MEASURING THE HOMOGENEITY OF INTERVIEWERS' PRACTICES IN READERSHIP SURVEYS

Antoine Moreau, IMS

Introduction

Ensuring that all interviewers have the same way of interrogating interviewees is a key element in the field management of a survey. Supervisors and standardized training and briefing are there to homogenise the interviewing practices. But, much probably, interviewers tend not to work in the same way at the beginning of a survey and after having worked for a few days on it.

Controlling for this is all the more important for readership surveys, for which field may last several months. In France as in most countries, interviewers are on short time contracts, which ensures a great flexibility in the way the studies can be scheduled, anticipated, postponed, cancelled… But this allows also interviewers to leave the company on very short notice. Due to differential turnover in the population of interviewers, the number of weeks a given interviewer is working on a readership survey may vary greatly across the team of interviewers allocated to that survey.

In this paper, we try to answer the following questions: What is the influence of the number of questionnaires already done by the interviewer on the way he is collecting data? Should there be more briefings than the initial one for interviewers who work more than a certain amount of weeks on the survey? More generally, are some interviewers systematically different, in terms of measured audience, from the majority of the interviewers population?

Simple tabulations cannot answer correctly these questions, since they do not allow one to give an appraisal other things being equal. Using data from the French readership survey on dailies for 2000, we estimate a model linking the audience with the characteristics of the interviewee and variables neutralising the effect of the interviewer. This allows us to estimate the individual effect of each interviewer on the measured audience as well as assessing the way the duration of the spell of work on the same survey for a given interviewer may affect the results.

The PQR readership survey

The PQR readership survey is conducted each year by Ipsos in France. It measures the readership of all national and local dailies and is the reference survey for this type of medias. The field takes place throughout the year, with a given number of interviews to be done each month. The audience data are all collected during a CATI interview. Then a questionnaire is sent to each respondent, which is a purely self administered phase. Here, of course, we are only interested in the CATI phase, when the interviewee answers about the readership.

There are two main sessions of briefing for this survey: one in January/February and one in September/October. Interviewers are mainly allocated to the survey at these two times. Thus, an interviewer might work during the whole year for the survey, or work only for the last quarter, with various intermediate possibilities arising in between.

The variance in the duration of work of interviewers on the survey may be measured in various ways. First of all, one can look at the duration between the first interview and the last interview done by an interviewer, which is done in table 1 below.

Table 1
Distribution of the number of interviewers according to the duration between the first and the last interview

Less than 15 days	6
Between 16 and 30 days	5
Between 31ans 60 days	2
Between 61 and 90 days	9
Between 91 and 110 days	17
Between 111 and 130 days	6
Over 131 days	33
Total number of interviewers	**78**

Table 2 below gives another insight on that matter: it displays the distribution of the number of questionnaires done by one interviewer across all interviewers.

Table 2
Distribution of the number of questionnaires across interviewers

Less than 40 questionnaires	8
Between 41 and 90 questionnaires	8
Between 91 and 159 questionnaires	9
Between 160 and 289 questionnaires	26
Over 290 questionnaires	27
Total number of interviewers	**78**

The model

We want to model the answer of individuals to audience questions. These are 0/1 questions: either the individual has read the daily or he has not. Thus we have to turn to dichotomous models to analyse our data. We want to estimate the response function of the individual, linking the fact he has read or not the daily with his characteristics, x. We also of course want to add to the model a measure of the workload of the interviewer, y. Thus the model is of the following form:

$$\text{Audience of individual i interviewed by interviewer j} = f(x_i, y_j)$$

The simplest way to do that is to estimate a logit model. The probability that an individual has read the daily will be written using a logistic distribution function as followed:

$$\text{Probability of reading of individual i interviewed by interviewer j} = 1/(1+\exp(x_i'\beta+y_j'\gamma))$$

The model is estimated using the maximum likelihood method, using a stepwise method to eliminate variables which are not significant at the 5% level.

The audience variable we use in the model is the audience over the last twelve months: has the individual read the paper during the last 12 months?

The explanatory variables entered in the model are the following:

1. Variables characterising the individual:

- sex,
- age in 12 classes,
- occupation of interviewee,
- educational level of interviewee,
- income of interviewee,
- size of household,
- child less than 15 years old in household,
- type of housing,
- size of town,
- renter/owner,
- département (French local district),
- time out of home the day before,
- day of the week,
- week of interview.

2. Variables characterising the interviewer:

- number of questionnaires done since the beginning of the survey at the time of the interview,
- duration since the first interview,
- a dummy variable which is equal to 1 for a given interviewer and 0 for the others. This variable will summarize the individual effect of the interviewer, controlling for the characteristics of the individual and the workload to date of the interviewer on the survey.

In fact we estimate 4 models. Preliminary estimations have shown that the effect of sex and age are not additive with some other variables. Instead of introducing crossed effects in the model, we preferred to split the estimation according to these two criterias. Thus the assumption about the link between these two variables and the probability of readership is less stringent. We split the initial sample between four sub samples: women under 34, women over 35, men under 34 and men over 35. All the above variables, including the relevant classes of age, are introduced in those models.

Table 3 below give the size of our sub samples.

Table 3
Size of samples used for the estimation

Women under 34	3 366
Women over 35	6 772
Men under 34	3 401
Men over 35	5 989

Measuring the quality of the model

As in a regression model, it is possible to calculate, within the framework of a logit model, statistics on the accuracy of the model. A is well known, these statistics, if calculated within the model, are widely misleading. They overestimate the quality of the model because the estimation method is built to be optimal for the set of data analysed.

Thus, to measure the quality of our estimation, we used cross validation methods. A group of individuals is set apart. The model is estimated without these individuals and their readership probability is then computed. Comparing the probability given by the model and their actual answer gives a quite accurate measure of the adequacy of the chosen model.

To fully implement this cross validation method, two problems must be tackled:

- We set apart 10% of the sample as the control cell which will be used for cross validation. But the individuals in this control cell cannot be chosen completely at random. Since we want to include in our model a variable summarizing the effect of each interviewer, we must have at least one observation for each interviewer in our final sample. Thus, we choose at random 10% of the interviews of each interviewer to be part of the sub sample which is set apart.

- The response of the individual to the audience question is a 0/1 variable. How are we to compare it to the probability computed with the model? The way we choose to do that is the following. We can compute the average audience of the total sample, which gives the average probability of readership. Then we consider that the model is correct in predicting the response of an individual if:

1. He has read the paper and the probability given by the model is greater than the average probability of reading.

2. He has not read the paper and the probability given by the model is smaller than the average probability of reading.

The model is in fact estimated 25 times and the cross validation method is used each time. Thus we ensure that different sets of individuals have been eliminated from the estimation. One should recall that most of the data analysis methods, including the estimation of logit models, are sensitive to outlier values in the sample. If one estimation is, in a positive or a negative way, influenced by the elimination of one particular individual, this is mitigated by the estimation of several models and the computation of the quality statistics on all of them.

We have estimated our models for two groups of dailies, for which the quality of the models are presented below.

Table 4
Percentage of correct predictions by the model

	Group 1	**Group 2**
Women under 34	70,5%	83,2%
Women over 35	76,2%	58,8%
Men under 34	80,1%	59,8%
Men over 35	60,2%	67,3%

The cross validation method is a quite stringent test of the accuracy of the model. A 70% percentage of good predictions is a good result. 4 of our models thus give quite accurate predictions of the audience. In the other 4, it is clear that some variables are missing in the model. Since we only are interested in the effects of the variables related to interviewers, this is not crucial. The missing variables most probably relate to the interviewee and are certainly not correlated with the variables characterising the interviewer.

A first insight in the issues we are interested in can be made at this point. We have estimated the models without the interviewers variables. By computing in the same way percentages of correct predictions, one can asses the explanatory power of these variables. This is done in the following table.

Table 5
Explanatory power of the interviewers variables: percentage of good predictions

	Group 1 – with interviewers variables	**Group 1 – without interviewers variables**	**Group 2 – with interviewers variables**	**Group 2 – without interviewers variables**
Women under 34	70,5%	71,3%	83,2%	81,2%
Women over 35	76,2%	77,3%	58,8%	60,2%
Men under 34	80,1%	80,1%	59,8%	58,8%
Men over 35	60,2%	57,4%	67,3%	67,5%

It may seem strange that, for some models, the estimation without the interviewers variables may give better results than the estimation with them: adding variables to a model does not necessarily improve it, but it certainly cannot deteriorate its performance. Recall here that we do not present any standard errors for the above. From table 5, we can safely conclude that the performance of models with and without variables related to the interviewer are roughly the same.

This does not mean that there is no heterogeneity in the measured audience across interviewers, as will be shown in the next section. It rather indicates that, even if this heterogeneity exists, there is also a lot of variance within the work of interviewers and that, on an aggregate level, interviewers effects tend to cancel out. This was not at all certain and comes as good news for the quality of the survey.

Estimation Results

The output of a logit model, as in a regression model, is a coefficient for each variable, associated with a standard error, which allows one to calculate if the variable is significant. Using the estimated values of the coefficients, one can then compute the estimated probability of reading of each individual.

Of course, it is much easier to assess the results using these estimated probabilities than the coefficients: these are linked in an intricate way to the probability and, except for its sign, it is difficult to know only from the value of a coefficient the magnitude of the effect of the associated variable on the readership probability.

Thus we prefer to present our result in terms of probabilities. These are computed in the following way. Suppose for example the first class of age is significant at 5%. For all individuals in the sample, we compute the estimated probability of readership if he were in this class of age and if he were not, others things been equal. We then average these probabilities other all the sample, which gives the effect of the variable, with the other variables set to their average.

This is done 25 times, as before. The minimum et maximum probabilities obtained for a given variable allows one to assess the precision with which the effect of the variable is estimated. These are presented in the appendix.

In this paper, we concentrate on the variables related to the interviewer. We only present results for these variables. A full modelling of the audience, in which one would focus also on the variables related to the individual, should take into account duplication, which is beyond the scope of this paper. Moreover, in the tables below, we only display the variables which are significant at the 5% level and that appear in at least 20 of the 25 estimated models. Quite a few variables appear in a small number of models, which justify a posteriori our cross validation method. It is only by sampling replication that one can eliminate spurious effects that are particular to one given sample.

Since we concentrate on the variables related to the interviewer, 3 types of variables can be found in the following tables:

- number of questionnaires done since the beginning of the survey at the time of the interview. This variable is split in 5 classes: less than 40 questionnaires, between 41 and 90 questionnaires, between 90 and 160 questionnaires, between 160 and 290 questionnaires, over 291 questionnaires

- duration since the first interview. This variable is split in 7 classes: less than 15 days, between 16 and 30 days, between 31 and 60 days, between 61 and 90 days, between 91 and 110 days, between 111 and 130 days, over 131 days.

- dummy variables summarizing the individual effect of the interviewer. For each interviewer, we indicate in brackets how many questionnaires he has conducted.

Table 6
Effect of the significant variables on the probability of reading
Group 1

	Number of models in which the variable is significant	Probability
Women under 34		
Average probability of reading		**69,1%**
Interviewer 8391 (726)	25	60,7%
Interviewer 8420 (571)	25	60,8%
Interviewer 8593 (107)	25	46,4%
Interviewer 8880 (175)	25	53,4%
Interviewer 11238 (375)	25	80,3%
Interviewer 11832 (115)	25	86,4%
Interviewer 11863 (521)	25	58,3%
Interviewer 12139 (635)	25	61,4%
Interviewer 12835 (122)	25	41,1%
Interviewer 13617 (275)	25	56,6%
Interviewer 13834 (112)	25	84,5%
Interviewer 16258 (96)	25	92,6%
Interviewer 16553 (188)	25	51,3%
Interviewer 16925 (189)	25	45,1%
Number of questionnaires: Between 91 and 159 questionnaires	24	65,1%
Duration: Between 91 and 110 days	22	66,1%
Women over 35		
Average probability of reading		**59,1%**
Duration: Between 111 and 130 days	25	52,9%
Interviewer 16925 (189)	25	41,3%
Men under 34		
Average probability of reading		**80,9%**
Duration: Between 31 and 60 days	25	72,6%
Interviewer 4963 (467)	25	69,5%
Interviewer 12528 (457)	25	69,9%
Interviewer 13602 (456)	25	91,2%
Interviewer 13606 (637)	25	70,5%
Interviewer 16010 (56)	25	45,0%
Duration: Less than 15 days	23	75,0%
Men over 35		
Average probability of reading		**72,1%**
Number of questionnaires: Between 31 and 60 questionnaires	25	74,7%
Duration: Less than 15 days	25	77,1%
Interviewer 8463 (475)	25	65,3%
Interviewer 11651 (320)	25	64,6%
Interviewer 11863 (521)	25	79,2%
Interviewer 11864 (400)	25	83,0%
Interviewer 12139 (635)	25	80,3%
Interviewer 12365 (21)	25	93,4%
Interviewer 16053 (258)	25	84,6%
Interviewer 16552 (282)	25	80,9%
Interviewer 16880 (358)	25	88,7%
Interviewer 16924 (294)	25	80,6%
Interviewer 31191 (29)	25	53,7%
Interviewer 8420 (571)	24	74,9%
Interviewer 12297 (162)	24	65,6%

Table 7
Effect of the significant variables on the probability of reading
Group 2

	Number of models in which the variable is significant	Probability
Women under 34		
Average probability of reading		**83,9%**
Interviewer 8391 (726)	25	94,7%
Interviewer 8787 (409)	25	68,1%
Interviewer 8880 (175)	25	71,0%
Interviewer 11651 (320)	25	71,0%
Interviewer 12366 (180)	25	60,1%
Interviewer 13617 (275)	25	73,1%
Interviewer 373 (4)	24	81,7%
Interviewer 12227 (297)	20	78,8%
Women over 35		
Average probability of reading		**82,4%**
Interviewer 8783 (527)	25	71,5%
Interviewer 11675 (163)	25	62,5%
Men under 34		
Average probability of reading		**81,9%**
Interviewer 5144 (174)	25	64,3%
Interviewer 12366 (180)	25	67,6%
Interviewer 12422 (80)	25	51,5%
Interviewer 13617 (275)	25	64,5%
Interviewer 13835 (246)	25	71,5%
Interviewer 13836 (357)	25	67,9%
Interviewer 11651 (320)	24	74,9%
Interviewer 12835 (122)	21	73,2%
Interviewer 8880 (185)	20	73,6%
Men over 35		
Average probability of reading		**84,9%**
Duration: Between 91 and 110 days	25	90,4%
Interviewer 11651 (320)	25	75,7%
Interviewer 12135 (553)	25	68,5%
Interviewer 12227 (297)	25	75,5%
Interviewer 13600 (59)	25	60,6%

Conclusion

The conclusion from these estimations can be summarized as follows:

There is indeed some heterogeneity in the way interviewers are collecting audience data in the PQR survey. On the aggregate level, there is no evidence that this heterogeneity leads to serious biases of the measured audience.

Although sometimes, the length of time since the first interview and the number of questionnaires done to date by the interviewer have some influence on the measured audience, the effects are not clear cut and of less magnitude than the individual effects of interviewers.

Some interviewers measure the audience in significantly different ways than the majority of them. One should note that, across different type of interviewees, the effect is not always the same: some interviewers seem to underestimate the audience on a certain group and to overestimate it in other groups. The same conclusion holds for the type of dailies. More than repeated training sessions, the supervisor's job is then crucial to the quality of the fieldwork.

Appendix: Precision of the results

All the models have been estimated 25 times on a subsample of the original sample. In the Estimation Results section of the paper, we presented the effect of each significant variable in the model by computing the average readership probability associated to that variable, others things being set at their average.

Since we computed 25 probabilities, the minimum and maximum probability across the 25 estimations give a good indication on the precision of our estimations. The tables below give these minimum and maximum probabilities.

Table 8
Effect of the significant variables on the probability of reading
Group 1

	Minimum	Maximum
Women under 34		
Interviewer 8391 (726)	59,6%	61,8%
Interviewer 8420 (571)	60,1%	61,6%
Interviewer 8593 (107)	43,0%	49,8%
Interviewer 8880 (175)	51,4%	55,5%
Interviewer 11238 (375)	79,6%	81,0%
Interviewer 11832 (115)	85,5%	87,4%
Interviewer 11863 (521)	57,0%	59,7%
Interviewer 12139 (635)	60,2%	62,6%
Interviewer 12835 (122)	39,2%	43,0%
Interviewer 13617 (275)	55,1%	58,1%
Interviewer 13834 (112)	82,4%	86,6%
Interviewer 16258 (96)	91,9%	93,2%
Interviewer 16553 (188)	49,9%	52,8%
Interviewer 16925 (189)	43,3%	47,0%
Number of questionnaires: Between 91 and 159 questionnaires	61,2%	69,1%
Duration: Between 91 and 110 days	63,2%	69,1%
Women over 35		
Duration: Between 111 and 130 days	52,8%	53,0%
Interviewer 16925 (189)	40,0%	42,6%%
Men under 34		
Duration: Between 31 and 60 days	70,4%	74,9%
Interviewer 4963 (467)	67,4%	71,6%
Interviewer 12528 (457)	68,0%	71,7%
Interviewer 13602 (456)	90,7%	91,8%
Interviewer 13606 (637)	69,3%	71,8%
Interviewer 16010 (56)	37,6%	52,4%
Duration: Less than 15 days	69,2%	80,9%
Men over 35		
Number of questionnaires: Between 31 and 60 questionnaires	74,4%	74,9%
Duration: Less than 15 days	76,4%	77,9%
Interviewer 8463 (475)	64,8%	65,9%
Interviewer 11651 (320)	63,7%	65,6%
Interviewer 11863 (521)	79,0%	79,5%
Interviewer 11864 (400)	82,8%	83,3%
Interviewer 12139 (635)	80,0%	80,6%
Interviewer 12365 (21)	93,1%	93,7%
Interviewer 16053 (258)	84,2%	85,1%
Interviewer 16552 (282)	79,1%	82,7%
Interviewer 16880 (358)	88,4%	89,0%
Interviewer 16924 (294)	79,9%	81,4%
Interviewer 31191 (29)	52,3%	55,1%
Interviewer 8420 (571)	72,2%	77,5%
Interviewer 12297 (162)	59,1%	72,1%

Table 8
Effect of the significant variables on the probability of reading
Group 2

	Minimum	Maximum
Women under 34		
Interviewer 8391 (726)	94,6%	94,9%
Interviewer 8787 (409)	67,5%	68,8%
Interviewer 8880 (175)	70,2%	71,7%
Interviewer 11651 (320)	70,1%	71,8%
Interviewer 12366 (180)	56,2%	64,0%
Interviewer 13617 (275)	71,8%	74,5%
Interviewer 373 (4)	79,4%	83,9%
Interviewer 12227 (297)	73,7%	83,9%
Women over 35		
Interviewer 8783 (527)	71,3%	71,7%
Interviewer 11675 (163)	62,2%	62,8%
Men under 34		
Interviewer 5144 (174)	61,7%	66,9%
Interviewer 12366 (180)	66,1%	69,0%
Interviewer 12422 (80)	47,6%	55,4%
Interviewer 13617 (275)	63,1%	65,8%
Interviewer 13835 (246)	70,7%	72,4%
Interviewer 13836 (357)	66,7%	69,1%
Interviewer 11651 (320)	67,9%	81,8%
Interviewer 12835 (122)	64,7%	81,8%
Interviewer 8880 (185)	65,2%	71,8%
Men over 35		
Duration: Between 91 and 110 days	90,1%	90,7%
Interviewer 11651 (320)	74,7%	76,7%
Interviewer 12135 (553)	67,9%	69,1%
Interviewer 12227 (297)	74,9%	76,2%
Interviewer 13600 (59)	58,9%	62,3%