DATA QUALITY MONITORING USING THE BLAISE AUDIT TRAIL

Jennifer Ali[1]

ABSTRACT

Statistics Canada's Canadian Community Health Survey has expanded its approach to data quality monitoring and developed an innovative program for the ongoing evaluation of the integrity of data during data collection. This program adds to existing methods of data quality monitoring and allows for prompt interventions during data collection to improve data quality. The program uses the Blaise audit trail which yields a richly detailed file of times and trailing within the questionnaire. It records the duration and content of each keystroke at the case level. Two primary types of indicators are duration timings and non-response. For the CCHS 1.2, analysis was conducted at the national, regional, interviewer and case levels. This paper describes the process, identifies appropriate applications, discusses statistics that can be generated, and outlines actions taken by Statistics Canada to improve data quality. Limitations and challenges in the implementation of a data quality monitoring program based on the Blaise audit trail are discussed.

KEY WORDS: Audit trail, Blaise, Canadian Community Health Survey, Data quality, Monitoring, Survey methods

RÉSUMÉ

L'Enquête sur la santé dans les collectivités canadiennes de Statistique Canada (ESCC) a étendu son approche de la surveillance de la qualité des données et mis en place un programme novateur d'évaluation permanente de l'intégrité des données durant la collecte. Ce programme étoffe les méthodes existantes de surveillance de la qualité des données et permet d'intervenir promptement durant la collecte des données afin d'améliorer la qualité de ces dernières. Il s'appuie sur l'option Piste de vérification de Blaise qui fournit un fichier de données très détaillées sur le minutage et le cheminement des interventions dans le questionnaire. Il permet d'enregistrer la durée et le contenu de chaque entrée dans un champs du questionnaire au niveau du cas. Les deux principaux types d'indicateurs sont ceux de durée d'interview et de non-réponse. Pour le cycle 1.2 de l'ESCC, l'analyse a été réalisée aux niveaux national et régional, ainsi qu'aux niveaux de l'intervieweur et du cas. L'article décrit le processus, les applications appropriées, les statistiques qui peuvent être produites, ainsi que les mesures prises par Statistique Canada pour améliorer la qualité des données. Sont également décrits les limites et les défis de la mise en œuvre d'un programme de surveillance de la qualité des données fondé sur l'option Piste de vérification de Blaise.

MOTS CLÉS : Qualité des données; Blaise; Enquête sur la santé dans les collectivités canadiennes; méthodes d'enquête; piste de vérification; surveillance.

1. INTRODUCTION

Monitoring the quality of the data collected in a survey is an essential component of any survey process. At Statistics Canada, data quality monitoring is an activity that occurs throughout the survey process. Pre-testing and pilot tests address data quality before data collection. During data collection, field personnel monitor data as they oversee the daily activities of project management such as monitoring response rates, reasons for non-response, number of trials for phoning and examining statistics on timers built into the collection instrument. During data processing, data quality is monitored through consistency checks, calculation and comparisons of response rates, and analysis of sampling and non-sampling errors.

Monitoring during data collection is essential to ensure that any deviations from procedure or issues presenting during collection can be resolved in a timely manner to avoid impacting the quality of the data collected. This is a challenging phase of the survey to monitor because of the fast pace of data collection and the time-consuming, labour-intensive, and decentralized nature of standard monitoring procedures. Moreover, although timings can be monitored at the module level though timers built into the questionnaire at the beginning and end of each module, these timers do not provide micro level information at the question level and cannot take into account back and forth movement between modules in the questionnaire. Therefore, while providing valuable information, the detail and scope of monitoring using standard approaches is restricted. New technology in the form of the Blaise audit trail allows the expansion of monitoring to the

1 Jennifer Ali(Jennifer.ali@statcan.ca), Health Statistics Division, Statistics Canada, Ottawa, Ontario, Canada, K1A OT6

micro level of data during this critical stage of data collection. Specifically, it provides data for case by case analysis of timings and edits at the question level.

The Canadian Community Health Survey, cycle 1.2 (CCHS 1.2) presented an opportunity to implement this expansion. The survey was a computer-assisted face-to-face interview of about 37,000 Canadians. Data was collected between May 2002 and December 2002. The structure of the questionnaire instrument and the subject matter raised some challenges that motivated more extensive monitoring while the data was in the field. This was the first national survey on mental health in Canada to assess prevalence rates for five mental disorders (major depression, mania, agoraphobia, social phobia, panic disorder). There was concern about possible respondent reactions to the personal questions. Feedback during data collection about non-response and edits could help monitor that. Second, many questions were quite long and contained multiple concepts. Monitoring at the question level could help ensure that questions were being asked consistently and slowly and that the multiple concepts in the questions were conveyed clearly. Finally, the question organization was hierarchical, using screener questions to determine which modules would be asked to a respondent. The result was countless possible flows through the questionnaire and great variation in the interview length. Since it was impossible to pretest every possible path, detailed monitoring of timings during collection could provide details on the range of interview durations. As a result, this survey had much to gain from a detailed, ongoing evaluation of the data integrity during data collection. This paper discusses the new processes as they were applied to the CCHS 1.2.

2. METHODS

The CCHS 1.2 data quality monitoring program is more systematic and automated than standard approaches used at Statistics Canada. Information was drawn from the raw data as well as information from the Blaise audit trail, which provides a second form of data for analysis. The raw data provides the actual responses to the questions for each respondent and can be used to calculate rates of non-response in a questionnaire.

2.1 Blaise audit trail

Blaise is the software used to program the questionnaire questions, flows, edits, and skips into the computer. One option of this software is the ability to generate an audit trail for each case simultaneous with the data capture at the time of interview. The Blaise audit trail can be turned on for any questionnaire. It creates an individual file for each case that includes much of the same information as the raw data plus substantial additional information on timings and progression of questions, edits, and flows. Each file includes the following information for that respondent: the time each question was entered; the value the question had before the question was entered; the value the question had after the response was entered; any edit or suppression activity that occurred; and the time the question was exited. Time is to the second and there is a separate line for each activity, making it possible to follow the progression through the interview, including backtracking and changes to values. Although the files were used as aggregates to calculate statistics, the individual reports were available for examination of outlying cases.

2.2 Process

The files for individual cases were aggregated and merged every two weeks during data collection. The dataset of merged files are in CSV (comma separated value) format and are read into SAS. The file is flattened to calculate the time for each question. This feature takes into consideration the total time spent on each question, including any time spent in revisiting the question later in the interview. Each case is linked to the Interviewer ID using information from a separate log file attached to the interview. In this form, the audit trail data is available as a SAS dataset and can be analyzed for questions related to timings and edits in the questionnaire. The series of statistics used for the data quality monitoring program were identified prior to the start of data collection.

In supplementing the raw data, the audit trail data is very flexible regarding the type of statistics and the level of analysis available for producing reports. Statistics on non-response, edits, and timings were examined at the national, regional, interviewer and case levels. Non-response analysis was based on the raw data while analysis on edits and timings was based on the audit trail data. In general, reports were produced in series, presenting the same type of statistic first at the national and regional levels, followed by lists of interviewers whose average across interviews was outlying, followed by outlying individual cases.

Because there was no precedent with established criteria for determining outliers, levels were set arbitrarily at first and then refined as collection continued and results were examined. For example, the threshold for identifying outliers for many reports was set at the most extreme 2.5% of cases (2 standard deviations from the mean assuming a normal distribution). This was then modified when feedback from reports suggested a threshold that better suited the data.

2.3 Non-response reports

Non-response was considered from several different perspectives. First, many modules were set up so that a response of "don't know" or "refuse" to the first question in the module would lead to an immediate skip out of the module. This "module non-response" was monitored to verify that the subject matter of any module was not generating patterns of respondent rejection. Rates of module non-response were calculated based on the rate of "don't know" and "refuse" to the first question in each module where such a response led to an immediate skip out of the module.

Second, to check whether respondents might go through the modules but answer most of the questions as non-response, rates of item refusal were calculated as the average number of items refused in the entire interview. Third, specific questions were selected to be monitored more closely because of their potential sensitivity or because of concern about misunderstanding. Items selected for individual monitoring included questions on income, suicidal behaviour, and some of the introductions to the disorder modules. Parallel statistics were also generated for overall and specific item responses of "Don't know".

Finally, a pattern of "don't know" responses within a module might be indicative of a soft refusal pattern to that subject. To examine the potential for such a soft refusal pattern, the rate of "don't know" responses within each module were calculated. A high overall rate of "don't know" responses within a module could indicate that the questions may be perceived by respondents as sensitive or confusing or not applicable. A high rate of "don't know" responses for one particular interviewer could indicate that that interviewer might be coding refusals as "don't know" or some other pattern that could signal a need for reviewing procedures.

2.4 Timings reports

The major contribution of the Blaise Audit trail over standard monitoring is to provide timing and edit information at the question level. This information was used to compile reports on durations at different levels. The most important was the monitoring of interview duration. Interview duration excluding the time for the longest question was also examined since it is possible that many interviews might contain one long question that represented a period where the respondent took a break (e.g. to go to the bathroom, to get a drink, to answer the phone, to tend to a child, etc.) and that time should not be reflected in the overall interview duration. The ability to disregard the longest question is an advantage that the audit trail provides over simply looking at the timers built into the collection instrument.

In addition to examining straight interview duration, the audit trail provided the opportunity to examine interview durations standardized for the number of questions the respondent answered. This provides a more accurate basis for comparison across interviews since the survey was structured so that respondents received widely differing numbers of questions due to many complicated internal skip patterns. In order to standardize across interviews containing varying numbers of questions, interview rate was measured by seconds per question. This statistic gives a sense of the pace of the interview.

Table 1 illustrates a report on overall timings generated at the national and regional levels. In addition to the average duration, the minimum, 10^{th} percentile, median, 90^{th} percentile, maximum and standard deviation were presented to provide a better sense of the distribution of timings. This report was produced every two weeks and could also be generated on a cumulative basis for all cases completed to date. In this two- week period, the national average and median were comparable to previous reports and to predictions from the pretest so the interpretation would be that the overall timings were acceptable and as expected. Taken together, the statistics addressing the distribution indicate that, as expected, there was a lot of variability in interview length. The statistics on average seconds per question indicate that there is also a fair bit of variability in the pace of interviews. One reason for examining regional variations was because interviewer training was conducted separately by each region and variations across regions might indicate inconsistencies in training procedures that could be addressed and standardized.

Table 1 - Interview Duration - National and Regional Averages

Category	National	Atlantic	Quebec	Ontario	Prairies	BC
Average Duration (min)	63.4	60.0	66.1	64.2	60.3	69.1
Median Duration (min)	57.4	55.1	60.4	58.3	54.1	63.2
Min Duration (min)	17.3	19.5	22.0	17.3	18.6	18.1
10th Percentile (min)	46.5	43.4	40.1	38.1	44.1	49.3
90th Percentile (min)	95.0	91.2	98.6	96.6	90.0	103.0
Max Duration (min)	315.9	246.2	239.2	315.9	302.6	264.2
SD (min)	30.2	25.5	27.8	29.4	35.6	32.0
Average Duration (excl. longest question) (min)	58.4	55.2	61.1	59.5	54.8	64.4
Min Duration (excl. longest question) (min)	17.0	18.3	21.1	17.0	18.2	17.7
Max Duration (excl. longest question) (min)	275.8	213.7	228.9	275.8	248.2	235.2
SD (excl. longest question) (min)	24.6	23.2	25.1	25.2	22.2	26.8
Average Seconds per Question	10.1	9.7	10.6	10.3	9.7	10.6
Min Seconds per Question	4.5	4.7	4.5	5.4	4.6	4.6
Max Seconds per Question	64.2	45.0	35.2	44.5	64.2	50.9
SD of Seconds per Question	4.3	3.5	3.8	4.0	5.6	4.2
Average Seconds per Question (excl. longest question)	9.3	8.9	9.8	9.5	8.8	9.8
Min Seconds per Question (excl. longest question)	3.2	3.2	3.3	3.4	4.6	4.5

Typically, a report like this was followed by additional reports for the same statistics at the interviewer and case levels, which identify outliers. These reports provide more information about both the short and long cases reflected in the overall table. Reports identifying interviewers whose average interview duration over all their cases was extremely long or short were examined and compared with previous reports to see whether some interviewers appeared to have developed patterns of either short or long interviews that might require further investigation.

A final method used to compare timings across interviews was by defining miniblocks. A miniblock is a series of questions identified as consecutive with no internal skips. Because respondents who answer the first question in the miniblock will also answer all the other questions in the series, it is possible to compare equivalent timings for miniblocks across interviewers and across respondents. This provided an alternate method to assess the pace of interviewing for questions in different modules. Examining the averages for miniblocks could identify overall changes in the pace of the interview between different modules. For example, is there a module where the pace slows down considerably, potentially indicating that the questions are difficult for respondents? Conversely, is there a module where the pace seems very fast, possibly indicating that interviewers are reading more quickly than was recommended in training?

2.5 Edit Reports

The audit trail also provides data on the incidence of hard and soft edits, and suppressions. When an edit is triggered, the interviewer is presented with a dialog that explains the problem and allows the interviewer to return to one or more previously answered questions. A hard edit occurs when logically impossible data has been entered. A set of answers are in conflict with each other, and the conflict must be resolved before any other interviewing can be done. For example, a person may have replied that they sleep 12 hours a day, and work 1-4 hours a day. Individually, both answers are possible. However, in combination they are in conflict because they total 26 hours. A soft edit occurs when unusual data has been entered. This most frequently happens on an open-ended range question, where high values are unusual but not impossible. For instance, a soft edit might be triggered if a person replied that they sleep 20 hours a day. The interviewer can either change the answer or suppress the edit if the respondent confirmed that they really do mean 20 hours a day.

Monitoring of hard edits, soft edits, and suppressions is valuable as feedback for questionnaire design. A high number of edits for a particular question suggests that the range of responses for that question be reconsidered since many respondents initially provided a response that was not consistent with those anticipated by those who designed the survey. A high number of suppressions for soft edits suggests the same or may suggest that further analysis is needed to understand why respondents are reporting seemingly contradictory information across questions. This type of information can be used to modify the questionnaire for future use, and would be a particularly valuable tool if used as part of a pretest.

3. RESULTS AND DISCUSSION

As outlined, the monitoring program generated many reports. Potentially sensitive or confusing questions were examined for duration, non-response, and edits to ensure that they were not posing problems for respondents. Interview timings were examined for changes as data collection progressed and for regional variations to identify potential differences in training and procedures that impacted the data. Rates of non-response were monitored to ensure that there were no systematic or individual issues that needed addressing. For each type of statistic, interviewer and cases with outlying patterns were identified and examined. Interviewers with outlying patterns on multiple reports were identified as well. Record was made regarding whether interviewers with outlying patterns during a particular reporting period had been previously identified as outlying in an earlier reporting period. In this way, interviewers and cases requiring follow-up were determined.

Overall, the monitoring program reassured us about the good quality of data being collected. For example, the results showed that concerns about stigma or sensitive questions did not translate to non-response. Few edits for questions identified as potentially confusing showed that multiple concepts in long questions were clear enough. Across multiple indicators, the reports suggested that respondent burden was not an issue.

The data monitoring program was designed as a tool for collection management to add to their existing quality control measures. Consequently, where individual follow up was necessary, actions were taken by those responsible for collection: the regional offices or those overseeing the survey from Survey Operations Division. Information about interviewers with outlying patterns was sent to collection management who integrated it with their own quality control procedures. For example, if a case with a particularly long interview was identified, the senior interviewer would ask the interviewer about it to see how the interview went and how the respondent had felt about the interview. We found that those cases were usually more strenuous for the interviewer than the respondent since respondents who had numerous problems tended to enjoy having someone to listen to them. In the example of an interviewer having a pattern of short interviews, an interviewer might receive retraining to improve the quality of their interviews. Emphasizing procedures such as talking slowly and asking questions in a non-directive manner are some examples of addressing issues raised through monitoring.

When general issues were identified, they were addressed as part of a newsletter received by all interviewers. For example, where the miniblock timing of a module seemed to be faster than anticipated and further examination indicated that the response pattern was one of consistent “no” responses, it was speculated that the “no” pattern was a reflection of a soft refusal where the respondent refused to listen to each question being read in full and wanting to skip over similar questions. The newsletter reminded interviewers that the correct procedure in a situation where the respondent forces them

to skip questions was to enter a “refusal” response instead of a “no” response. In addition, such information about how the instrument is being used by interviewers can be used to modify future versions of the module.

Information on flows and edits provided feedback on whether skips provided a smooth interview experience for the respondent and whether the response categories were consistent with the types of answers respondents were giving. For the most part, this information confirmed that flows were not problematic, but was also useful in identifying a couple places that might be more respondent-friendly if a skip were added at a certain question or if a new category was added to the response set. Questions that had a high number of edits or that tended to be revisited can be modified to be more respondent-friendly. This information will be used to improve future surveys.

While a strength of the audit trail is the volume, depth, and breadth of information available, this can also be a challenge in terms of sorting through and evaluating all the data generated by the monitoring program. It is necessary to focus on key indicators and ensure that resources are available to review and act on the reports without delay. Resources are also required to set up the statistical programs, and to unpack the data and run reports on a regular basis. Similarly, it is important to set up mechanisms to deal with any issues encountered through monitoring. One area that requires care is how to deal with interviewers who are identified with outlying patterns. In this regard, the results of the monitoring program were used in conjunction with other data quality monitoring and interviewer management activities that collection management employs on a regular basis in order to determine whether the interviewer’s procedures were problematic and, if so, what type of remedial action would be most appropriate. Results of the monitoring program were never used in isolation from other collection management activities.

The audit trail information represents a valuable addition to standard monitoring procedures while the data was in the field for a number of reasons. It was useful for demonstrating that, by and large, estimates made during questionnaire design were accurate and that survey procedures were followed consistently. It identified the odd break with standard procedures requiring follow up at the interviewer or case level. It provides the ability to identify or confirm more informal evidence about interviewers who require retraining because of poor performance. The audit trail provided for better evaluation of module length and other issues of timing. Because timings are available to the question level, it is valuable as a tool for evaluating timings for complex questionnaire structures. It can be used in questionnaire planning where there are different options for module inclusion or exclusion. This will help design questionnaires that can meet the various demands of more and more stakeholders who have different requirements for content (e.g. optional modules for provinces). The tracking of edits and suppression enables survey-makers to get a better understanding of aspects of their survey that may be of lingering concern. Finally, the data from the audit trail can be analyzed quickly, allowing for prompt turn-around (identification of and addressing of issues). This is very important for maintaining high quality of standards throughout the collection process.

In summary, the use of the new technology of the Blaise audit trail enriched the existing data quality control program. Perhaps most importantly, it provides a level of detailed information on what actually happens in the field not previously available for computer-assisted personal interviews. Previously, feedback about how the questionnaire was received and how respondents reacted was anecdotal through interviewer feedback. The information from the audit trail adds concrete data that can be analyzed to provide more specific information on how the instrument works across all interviewers and across different types of respondents. Together, the new information on non-response, timings, and edits can be used to improve future questionnaire designs.