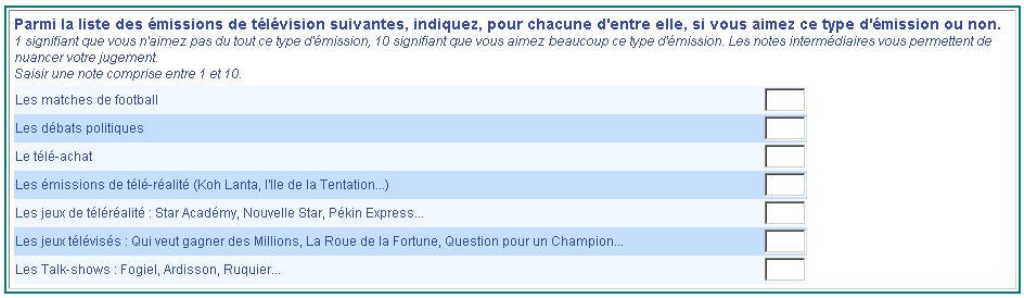


Horizontal cursor



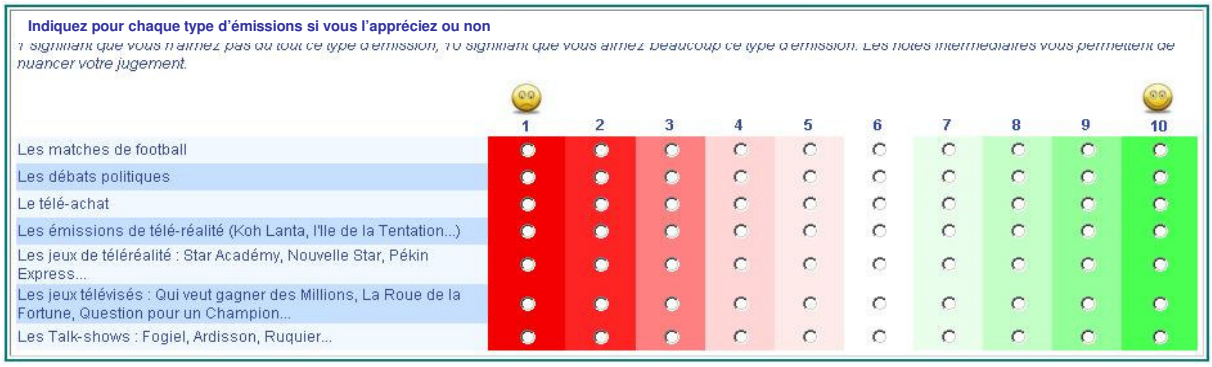
Drop-down menu

The observation concerning the “effort” to answer seems also true when we ask for direct keyboard input : because the number "10" requires 2 keystrokes, we registered 15% less scores of 10 than in average. And when we ask interviewees to respond freely by using their keyboards to enter scores in boxes, we also record uneven peaks, notably for the score of "5" (25% more responses than with formatted scales), which was again, and incorrectly, seen as the median.



In contrast, a bipolar horizontal scale seems to produce a greater variety of responses.

But, once again, the presentation of the scale is not neutral: color shading (cool tones for "1" and warm tones for "10") can have an influence on responses. With shading, the score of "6", which in our experience corresponds to the color "white", was undoubtedly seen as the median and was checked 25% more often than when using scales with no shading.



b) Requiring participants to absolutely answer questions:

A second experience involved comparing responses concerning the frequency of doing a number of leisure activities, this evaluated on a 3-point scale (regularly, occasionally or never) and either requiring interviewees to answer[3] or leaving them free to answer as they liked. When respondents were left free to answer (see table below) we recorded a very small proportion of blank answers (an average of 3%), and we found no variation in the "regularly" scores regardless of whether they were free to respond or required to do so. The gain in information when we required Internet users to answer was thus marginal.

Frequency of doing the following activities: sample results as %

	Regularly	**Occasionally**	**Never**	**NA/DK**
Gardening, DYI (unprompted)	33	39	26	2
Gardening, DYI (prompted)	***32***	***41***	***27***	**-**
Home decorating (unprompted)	26	56	15	3
Home decorating (prompted)	***26***	***57***	***17***	**-**
Photography (unprompted)	26	50	21	3
Photography (prompted)	***25***	***52***	***23***	**-**
Other manual activities (unprompted)	13	32	51	4
Other manual activities (prompted)	***14***	***33***	***53***	**-**

Further, requiring them to answer can be counter-productive because it can force respondents (whose participation, we need to remember, is based solely on good will) to interrupt filling out the questionnaires. In fact, when responses are mandatory, 55% of the interviewees say they sometimes stop answering and 56% say they sometimes "answer anything at all".

We can thus question the usefulness of verifications and constraints, which can be sources of false security, and retain them only for major questions or those for which a clear answer is essential: for example, variables used to correct the sample or those in press readership surveys that deal with information that is essential in calculating readership rates.

c) Maintaining interviewee interest:

One of the difficulties in online studies is ensuring a high participation rate: you have to motivate people to participate in the first place but you also have to maintain their interest and level of involvement throughout the questionnaire in order to limit abandons during the process of filling it out. We found out during "qualitative" re-interviewing of Internet users who had previously taken part in Web surveys that 77% had already stopped filling out an online questionnaire. Questionnaire length, technical problems (slow response, errors…) and even the subject of the questionnaire were of course all pointed to as the main reasons for abandoning the process. **But among the other important reasons for abandon were poor presentation or layout of the questionnaire (15% of Internet users) and the overly-directive tone of the instructions (13%).**

d) Various methods of filling out questionnaires:

Respondents in the "qualitative" re-interview phase also signaled very different degrees of concentration and reflection.

Read the question entirely before answering	60%
Go to the essential and read online questionnaire questions quickly	32%
Answer with my initial impression	46%
Take time to think before answering	45%

This variance in methods used to fill out questionnaires undoubtedly has an effect on the quality and precision of the responses. Optimizing questionnaire layout and ergonomics would certainly be one good way to reduce these risks and lead interviewees, without boring them, to take the most uniform approach possible to the entire questionnaire. This condition is capital for a

[3] They were sent an alert message if they failed to answer and were unable to continue the questionnaire without completing the question.

standardized measurement survey such as that for Press readership, one of whose cornerstones is providing comparable data regardless of the publication or interviewee involved.

4 – In 2009: tests on press readership measurements:

a) Presentation of the tests methodology

These were conducted over 3 weeks in May, 2009.
We reproduced the complete sequence of the 3 key audience measurement questions based on the methodology currently used in France:

- Starting with **a filter** concerning readership over the past 12 months,
- Followed by all publications that passed the filter:
  - a question **about readership habits (frequency)** – scale with 5 items,
  - finishing with a question about **Recent Reading** (scale with 7 items) making it possible to calculate Average Issue Readership.

Two distinct layouts for this sequence of 3 questions (filter, frequency, date of RR) were proposed randomly to the two sub-populations.

Since these tests were included in surveys for which they were not the main subject, and because of the large number of questions in each survey, we chose to limit the number of publications proposed to interviewees. But we also felt it was important that the publications we chose were evaluated within a competitive universe that was as complete as possible.

We chose to limit ourselves to 2 press families that were large enough to ensure a sufficient number of respondents, with one of the families presenting some risk of confusion.

In addition, in order to isolate the effects of questionnaire layout on individual responses as much as possible, and with "everything else being equal", we decided to avoid increasing the number of outside factors that could lead to variance in the results. We therefore decided to concentrate on a single issue frequency, which presented the advantage of requiring only one answer scale each for the average and recent readership questions.

We decided on two families of weekly publications: TV magazines (12 publications) and news magazines (6).

- **Design of the 2 test questionnaires:**

In France, color logos of the survey publications are currently used as visual stimuli in major Press readership surveys to aid recognition of the test publications. In order to minimize the effects of presentation order, we randomly rotated the publication families and the magazines in each family. Finally, we required responses to readership questions. We chose to adopt all of these choices in both test versions.

What changed in the 2 versions:

The first layout could be considered "classic". It presented questions in the form of tables with the names of the publications and their color logos on the same line with the item responses in columns. This layout is the same as that used for the Audipresse Premium survey covering managers, executives and high-income individuals, and whose questionnaire is on line.

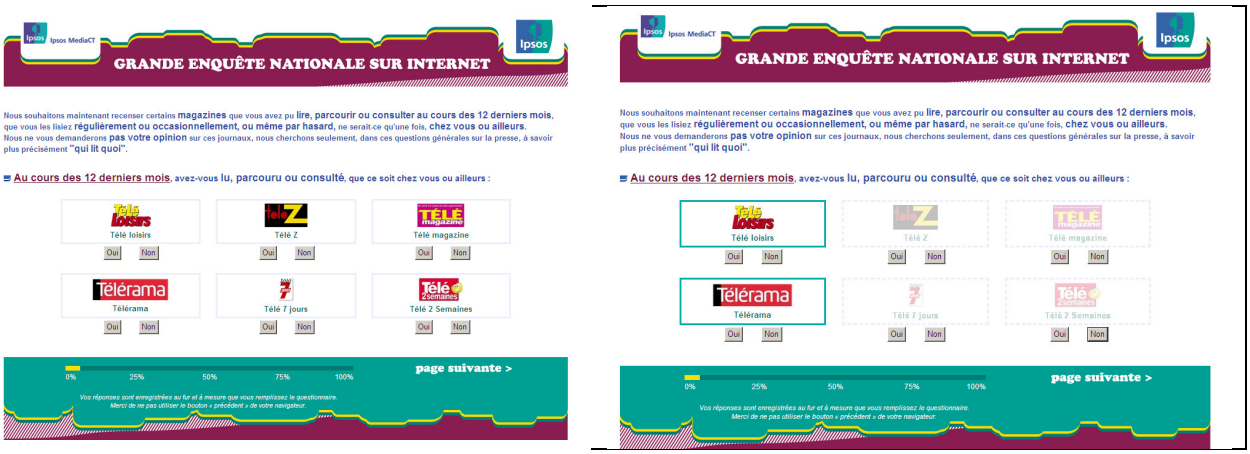
Version 1: 12 months filter question

Version 1: frequency

The second layout was somewhat innovative because it used flash animation and a unique screen format. In some ways it was

similar to the AEPM survey (national press readership survey)[4]: a filter with simultaneous presentation of the logos, with the average and recent readership questions asked publication by publication.

At the filter question level: the publications were presented in groups of 6, but the question was asked individually with a required "yes" or "no" answer for each publication/logo presented. In order to help interviewees navigate the questionnaire, the publications they said they had read were framed in blue and those they had not read were lightly shaded.



Version 2 : 12 months filter question

For the frequency and recent readership questions:

Every publication read in the past 12 months appeared in the center of the screen with its logo. The answer scale remained in the same position at the bottom of the page. Respondents were asked to click on the buttons corresponding to their choices and the animated logo disappeared behind the chosen score.

Version 2 – frequency question

[4] Presented by JL Marx et O Lê Van Truoc at the Vienna WRRS - 2007

Version 2 – recent reading question

- **Sample populations obtained:**

The total sample population included 2285 interviews, with the first sub-population ("classic" method) including 1136 respondents and the second one ("innovative" method) including 1145 respondents.

The two sub-populations were corrected (weighted) posteriorly based on a battery of socio-demographic questions (gender, age, occupation, residential region and agglomeration category …).
After those corrections, we verified that they were perfectly matched for other criteria not included in the quotas and linked by media behavior (especially in the area of news magazines and TV : Internet connection frequency, sites visited, centers of interest (politics, economy, travels, technology, etc…), tv consumption habits…. We thus had two closely comparable sub-populations.

b) Selected results:

b.1) Comparison between online tests readership and AEPM (National Readership Survey) references:

The tests were not designed to be directly comparable with the AEPM survey of magazine readership in France: the way in which the test sample populations were constructed included only Internet users, and in light of the recruitment method, probably mainly regular Internet users. We thus excluded from the test samples a large part of the French population (at least 35% non-Internet users) with specific readership habits.

The test fieldwork was also conducted during a short period and not continuously over 12 months. Finally, the list of test publications (18) was much smaller than that covered by the AEPM magazine readership measurement study (190), and the readership questions were included in a questionnaire covering a large number of subjects unrelated to press readership.

Beyond the data collection method (online and self-administered vs. offline and administered) there were therefore notable differences in interview conditions and undoubtedly in the respondent population.

Even so, it seemed incongruous to us not to investigate the readership figures in the tests relative to the reference ones at a time when French press media are considering a make-over of all their audience measurement surveys and when the implementation of online readership data collection can be considered. Even if we don't compare results in detail, we can focus on the orders of magnitude of the various indicators measured. The major trends that emerge are often instructive when considering the implications of a survey method.

In order to account for the effects of online sampling as much as possible, we chose to isolate regular Internet users (at least one connection per day) in the AEPM study. The AEPM readership figures are, of course, calculated solely on the basis of the test publications.

Readership filter results that are markedly lower, different regular and recent readership rates, but comparable orders of magnitude:

The number of publications interviewees said they had read was sharply lower in the test than in the AEPM, regardless of the test version: this result was not totally surprising except in terms of amplitude. This is undoubtedly a result of the large differences in interview conditions: AEPM is administrated and the interviewer is very present. The CAPI Double Screen system

requires the interviewer to give numerous standard prompts, and the length of time devoted to the filter questions is monitored closely to ensure that sufficient attention is paid to each publication. The filter is thus a fairly long and repetitive litany of "yes" and "no" answers and, as pointed out by JL Marx (…), the large number of publications that pass the filter can possibly be attributed to "acquiescence bias".

In an online interview, the interviewee is free to pay attention to the question he or she wants. And even if we require interviewees to respond to each publication, they are free to focus on the essential, meaning the publications they are most familiar with. For example, the additional readers over 12 month in the AEPM are chiefly (very) occasional readers.

	Online Tests (average)			**AEPM base: Regular Internet users**
Index: Base 100 = AEPM regular Internet users	***Number of publications passing the 12 month filter***	***Number of publications read regularly (every week + 2 to 3 times monthly)***	***Number of AIR publications read (in past week)***	
Total magazines	*59*	*85*	*94*	*100*
TV weeklies	*55*	*71*	*79*	*100*
News	*64*	*150*	*158*	*100*

Even if there are differences between publications, the numbers of regular readers are much more comparable (index of 85 vs. AEPM) and AIR readership even more so (in past week). This observation appears to vary significantly by publication family (TV vs. News), families whose readership structures vary considerably (a much larger percentage of regular and recent readers for TV publications, which accompany viewers on a daily basis). But it is difficult to draw formal conclusions about the data collection method from these results because the variance could stem from differences in sample populations or question environments.

b.2) Comparison of the 2 layouts ("classic" / "more innovative"):

At first look, there are few evident differences in readership levels between the 2 layouts for the aggregated results. On average, interviewees declared a few more publication for the 12-month filter in the "classic" version 1, and fewer publications read regularly and recently.

	Version 1: "classic" layout (tables…)	***Version 2: more "innovative" layout***	***ratio layout 2/ layout 1 (index)***
Average number of publications read			
12 months			
total	3,35	3,19	***95***
TV	1,78	1,79	***100***
News	1,57	1,41	***90***
N° of publications read regularly			
total	1,34	1,45	***109***
TV	0,82	0,96	***117***
News	0,52	0,49	***95***
N° of publications read recently (AIR)			
total	1,40	1,45	***104***
TV	0,83	0,95	***114***
News	0,56	0,49	***90***

But we also see more marked differences by publication family studied, which we should not be able to clearly attribute to sampling effects: it is of course a possibility, but the two sub-samples appeared to be quite perfectly matched for criteria linked to centers of interest (politics, economy,…) and TV habits.

Publication by publications, there are only few significant differences between the results obtained with each version: 3 out of 18 publications for past 12 month filter, and 4 out of 18 for regular and recent readership.

- **A few results to consider further:**

These overall results mask the differences we find when we evaluate the results in detail, differences that possibly stem from the layout:

Two examples:

1- Thus, when we look at the number of publications read at the level of the 12-month filter, we find two significant differences for TV magazines: substantially more people say they have read just 1 publication over the past 12 months with the "classic" method (version 1) and significantly more say they have read 2 publications over the same period with the method 2 using screens and flash animation. The rest of the distribution is perfectly similar.

N° of TV magazines read over the past 12 months		
	layout 1	**layout 2**
	%	*%*
0	*22*	*22*
1	***34***	***30***
2	***19***	***24***
3	*9*	*9*
4	*7*	*6*
5+	9	9

In fact, there is a single screen for all 12 TV magazines with method 1 while method 2 requires 2 (we presented 6 logos per screen). The very large majority of people who said they read 2 publications with method 2 chose their magazines from 2 distinct screens, almost as if they weren't sure there would be a second screen and they chose a magazine from screen 1 just in case because they knew they were readers of one TV weekly or another.

Method 2, which was intended to be more interactive and modern, possibly created bias concerning this family of publications that is subject to confusion.

2- When we run a more detailed analysis of the distribution of responses for questions with scoring scales (habits – 5 bars, and RR, 7), we also find surprising differences between the two layouts.

Average distribution (%)	**Version 1**	**Version 2**	**Ratio 2/1**
	100%	*100%*	***100***
. Every week	***24***	***30***	***124***
. 2 /3 times per month	***16***	***16***	***100***
. 1 time per month	***15***	***15***	***103***
. 5 to 6 times per year	***21***	***21***	***99***
. Less often	***24***	***18***	***74***

We find almost 25% fewer scores for the first level of "every week" in the "classic" version and almost 25 % more scores for the last level of "less often".

This observation holds true both for TV and for News publications.

We find the same trends, of a lesser amplitude but in the same direction, for Recent Reading responses: the first item, "yesterday", was checked approximately 15% less often in the "classic" version while the opposite was true of the last item, "a long time ago".

Of course, we might say that this is a case of unfortunate luck and that we are dealing with sample populations with different readership structures. But these differences are so significant while being symmetrical at the scale extremes that we should also investigate the possible influence of the layouts.

The layout differences between the two methods touch several aspects:

First, let's look at tables vs. stationary scales. In the second instance, interviewees simply have to move their cursors to the proposed items rather than move down a table line by line. In addition, all of the publications that pass the filter are presented together in a classic table while the question is really asked magazine by magazine with the second option. To caricaturize, we could say that we are more in a position of comparing publications and thus asking for relative responses in version 1 and perhaps according more attention to individual publications in version 2.

In fact, when we analyze the variability of responses to the readership habit question as well that for Recent Readership for given individuals, we find significant smaller standard variations with method 2 which uses scales than with the "classic" method using tables.

Next, in our test the two layouts had an incidence on the "visual" amplitude of the scales: there is a smaller distance between the scale extremes in the "classic" table version (approximately 50% less "distance" for habits and 25% less for recent readership) compared to the scale system used in version 2. We saw earlier that the effort required to select items could have an impact on the answers. So, we can reasonably ask ourselves whether this difference in distances when "searching" for the extreme items might have had a slight incidence on their selection and on the dispersion/ the variation of the answers.

Moreover, when programming the questionnaire, there was a form mistake, as the "answer buttons" have not the same size: for instance "every week" button was larger than "less often" one. We can not prove that it had an influence on the answers, but the results we found should remind us to be cautious when considering formal aspects, even tiny ones.

5) In rapid conclusion:

Our program of tests showed that layout has an incidence on the results collected. This is fairly obvious in certain of the experiments we conducted, notably those covering scoring scales.

These effects, as well as online respondents comments concerning their attitudes toward out questionnaires, confirm our thinking that research institutes should not be satisfied to merely *"ask the right questions to the right people"* but should also ask the questions in the *"right"* way. This calls for investigating, as has been done in previous work done with interviewers effects and interview situations, the formal aspects – in a large sense – of the questionnaires, the prompts, the instructions, the ergonomics, etc… in our self-administered online surveys. In fact, these elements are our sole means of interacting with / persuading / guiding respondents.

These observations are undoubtedly particularly well suited to readership surveys that are primarily standards and that require, more than other surveys, a high degree of uniformity in the conditions under which the information is collected.

Our first experiments covering readership questionnaires have not yet yielded conclusive results. Our tests were undoubtedly (too) ambitious and could have required other means: in particular, we can suppose that the repetitive readership mechanism generates noticeable effects over time, that layout problems are thus more evident when the list of publications is longer, leading to boredom or fatigue, and that our selection wasn't sufficiently large to clearly uncover these effects. In addition, we consciously chose layouts that were similar to those of currently used methodologies, which are not radically different in conception. This undoubtedly made it more difficult to discern variations.

In fact, we probably should have run an experiment that was entirely dedicated to readership measurement, using a large publications list, and with sharper differences in our layout choices.

Even so, these first experiments uncovered a few paths for reflection. Firstly, a word of warning: switching from a system that is tightly controlled by the interviewer to one that is self-administered (this seems evident but merits repeating) presents risks of significant divergence: the large drop in the 12-month filter figures in our tests is undoubtedly an indicative symptom.

Second, even though our exploration is incomplete, it appears that differences in layout, even though they were apparently fairly subtle in this case, this from an organizational point of view (presence or absence of other publications / number of publication per screen) or from an ergonomic point of view (distance to be "traveled" / "effort" needed to respond), can have an impact on certain press readership results. We believe it is important to be aware of this when designing or modifying the methodologies of surveys that are forms of "currency".