Measuring Interviewers’ Performance in Telephone Surveys

By

Claire Durand,
department of sociology,
University of Montreal

Paper presented at the 57th Annual AAPOR Conference,
St-Pete’s Beach, Florida
May 20, 2002

© Claire Durand, 2002
Table of contents

Research goals ........................................................................................................ 1

Data: ......................................................................................................................... 2

Elaboration of the performance index ................................................................. 2
  First step : characterize each call ..................................................................... 2
  Second step : produce the interviewers' daily work load file ......................... 3
  Third step : produce the performance index .................................................. 3

Results : Properties of the performance index .................................................. 3
  Missing values ..................................................................................................... 3
  Variation .............................................................................................................. 3
  Comparison between the performance index and the cooperation rate ............ 4
  Possible bias of the index : length of questionnaire ......................................... 5

Results: relationship with external predictors ...................................................... 6
  Relationship with the interviewer’s experience in the projects ....................... 6
  Relationship with the characteristics of the task .............................................. 6

Conclusion: .......................................................................................................... 7
Measuring Interviewers’ performance in telephone surveys
Claire Durand, University of Montreal

Most measures of performance in the survey field have been devised for face to face surveys, where each interviewer has a number of areas and households to visit and “follow” each “client” until finalization into either completion, refusal, no response, unavailability or some “out of sample” categorization. The situation is quite different with telephone surveys conducted from centralized call centers. The concentration of interviewers as well as the computerization of all the field work using CATI has gradually brought a different way of conducting surveys where any interviewer may inherit of another interviewer’s appointment, depending on its availability at one moment and of his work schedule. Interviewers also conduct refusal conversion: they call households where a previous refusal has been recorded and try to reverse the previous decision and complete a questionnaire with the selected respondent.

When researchers started to examine the relationship between interviewers' characteristics, behaviors and attitudes and their performance, much emphasis was put on which determinants had to be examined and how to measure different attitudes and behaviors. Less emphasis was put on how to measure performance. The measure of performance used in most research is the cooperation rate at first contact. It is computed as the ratio of completed interviews on the sum of completed plus refusals but only the calls where the interviewer is the first to have a contact with the household are taken into account. This prevents from attributing to an interviewer a good or bad performance that was "prepared" by another interviewer. The obvious problem with this measure is that it does not allow the measurement of performance when the result is not obtained from a first contact. Therefore, when interviewers convert refusals, their performance is not assessed. It is even impossible to calculate any cooperation rate for the interviewers who conduct refusal conversion for a whole shift. Therefore, the performance of the best interviewers is not measured properly. It becomes obvious that a good measure would have to take into account the work load of each interviewer in terms of the type of task that they get attributed.

Research goals

This research aims at solving the identified problem. In order to do this, a new measure of interviewer performance is elaborated. An appropriate measure should, in our view, have the following properties
a) has to take into account not only the actual result of a call but also the previous result for each phone number that has been called before.
b) has to be valid for all types to tasks/ work loads
c) has to vary and discriminate between interviewers (reliability- precision)
d) has to be related to known predictors of performance (external validity)
e) has to be unbiased for different projects, shifts lengths, be comparable within projects and ideally between projects and between firms (reliability across time and space)
**Data:**

The data for this research comes from 10 different surveys conducted by three different pollsters during the Canadian electoral campaign of October-November 2000. The data sets from each survey comprise all the calls that were made as well as the details of each call: time and length of call, result, identification of the interviewer who called, etc. The data from interviewers who worked only one day and the data for the days where an interviewer worked less than two hours, were considered outliers and dropped from the data sets. We end up with the data for 10 surveys, 190 interviewers (26, 124 and 40 respectively for the three firms) and 977 interviewer-days.

**Elaboration of the performance index**

Three steps have to be performed in order to compute the performance index. First, each call has to be characterized in order to determine its status; second, the measures of each interviewer’s daily work load have to be computed and finally the daily performance index can be computed.

*First step: characterize each call*

In order to achieve our goal, the first step is to portray each call not only in terms of the result of the call but also in terms of the initial state of each call. In order to do this, technically, it is necessary to sort the administrative data set1 in order to input the previous result of each call to the actual call. This can be done rather easily using the lag function, as long as the data set is sorted by phone number and then sequentially. When the call is the first for a given phone number – and therefore there is no lag – a special code is attributed. Therefore, in the final data set, each line contains the actual as well as the previous result of each call. It is then possible to compute a variable giving the task performed each time a number is dialed. A call resulting in a completed interview can be categorized as completed from a never reached phone number (new or never answered), completed from appointment or completed from refusal. A refusal can similarly be categorized as a refusal from a never reached phone number, a second refusal, a refusal from an appointment.

Concretely for example, if a ring no answer gets the code ‘10’, a new phone number, the code ‘20’, a refusal, the code ‘08’ and a completed interview, the code ‘01’, the new variable gives the code ‘1001’ to a completed interview from a never answered phone number, the code ‘2008’ to a refusal from a new phone number and the code ‘801’ to a completed interview from a previous refusal.

1We take for granted that, either the survey was conducted using a CATI system that automatically produces a file of call results including time and date of call, length of call, interviewer who performed the call and result of the call, as is the case for most available software.
Second step: produce the interviewers' daily work load file

The second step consists in producing the measures necessary to determine each interviewer's daily work load. In order to do this, it is sufficient to sort out the data set with the new variables by interviewer and then by date and time of call. This sorted file can then be aggregated by interviewer and day in order to give a new file which will contain the daily work load for each interviewer: total number of calls, number of completed interviews from never reached phone numbers, number of completed interviews from previous refusals, number of refusals from never reached phone numbers, etc. At the same time, it is possible to compute the daily shift length by subtracting the time of the first call from the time of the last call and adding the length of the last call. This file contains all the information necessary to characterize each interviewer's daily work load and performance.

Third step: produce the performance index

The final step consists in elaborating the performance index. In order to elaborate this index, inspiration came from sport. Baseball, hockey and other sports compute statistics based on the goals scored by a team or against a team when a given player is in the alignment. Similarly, in order to compute the performance index, one point is attributed for each questionnaire completed from a never reached phone number or from an appointment and one point is subtracted for each refusal from a never reached phone number. In addition, calls that were completed from previous refusals add two points to the index since they not only contribute one more completed to the pool but they also reduce refusals by one. The index proposed here can be seen as an index of net performance in the sense that it gives the net improvement in cooperation rate given by the work of an interviewer. This raw index needs then to be divided by the number of hours worked in order to be on the same basis for all interviewers.

Results: Properties of the performance index

Missing values

A first criteria of a good measure is its capacity to provide information whatever the task performed. In this regard, the performance index provides information for all the days and all the interviewers whereas the cooperation rate at first contact has almost 2 percent missing values (19 out of 977) days. The missing values occur mostly for firm number one (Jolicoeur) where a rolling cross-section type of sample was used. In this firm, for 18 of the 222 interviewer-days – which is 8 percent – the cooperation rate at first contact cannot be computed.

Variation

The quality of a measure may be assessed in different ways. The first one is whether it varies in itself, between individuals and with varying situations. In the present case, since it is possible to compute daily indices, it is also possible to assess whether the measure varies between
Interviewers and within interviewers between days. Multilevel analysis normally allows to figure out what part of the daily variation may be attributed to differences between interviewers as compared to differences from day to day within interviewers.

The daily index varies from a minimum of minus 10 to a maximum of plus 11, the mean daily index being equal to 0.26. However, when the multilevel structure of the data is taken into account, the mean is equal to 0.05. Table 1 provides information on the mean and share of variation between the two levels, namely the interviewer’s (level 2) and day’s (level 1), this for the 10 different files, the three firms and for the total. It shows that the mean index varies both between firms (.29, -.18 and .64) and within firms, particularly for firm number 2 (- .84, -78, +1.66). The variance is quite high, varying from .72 to 13.1. This indicates that the variance of the measure is generally appropriate. In the present case, a lower variance also indicates a greater homogeneity in the interviewers’ performance. The variance of the index is generally higher at Ekos and lower at Jolicoeur. However, it is related to the length of completed interviews so that it has to be interpreted with care. A number of other factors may also explain these differences. We will come back on this topic later on.

Multilevel analysis allows to examine the share of variance between levels, which is, what part of the variation is due to differences between interviewers and what part is attributed to day to day variation within interviewers. The intraclass correlation gives the proportion of variance attributable to interviewers. The lower the intraclass correlation, the higher the homogeneity in the pool of interviewers. In the present case, intraclass correlation varies from 0.31 at Jolicoeur to a range of .30 to .55 at Createc and .70 to .75 at Ekos. The variation is mostly between firms and not between projects of the same firm. This information, combined with the information on the variance, tends to confirm that the index performs well in giving an appreciation of the homogeneity of the pool of interviewers, this being characterized by a low intraclass correlation and a low variance, everything else being equal.

Comparison between the performance index and the cooperation rate

Figure 1. shows the plot of the relationship between the performance index and the cooperation rate at first contact. It shows that though the two indices are related, they are not similar.

The cooperation rate has the advantage and disadvantage of being bound i.e. it ranges from zero to 1, which makes it in a way comparable across projects, firms, etc. However, as stated before, as soon as an interviewer has completed whatever number of questionnaires from never reached households and has not produced new refusals, his/her cooperation rate is equal to one and if an interviewer has converted refusals during all his/her shift, the cooperation rate at first contact will be missing. By contrast, the performance index is bound by the number of interviews that can be completed in an hour.
Figure 2 presents the evolution of the mean index compared to the mean cooperation rate at first contact for one of the firm, Jolicoeur, where the two indices are most comparable (only one project, rolling cross-section, fieldwork every day for a whole month). It shows that the performance index is more variable than the cooperation rate at first contact. Is this a sign of quality of the index? Not necessarily, the two scales being quite different. Since a number of factors may influence the index for the 10 different projects, it is premature to conclude.

Possible bias of the index: length of questionnaire

One clear goal of the process is to end up with an index that is not biased by factors that should not be related to the measure. One such possible bias is the length of the questionnaire. The longer the questionnaire, the larger the difference in the time taken to complete a questionnaire versus get a refusal. This difference in the length of the questionnaire is likely to bias the index in the sense that, when you complete a long questionnaire, less time is available to call other phone numbers and therefore, there is less opportunity to get new refusals. However, this problem also affects the cooperation rate at first contact. Again, multilevel analysis was used to examine whether the mean length of each questionnaire was related to the daily performance index. However, this impact could be assessed only within firm since if the complete data set would be used, it would not be possible to differentiate the effect of firm from the effect of the mean length of questionnaire.

In table 2, the model 2 presents the results of a multilevel analysis of the impact of the mean length of the questionnaire on the index, this for the two firms where more than one project was conducted. It shows that, for one firm, Createc, there is no impact of the mean length of the questionnaire, whether using the index or the cooperation rate at first contact as dependent variable. However, such an impact exists for Ekos, the unexplained variance at level 1 (within days) decreasing by 32% (from 4.22 to 2.87) for the index and by 21% for the cooperation rate. This situation may be explained by the fact that Createc’s projects are of similar length of interview (317 to 442 sec.) and response rates while for Ekos, the third project differs highly from the two others by mean length of the questionnaire as well as by response rate (see table 1). Further verification allows to conclude that all the impact of length of interview for Ekos is explained by the third project. No difference exists between the two other projects. Therefore, the difference may be due as much to a higher response rate and possibly to a different way to manage field work as to the mean length of the questionnaire.
Results: relationship with external predictors

Relationship with the interviewer’s experience in the projects.

One available characteristic of the interviewers is the number of days they have worked on the pre-electoral surveys of each firm. For Jolicoeur, there was only one project and this information is rather reliable. For Createc, the variable gives information on the experience on all the projects. However, since the pre-electoral projects are quite similar in terms of demand and of type of questionnaire, the information is also quite reliable. In the case of Ekos, however, we do not have all the pre-electoral projects on which interviewers worked so that the information is less reliable. Therefore, only the data from Jolicoeur and Createc will be used. Finally, it has to be noted that this measure does not give information of the interviewers’ general experience, whether with previous projects, in other firms, etc.

The results show a positive impact of experience on the project for Jolicoeur (see table 3). This impact is significant and reduces the unexplained variance at the interviewer’s level by 17 percent. A similar analysis for Createc also gives a significant impact of interviewer experience and reduces the unexplained variance at the interviewer’s level by 7.8 percent.

By contrast, if the cooperation rate at first contact is used as the dependent variable (results not presented), there is no impact of experience on the rate for Jolicoeur and the impact is less substantial for Createc.

Relationship with the characteristics of the task

One question is whether the type of task that has to be performed by interviewers is related to the performance index. The different measures of the characteristics of the task are highly correlated so that it is necessary to make a choice among them. Some measures identify the task by the characteristics of the available phone numbers, i.e. whether they were never tried, never reached, were appointments or refusals. Other measures identify the results of the calls. Of particular importance is the proportion of ring no answers during a particular shift. Others are the proportion of appointments, refusals, invalid phone numbers. A screening process brought us to keep two measures: the proportion of “fresh” sample (never tried phone numbers) as input and the proportion of ring no answer as output.

Table 3 shows that for Jolicoeur, the introduction of the measures related to the task modifies the equation substantially. First of all, the task variables are related to the interviewer experience in such a way that the coefficient for the latter variable becomes non significant. In the new equation, the impact of the introduction of fresh sample as well as the proportion of ring no answer obtained is highly significant and negative. Finally, the contribution of these variables reduces the unexplained variance at the interviewer level by 21%. A similar situation appears for Createc. In this latter case, both the unexplained variance at the interviewer level (reduction of 10 percent) and at the day level (5 percent) are reduced.
Conclusion:

The performance index that has been elaborated has a number of good properties: first of all, it makes it possible to assess interviewers’ performance whatever the type of task performed. Second, within certain limits, it does not vary much with the length of the questionnaire. Third, it varies positively with the interviewer’s experience and is related to the task performed in a plausible way. However, this latter finding means that interviewers’ performance have to be assessed controlling for this impact. The index seems indeed to be an improvement over the previously used index.

A number of questions are not yet solved however. The actual index allows for comparison within firm and between projects of the same firm, provided that there is not much difference in the characteristics of the various projects. However, it is premature for now to conclude that the index is comparable across firms.

Finally, further analysis should bring us to assess whether some behaviors and attitudes hypothesized to impact on interviewers’ performance are indeed related to it.
### Table 1

**Basic information on the characteristics of the projects, firms and performance index**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Jolicoeur</th>
<th>Ekos total</th>
<th>Ekos 1</th>
<th>Ekos 2</th>
<th>Ekos 3</th>
<th>Createc total</th>
<th>Create 1</th>
<th>Create 2</th>
<th>Create 3</th>
<th>Create 5</th>
<th>Create 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix effect (mean)</td>
<td>.05</td>
<td>.29</td>
<td>-.18</td>
<td>-.839</td>
<td>-.78</td>
<td>1.66</td>
<td>.64</td>
<td>.61</td>
<td>.99</td>
<td>.71</td>
<td>1.0</td>
<td>.68</td>
</tr>
<tr>
<td>Random variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 2 (interviewer)</td>
<td>4.16</td>
<td>.23</td>
<td>5.2</td>
<td>3.88</td>
<td>5.73</td>
<td>9.82</td>
<td>1.54</td>
<td>1.64</td>
<td>1.26</td>
<td>1.15</td>
<td>1.05</td>
<td>1.85</td>
</tr>
<tr>
<td>Level 1 (date)</td>
<td>2.68</td>
<td>.49</td>
<td>4.2</td>
<td>1.61</td>
<td>2.29</td>
<td>3.32</td>
<td>1.96</td>
<td>1.80</td>
<td>1.01</td>
<td>1.50</td>
<td>2.47</td>
<td>2.73</td>
</tr>
<tr>
<td>Total variance</td>
<td>6.84</td>
<td>.72</td>
<td>9.4</td>
<td>5.49</td>
<td>8.02</td>
<td>13.14</td>
<td>3.5</td>
<td>3.44</td>
<td>2.27</td>
<td>2.65</td>
<td>3.52</td>
<td>4.58</td>
</tr>
<tr>
<td>Intraclass R</td>
<td>.61</td>
<td>.31</td>
<td>.55</td>
<td>.71</td>
<td>.71</td>
<td>.747</td>
<td>.44</td>
<td>.48</td>
<td>.55</td>
<td>.43</td>
<td>.298</td>
<td>.405</td>
</tr>
<tr>
<td>Length of questionnaire</td>
<td>-</td>
<td>1661</td>
<td>-</td>
<td>788</td>
<td>563</td>
<td>212</td>
<td>-</td>
<td>442</td>
<td>410</td>
<td>463</td>
<td>338</td>
<td>317</td>
</tr>
<tr>
<td>Cooperation rate</td>
<td>-</td>
<td>56.2</td>
<td>-</td>
<td>41.8</td>
<td>43</td>
<td>60.7</td>
<td>-</td>
<td>52.6</td>
<td>53.5</td>
<td>48.9</td>
<td>52</td>
<td>47.7</td>
</tr>
<tr>
<td>Response rate</td>
<td>-</td>
<td>50.5</td>
<td>-</td>
<td>41</td>
<td>42.6</td>
<td>60.4</td>
<td>-</td>
<td>35.6</td>
<td>35.7</td>
<td>34.8</td>
<td>36.6</td>
<td>35.2</td>
</tr>
<tr>
<td>N questionnaires</td>
<td>10756</td>
<td>1252</td>
<td>7725</td>
<td>2001</td>
<td>2237</td>
<td>3487</td>
<td>1779</td>
<td>411</td>
<td>343</td>
<td>365</td>
<td>314</td>
<td>334</td>
</tr>
<tr>
<td>completed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nb interviewers</td>
<td>190</td>
<td>26</td>
<td>124</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nb days</td>
<td>977</td>
<td>222</td>
<td>500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>255</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 2
Impact of the mean length of the questionnaire on indices of performance

<table>
<thead>
<tr>
<th></th>
<th>Createc</th>
<th></th>
<th></th>
<th></th>
<th>Ekos</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Performance index</td>
<td>Cooperation rate 1st contact</td>
<td>Performance index</td>
<td>Cooperation rate 1st contact</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>Model 1</td>
<td>.64**</td>
<td>Model 1</td>
<td>.52***</td>
<td>Model 1</td>
<td>-.19</td>
<td>Model 1</td>
<td>-.005***</td>
</tr>
<tr>
<td></td>
<td>Model 2</td>
<td>1.49*</td>
<td>Model 2</td>
<td>.57***</td>
<td>Model 2</td>
<td>2.52***</td>
<td>Model 2</td>
<td>.68***</td>
</tr>
<tr>
<td>Length of questionnaire</td>
<td></td>
<td>-.002</td>
<td></td>
<td>-.0001</td>
<td></td>
<td>-.0003***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random variance</td>
<td>Interviewer level</td>
<td>1.54***</td>
<td>1.57</td>
<td>.014***</td>
<td>.014***</td>
<td>5.48***</td>
<td>5.95***</td>
<td>.033***</td>
</tr>
<tr>
<td></td>
<td>Day level</td>
<td>1.96***</td>
<td>1.94</td>
<td>.016***</td>
<td>.016***</td>
<td>4.22***</td>
<td>2.87***</td>
<td>.019***</td>
</tr>
</tbody>
</table>

*: p < .05; **: p < .01; ***: p < .001
### Table 3
Impact of external factors on the performance index

<table>
<thead>
<tr>
<th></th>
<th>Jolicoeur</th>
<th>Createc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>.29**</td>
<td>.12</td>
</tr>
<tr>
<td><strong>Experience - days in project (s)</strong></td>
<td>.03**</td>
<td>.01</td>
</tr>
<tr>
<td><strong>Proportion of fresh sample</strong></td>
<td>- .93***</td>
<td></td>
</tr>
<tr>
<td><strong>Proportion of ring no answer obtained</strong></td>
<td>-1.24***</td>
<td></td>
</tr>
<tr>
<td><strong>Random variance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interviewer’s level</strong></td>
<td>.23**</td>
<td>.19*</td>
</tr>
<tr>
<td><strong>Day’s level</strong></td>
<td>.49***</td>
<td>.47***</td>
</tr>
</tbody>
</table>

*: p < .05; **: p < .01; ***: p < .001
Figure 1. Relationship between the performance index and the cooperation rate at first contact - all projects.
Figure 2 : Evolution of the performance index and the cooperation rate at first contact during the project - Jolicoeur