Proposed Indicators to Assess Interviewer Performance in CATI Surveys

François Laflamme[1], Hélène St-Jean[2]
[1]Statistics Canada, francois.laflamme@statcan.gc.ca
[2]Statistics Canada, helene.st-jean@statcan.gc.ca

Abstract

Interviewer Performance, defined in this paper as the ability of an interviewer to contact and convince respondents, is generally assessed by survey research call centres in using descriptive measures such as the number of completed interviews, -the number of completed interviews per hour, etc. Other more comprehensive performance indicators such as the cooperation rate at first contact and Net Contribution to Performance Index have been developed over the past few years. However many factors might impact interviewers' performance in a centralized call centre environment. In addition to the interviewer's characteristics and environmental factors, the type and portfolio of cases called, the effort already put into these cases, the time the call is made and the general productivity of the survey at the moment at which the call is made are some of these potential influencing factors. This paper proposes a new objective interviewer's performance measure that takes into account the complexity of the survey data collection process as well as new factors to consider when assessing refusal conversion and tracing data collection activities.

1. Introduction

With extensive use of computer-assisted data collection methods and the increasing complexity of survey data collection procedures, it becomes more and more challenging to assess the performance of telephone interviewers. For example, the automated call scheduler (which brings up cases automatically to the next available interviewer) considers many parameters before allocating cases to interviewers. This impacts the individual performance measurements of the interviewer.

How can interviewer performance or productivity be measured and monitored for telephone surveys? First, it is worth explaining what 'Interviewer Performance' means. This concept can be defined in two different but complementary ways:

- By the ability to contact and convince potential respondents to complete a survey questionnaire (Tarnai and Moore, 2007);
- By the ability to conduct quality interviews (e.g. CATI monitoring program that evaluates the interviewer's skills and PoINT system that assesses the pace of interviews (Egan 2009)).

Interviewer Performance (IP) is defined in this paper as the ability of an interviewer to contact and convince respondents to complete a survey questionnaire. Basic performance indicators include: number of calls made, number of completed interviews, number of refusals, average length of interview, total number of minutes worked and so on. Other time-based measures can be derived, for example, calls made per hour, completed

interviews per hour, refusals per hour, etc. Finally, other more comprehensive indicators such as the cooperation rate at the first contact (Cooprt1), defined as the number of completed interviews divided by the sum of completed interviews and refusals at the first contact is an example (Durand, 2005). Another example is the Net Contribution to Performance Index (NCPI) which results from the characterization and the scoring of each type of call (Durand, 2008).

Many external and internal factors might impact interviewers' performance in a centralized call centre environment. In addition to the interviewer's experience and profile, the type and portfolio of cases called, the effort already put into these cases, the time the call is made and the survey productivity at the moment of the call are some of these factors. The survey productivity is defined as the ratio of the system time devoted to the interviews themselves to the total system time which includes all unsuccessful and successful calls (Laflamme 2009). Total system time represents the total time logged onto the system once a case is open. All these factors need to be considered in any fair and objective IP measure to take into account the complexity of the data collection process of a centralized call centre.

The main objective of this paper is to present a new interviewer's performance measure that takes into account most of the complexity of survey data collection process. The secondary objective is to present a way to assess the success of special and difficult operational interviewers' tasks such as refusal conversion and tracing activities. The paper begins with an overview of the data collection process and procedures used in the centralized call centre environment of Statistics Canada. The next section describes the influencing factors that impact interviewer performance while section 4 presents the expected characteristics of the new interviewer performance indicators. Section 5 describes in details the new IP indicator and its characteristics and particularities including other derived measures that can be used to enhance the analytical value of the proposed IP indicator. Section 6 presents other measures to evaluate special collection activities such as the refusal conversion and tracing data collection activities for each interviewer. The research presented in this paper constitutes only the first phase in the development of a new IP indicator. The proposed measure is a prototype that will require more studies and analyses before implementing in production. Future research and ideas conclude this paper.

2. Data Collection for CATI surveys at Statistics Canada

Data collection for Computer Assisted Telephone Interviewing (CATI) social surveys is conducted and managed in Statistics Canada's five Regional Office (RO) call centres located across the country. All survey applications are built using the Blaise software and the call scheduler automatically[1] assigns individual cases to interviewers working out of a centralized environment. The call scheduler takes into account the interviewers' profile, paradata information collected since the beginning of the data collection period (e.g. outcomes of the previous calls) and some data collection parameters to assign a case to an interviewer.

[1] Interviewers have the opportunity to use a browser tool to access any in-progress case. This means that the interviewer can scroll the list of all cases and manually select a case thus skipping the call scheduler.

2.1 Interviewer profile

An interviewer profile is based on interviewer's characteristics (e.g. age and sex), skills and experience. It is an important component of the call scheduler. During the data collection period, a given interviewer can be identified to receive in priority (or exclusively) cases that belong to one primary Blaise group and one or more secondary Blaise groups. For example, experienced interviewers (or interviewers with very good convincing skills) are assigned to the Blaise Refusal group in order to try to convert those cases for which at least one refusal was recorded.

The assignment of specific interviewers to the Blaise Tracing[2] group is also another very good example.

2.2 Blaise Transaction History (BTH) record

A BTH record is automatically created each time a case is closed, whether it was opened for data collection or other purposes. The BTH record contains detailed information about each call made to contact each sampled unit during the data collection period. It also includes information on the survey and case identification, the date, the amount of time the case was open, the interviewer who worked on it, the resulting Blaise group (e.g. Refusal, Tracing, Regular, Home (finalized)), appointment information, the result of the call plus additional relevant information. The call scheduler considers, for example, the number and time of calls that have been made to an individual case, the result and the Blaise group of the last call to assign cases to a given interviewer. These rules essentially refer to the 'routing table' of the survey application.

2.3 Collection parameters

In addition to the routing table, other collection parameters are considered by the call scheduler such as time slices, cap on calls, appointments and other technical parameters. The time slice feature in the CATI call scheduler was utilized to assist in managing the new cap on calls policy which limits the number of calls[3] that can be made for each case. In practice, time slices ensure that a specific number of calls are attempted at different times of the day, and on different days of the week, before a case is finalized. It should be noted that only cases with a "no answer" outcome for the last call are subject to be influenced by the time slice parameters. The call scheduler also needs to manage appointments to make sure that cases are assigned at the appropriate time to an interviewer. In addition, some other technical parameters are considered by the call scheduler, for example, time between "busy" calls, minimum time between other "no answer" etc.

3. Impacting factors on interviewer performance

Several external and internal factors to the data collection process might impact interviewers' performance in a centralized CATI environment when cases are automatically delivered to interviewers according to a set rules used by the call scheduler. Among the external factors, the interviewer's experience, training, profile and skills, survey questionnaire, topic of the survey and quality of the survey frame (in particular, the contact information) represent some of the most important ones. Many of these

[2] Tracing consists of strategic and logical searches using all available resources to locate a respondent e.g. for those where the frame provided a wrong or missing telephone number.

[3] The cap on calls varies from 20 to 40 calls depending on the survey.

external factors are indirectly considered in the new performance indicator. In addition to these factors, the purpose of the calls for tracing or for any other reason, the type of cases called, the effort already put into cases, the time at which the calls were made and survey productivity at the moment of the call are among the most important internal factors that are used in the development of the proposed IP indicator. In fact, it is recognized that survey productivity decreases through time as the collection period progresses for CATI surveys (Laflamme 2009) since the proportion of 'difficult cases' increases. The proposed IP indicator takes advantage of the productivity variability by weighting each interview by a function of the inverse of its likelihood when considering the main internal factors. In practice, interviews that were less likely to be completed according to the realized low likelihood (based on the main factors) received a larger weight compared to those completed in a high likelihood environment. A detailed description of the proposed indicator is provided in section 5.

4. Expected characteristics of the new interviewer performance

The first objective of this research is to propose a new interviewer's performance measure that takes into account the most important factors impacting response and the complexity of the survey data collection process and procedures. To be operationally useful, this IP indicator should demonstrate some expected characteristics, in particular, it should:

- be objective and fair to all interviewers;
- be easily understood and interpreted;
- be measurable at any point in time during collection or for any duration;
- be adaptable to any type of CATI surveys;
- be fully automated and reproducible; and
- be decomposable into sub-components.

While the first five characteristics are self-explanatory, the last one needs some clarification. Even though the objective is to develop a single IP measure for each interviewer, it is also desirable to be able to determine the relative contribution of each component of the IP indicator to get a better appreciation of the operational context attached to the interviewer performance. The second objective of this research is to develop a tool to assess special operational interviewers' tasks such as refusal conversion and tracing activities. The intrinsic nature of the interviewer's work assigned to these two kinds of tasks is different and needs to be addressed separately.

5. Proposed IP indicator

Assessing interviewer performance in a centralized and complex environment is not straightforward and represents a real challenge with regards to developing a fair and objective interviewer productivity measure. A basic indicator can be simply defined as the ratio of interview system time (i.e. system time devoted to the interviews themselves) to the total system time for each interview. This indicator refers to survey productivity calculated at the interviewer level. However, this simple measure, called Raw IP in Section 5.4, does not take into account the dynamic and complexity of data collection process. As mentioned before, interviewer productivity for telephone surveys depends on many compounding factors. In particular, the purpose of the call, the characteristics of the case, the amount of work already performed on the case, its contact success to date and

the time the call was made are considered in the development of this proposed IP indicator.

5.1 In-scope calls

The purpose of each call is used to identify the in-scope calls used in the IP calculation. All calls are included except those made for some specific surveys that require tracing effort. In other words, the in-scope calls for the IP indicator are all calls except those made for tracing. By its nature, a tracing call or task is different than the other types of data collection activities. Once a case is traced successfully by a specialized interviewer, the case is generally redirected to a regular Blaise group to be called by another interviewer. As soon as a case is moved out of the Blaise tracing group, it is considered in-scope for the IP calculation. It should also be noted that some cases may remain in the Blaise tracing group for the entire data collection period meaning that no direct collection activities are ever performed on these cases.

5.2 Categorization of the impacting factors

CATI interviewer productivity depends on the characteristics of the case delivered and its contact success to date (factor 1, state of the case), the amount of work already spent on the case (factor 2, number of calls previously made) and the time the call was made (factor 3). In order to be easily used and understood, the factors need to be categorized in a meaningful and sound way.

The first factor essentially refers to the state of the case when delivered to the interviewer. The state of cases called by an interviewer can be determined by the sequence of calls previously made for each case. Currently, six states are used to differentiate cases at any point in time through collection:

1. Cases called for the first time (first call);
2. Cases with no previous contact and no tracing attempt;
3. Cases with no previous contact but that required some tracing;
4. Contacted cases for which no previous contact resulted in a refusal;
5. Contacted cases for which at least one previous contact resulted in a refusal;
6. Cases that transitioned at least once to the Senior Interviewer (SI) or Project Manager (PM) Blaise groups[4]. Once in, cases remain in this category.

A given case can go through many states during the course of data collection. For example, a case can go through states 1, 3 and 5 for the first three calls. However, some sequences of states are not possible e.g. 1, 6, 3 because a case remains in state 6 once it is in. Another impossible example is 1, 5, 3 because once a contact is made it is impossible for a case to go back to a 'no contact' state (i.e. 2 and 3). The above choice of states was based on practical and empirical reasons, resulting in a few exhaustive and mutually exclusive states. The productivity (i.e. ratio of the system time used to complete interviews to the total system time) attached to each state for a given survey varies (see Table 1).

The second factor, the amount of work already performed on a case, can be expressed by the number of calls previously made to the case before a certain point in time during collection. In the current research project, eight intervals of the number of calls are

[4] In special circumstances cases can be redirected to this group for in-depth review.

defined: “0” (first call), “1 or 2”, “3 to 5”, “6 to 10”, “11 to 15”, “16 to 20”, “21 to 25”, “26 calls and more”. The categories may vary from one survey to the next to avoid having intervals with a small number of cases. In the current example[5], a cap on calls of 40 was used. The survey productivity for the 2010 Survey of Labour and Income Dynamics (SLID) varies from a high of 54% for the first call category (see Table 2) to a low of 18% for the last interval.

The third factor is the time at which the call was made. All calls initiated before 16:00 belong to the day shift and all the others belong to the evening shift. Many past researches have demonstrated that evening shifts are generally more productive than day shifts for CATI social surveys, especially at the beginning of collection period. As shown in Table 3, the difference between the survey productivity of the two shifts for SLID 2010 was 4.3 percentage points, the evening shift outperforming the day shift.

Table 1

State of case	Productivity
First call	54.2%
No contact cases and no tracing	37.8%
No contact cases and some tracing	29.6%
Contacted cases with any previous refusals	41.4%
Contacted cases with at least one previous refusals	24.7%
Cases previously directed to SI or PM Blaise groups	29.0%

Table 2

Number of calls previously made	Productivity
0 call (first call)	54.2%
1 to 2 calls	52.7%
3 to 5 calls	44.3%
6 to10 calls	35.0%
11 to 15 calls	27.4%
16 to 20 calls	23.5%
21 to 25 calls	21.4%
26 calls and more	18.4%

Table 3

Time of the call	Productivity
Day (before 16:00)	37.5%
Evening (after 16:00)	41.8%

Putting the first call aside, there are 70 possible combinations (i.e. 5 x 7 x 2) of the three factors. As for the first call, only two groupings remain possible, the first call for day and evening shifts. Hence the combination of these three categorized factors results in a maximum of 72 groups or combinations. It should be noted that the portfolio of the calls made by an interviewer refers to the proportion of the total number of calls made for each factor (see table 4). It is used as the basic conceptual framework for the proposed IP indicator. In practice, each call, at a certain point in time during collection, is defined by one of these combinations. It should be noted that the day of data collection was also considered as one impacting factors but was discarded for two main reasons. Firstly, it would have created many combinations with a small number of calls, and potential outliers. Secondly, the fact that survey productivity is used as a weighting factor indirectly includes the survey day since productivity decreases as survey progresses.

5.3 Development of the proposed IP indicator

The proposed weighted IP indicator uses survey productivity attached to each combination of the three factors which defines a ‘type of case’ at different point in time during collection. Throughout collection, a given case can be included in different combinations depending on the characteristics of the calls made for the case. Each call resulting in a completed or partial interview is weighted by the square root of the inverse

[5] This research project used the Survey of Labour and Income Dynamics (SLID) data collected during the 2010 period. SLID is an annual longitudinal survey of about 34,000 sample units and one of the most complex CATI surveys at Statistics Canada.

of its likelihood of being interviewed (function of the productivity) then multiplied by the duration of the interview itself. In order words, the weight *w* of the combination is equal to the square root[6] of (1/productivity of the combination) while the weight of the call is *w**duration of the interview. All other calls received a weight of one[7]. More precisely, IP indicator for a given interviewer is derived in the following way:

i = Interviewer identifier
j = j^{th} combination of the three factors, $j = 1,\ldots,72$
k = Call identifier, $k = 1,\ldots, K$. *K is* the total number of in-scope calls made for the survey[8].
t_{jk} = System time of k^{th} call in combination j
d_{jk} = 1 if k^{th} call is in combination j, 0 otherwise
d_{jkc} = 1 if k^{th} call is in combination j and is an interview, 0 otherwise
d_{ijk} = 1 if k^{th} call is made by interviewer i and in combination j, 0 otherwise
d_{ijkc} = 1 if k^{th} call is made by interviewer i, is in combination j, and is an interview, 0 otherwise

Let $P_j = \dfrac{\sum_k t_{jk} d_{jkc}}{\sum_k t_{jk} d_{jk}}$ be the productivity of j^{th} combination with Max(P_j)=0.7 and Min(P_j)=0.1 set as fixed thresholds for the survey under study, and $w_j = \sqrt{\dfrac{1}{P_j}}$ be the weight of j^{th} combination. The weight of the k^{th} call in the j^{th} combination is w_{jk}. It equals w_j for an interview and 1 otherwise. The Weighted Interviewer Performance for interviewer i is defined as

$$WIP_i = \frac{\sum_j \sum_k w_{jk} t_{jk} d_{ijkc}}{\sum_j \sum_k w_{jk} t_{jk} d_{ijk}}$$

For example, completed cases from a combination with high productivity, hence more likely to be completed, receive a smaller weight than those that belong to low productivity combinations. Since the productivity decreases over the course of data collection, cases that are completed at the beginning of the survey are more likely to receive a smaller weight than those completed close to the end of collection period when the productivity is lower.

Given the proposed strategy, the productivity attached to any combination of factors will vary as data collection progresses until no more calls can be added to that combination or until the end of collection. At the beginning of the survey, some combinations (e.g. those requiring a large number of calls) are going to be empty. They will start to be filled in with few cases at some point in time during collection and will keep growing from day to day until a certain point in time (i.e. when it's impossible to add more cases) or until the

[6] Various functions were investigated before selecting the square root.
[7] It is possible to investigate another strategy where some 'weight' is given to other types of calls (e.g. appointment which can be seen as a partial success) but this will compromise the comparability with the raw survey productivity measure. It can also introduce more subjective information.
[8] K=332,364 in-scope calls for SLID 2010.

end of collection. For example, the combinations based on the 'first call' will be saturated once all sampled cases are called at least once (i.e. productivity and weight are fixed after that). Until that point in time, productivity of each combination varies. It also means that the weight attached to a given completed case can also vary from day to day based on the productivity of the combination. However, it is always possible to reproduce the overall collection process (i.e. the categorization of each call and the calculation of the weights of each combination) at the end of every data collection day.

When cases start to populate combinations, the productivity can be based on a very small number of records (calls) with the potential of creating some outliers i.e. a very high or very low productivity attached to the combination. To avoid the problem, minimum and maximum productivity boundary thresholds of 0.1 and 0.7 were respectively set in the prototype model (thresholds can vary by survey). In other words, the minimum and maximum weights that can be used for a given interview and combination are respectively 1.12 and 3.16[9]. It should be noted that the WIP indicator can be decomposed into its sub-components (up to 72) indicating on which 'type of cases' (combination) an interviewer has mostly worked during the survey and on which 'type of cases' an interviewer performed best (i.e. the biggest contributors to the WIP).

Finally, any WIP measure needs to be analyzed in its environmental context. A lower WIP does not necessarily mean that the interviewer is not as good as others. It can simply mean that the interviewer has worked on more difficult cases and this information needs to be provided and considered by the RO survey manager in the evaluation.

5.4 Derived IP measures

In addition to the proposed IP indicators, two other useful measures can be easily derived: Raw IP indicator and the standardized weight for complete and partial interview measure.

As opposed to the *weighted IP indicator*, the *raw IP indicator* is simply defined as the survey productivity indicator i.e. the ratio of interview system time (i.e. system time devoted to the interviews themselves) to the total system time. The weight w_j of each call is one in the formula of section 5.3. In other words, the raw indicator does not use the classification and the productivity attached to each combination *j*.

The *standardized weight for complete and partial* measure is defined as w_j for completed and partial interviews made by interviewer i ($w_j > 1$) divided by the mean w_j for completed and partial interviews made by all the interviewers in the same RO. If the *average standardized weight for complete and partial measure* for a given interviewer is higher than one, it means that on average the interviewer has completed more 'difficult' cases than the other interviewers in the RO.

5.5 IP Interpretation

Tables 4, 5 and 6 provide the results for a subset of four interviewers for SLID 2010 within the same RO. As shown in Table 6, Interv_1 has respectively a Raw and Weighted IP of 21% and 32.4% (+11.4 percentage points) and an average of standardized weights for complete and partial interviews greater than 1 (1.1750) meaning than this interviewer has completed on average more 'difficult' cases (i.e. cases in combinations with lower

[9] Less that 2% of the calls were subjected to these thresholds at the end of the collection period.

productivity) than the other interviewers in the RO (see Table 5). Even though, the Raw IP for Interv_1 is much lower than the overall RO productivity (40.7%), the Weighted IP recognized the fact that the interviewer was able to complete more difficult cases than others. Table 4 shows that this interviewer also worked on a portfolio of more difficult cases. For example, Interv_1 did not start to work at the beginning of the survey on the easier cases. Also this interviewer only worked on the day shift where survey productivity is lower.

Table 4: Proportion of the total number of calls made by factor

	Portfolio of calls made (Proportion of the total number of calls made)															
	States of the cases						Number of calls previously made on cases								Time of the calls	
Interv ID	First call	No contact & no tracing	No contact & tracing	Contact - no refusal call	Contact - at least one refusal call	From SI & PM groups	0	1-2	3-5	6-10	11-15	16-20	21-25	26+	Day 9:00-16:00	Evening 16:00-23:00
Interv_1	0.1%	22.8%	12.4%	59.2%	4.3%	1.2%	0.1%	3.3%	9.3%	31.1%	15.2%	14.2%	11.8%	15.0%	100.0%	0.0%
Interv_2	22.2%	33.3%	3.2%	41.4%	0.0%	0.0%	22.2%	21.0%	10.6%	26.8%	10.9%	8.5%	0.2%	0.0%	68.5%	31.5%
Interv_3	7.2%	31.4%	5.2%	55.1%	0.5%	0.5%	7.2%	10.8%	23.8%	13.0%	14.7%	9.9%	12.2%	8.5%	23.1%	76.9%
Interv_4	6.9%	18.4%	4.2%	41.6%	27.7%	1.3%	6.9%	14.1%	11.1%	21.1%	11.4%	11.3%	8.5%	15.7%	54.2%	45.8%

Table 5: Proportion of the total number of interviews by factor

	Distribution of interview made (Proportion of the total number of interviews)															
	States of the cases						Number of calls previously made on cases								Time of the calls	
Interv ID	First call	No contact & no tracing	No contact & tracing	Contact - no refusal call	Contact - at least one refusal call	From SI & PM groups	0	1-2	3-5	6-10	11-15	16-20	21-25	26+	Day 9:00-16:00	Evening 16:00-23:00
Interv_1	0.0%	17.9%	10.7%	57.1%	10.7%	3.6%	0.0%	3.6%	28.6%	42.9%	21.4%	0.0%	3.6%	0.0%	100.0%	0.0%
Interv_2	39.0%	22.1%	1.3%	37.7%	0.0%	0.0%	39.0%	42.9%	11.7%	3.9%	1.3%	1.3%	0.0%	0.0%	36.4%	63.6%
Interv_3	17.6%	24.7%	1.2%	54.1%	0.0%	2.4%	17.6%	20.0%	36.5%	9.4%	5.9%	7.1%	2.4%	1.2%	25.9%	74.1%
Interv_4	21.4%	16.2%	0.6%	41.6%	16.2%	3.9%	21.4%	23.4%	13.6%	20.1%	9.1%	2.6%	2.6%	7.1%	48.1%	51.9%

Table 6: Interview Performance (IP) indicator and other complementary statistics

		Number of Interviews			System Time			IP Indicators		
Interv ID	Calls made	Complete	Partial	Total	Interview (hrs)	Total (hrs)	Ratio	Raw IP	Weighted IP (WIP)	Average of standardized weights
Interv_1	927	27	1	28	12.7	60.6	21.0%	21.0%	32.4%	1.1750
Interv_2	568	76	1	77	22.3	45.2	49.4%	49.4%	57.9%	0.9261
Interv_3	1,274	81	4	85	28.2	76.8	36.8%	36.8%	46.9%	1.0027
Interv_4	1,634	141	13	154	58.8	160.5	36.6%	36.6%	48.1%	1.0708
RO	53,133	4,472	884	5,356	1,877	4,609	40.7%	40.7%	51.3%	1.0000

On the other hand Interv_2 has a Raw IP (49.4%) which is higher than RO survey productivity (40.7%). However, the workload composition of this interviewer was much easier than the first one which is reflected in the average of the standardized weights (lower than one). Finally, Interv_3 and Interv_4 have a similar Raw IP. However, the Weighted IP for Interv_4 is higher given his workload composition which contains more difficult cases (e.g. cases with at least one refusal) and the fact that he was able to complete cases in more difficult combinations of factors.

6. Other measures for special data collection activities

This section proposes various measures to evaluate the refusal conversion and tracing collection activities as well as other complementary descriptive indicators that can be

used in conjunction to the proposed IP indicators to assess the interviewer performance in providing additional contextual information.

6.1 Refusal

In order to assess more specifically interviewer performance in terms of refusal conversion rate, a series of complementary descriptive measures were derived. The number of cases called and the total number of calls made for refusal conversion (i.e. all calls following the first refusal), the overall conversion refusal rate, the conversion rate at first contact after the first refusal and the conversion rate at second contact after the first refusal are among the most important ones. It should be noted that Statistics Canada has a refusal conversion policy that allows a maximum of two refusal conversion calls.

6.2 Tracing

Tracing is a different data collection task for some CATI surveys which is not covered by the proposed IP indicators for many reasons. In particular, the work of the tracing activity by a given interviewer generally does not result in an interview. A 'traced' case is generally sent to another Blaise group for an interview later on by another interviewer and the tracing interviewer is not credited for this tracing work. In general, it is not easy to evaluate objectively and fairly this particular data collection activity. The in-scope calls for this series of measures are all tracing calls (i.e. those excluded in the IP calculation). The following descriptive measures are among the most important ones that can be used to evaluate the tracing performance of each tracing interviewer: the number of cases called and the total number of calls made for tracing, the number of cases successfully traced[10], the tracing rate, the number of cases successfully traced the first time, the number of cases that were thought to be successfully traced by the tracing interviewer the first time but sent back to tracing for a second time, the number of tracing cases worked on by the tracing interviewer resulting in an interview and the conversion tracing rate which is the ratio of these cases over the total number of tracing cases.

6.3 Other general measures

Other descriptive measures can be used in conjunction with the proposed IP to provide complementary contextual information to help interpret the IP measure. The average interview length, the average lag of time between consecutive calls, the proportion of browser calls, the refusal rate at the first contact, the number of calls, complete interviews and refusals are among the other descriptive measures that can be used to give supplemental information about interviewers' performance.

7. Summary

The proposed IP is a prototype model that aims at providing a comprehensive measure of the interviewer's performance in the complex data collection environment of Statistics Canada. The indicator is objective and seems to be fair at first glance to all interviewers. The WIP measure also meets most of the other expected characteristics. However, the indicator is not perfect. The WIP is somewhat dependant of the call scheduler parameters, interview length and staffing levels. For example, an interviewer with a higher average interview length is likely to obtain a higher WIP compared to another interviewer under similar conditions. The productivity boundary thresholds (which are somewhat arbitrary

[10] A case is considered successfully traced when it is redirected from the tracing group to another group.

at this point) used to take into account potential outliers might impact the WIP measure at the interviewer level. The fact that the productivity attached to any combination of factors can vary (and consequently the weight of the combination) as data collection progress, might also be seen as a disadvantage of this dynamic measure because it can introduce some variability in day to day comparison. Performance measure needs to be analyzed in its environmental context by survey manager in the evaluation of each interviewer.

8. Future work

This paper constitutes only the first step in the development of new IP indicators. More research is required. For example, it can be very useful to develop an analytical tool for survey managers that breaks the WIP into subcomponents to identify efficiently the most important contributors (combination of factors) to the indicator. It is also possible to identify combinations where the interviewer did not perform well and might require additional training. Investigations should be made towards the creation of another tool showing the evolution of the WIP indicator through intervals of time (e.g. weekly). To that extent, it will be essential to study the variability of the weights of the combinations during data collection. The feasibility of using such a tool for regular interviewer feedback also needs to be investigated. Some other research can focus on strategies to group interviewers with similar workload or portfolio to facilitate comparisons. More investigations are also required to compare this measure to other measures such as cooperation rate at first contact (Cooprt1) and Net Contribution to Performance Index (NCPI). In addition, the same type of analysis should be conducted across different types of surveys to assess its robustness. Given that interviewers can work on many CATI surveys during a period of time, it would also be interesting to evaluate the possibility of combining IP measures for two or more surveys. Finally, survey managers in Regional Offices need to be consulted with respect to the practicality and operationalization of such new performance indicator. In the long run, it would be more than desirable to integrate the various measures that assess the two dimensions of the interviewer performance (i.e. ability to contact/convince respondents and ability to conduct quality interviews (CATI monitoring and PoINT system)) into one consolidated tool.

References

Durand, C. (2005), “Measuring Interviewer Performance in Telephone Surveys”, Quality & Quantity Journal, Springer Netherlands, Vol. 39, No. 6

Durand, C. (2008), “Assessing the usefulness of a new measure of interviewer performance in telephone surveys”, Public Opinion Quarterly, Vol. 72, No. 4, Winter 2008, pp. 741–752

Egan, M. (2009), “POINT - A Method for Evaluating the Quality of Interviewing”. Internal report, Statistics Canada.

Laflamme, F. (2009), “Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada”. Proceedings from the 57th International Statistical Institute Conference.

Tarnai, J. and Moore, D.L. (2007), “Measuring and improving telephone interviewer performance and productivity”, Advances in telephone survey methodology, Wiley series, Chapter 17.