Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada's Symposium 2005
Methodological Challenges for Future Information Needs

DATA QUALITY CHALLENGES IN THE SURVEY OF PRINCIPALS

Martin Renaud, John Stardom[1]

ABSTRACT

The Survey of Principals collects information on various topics related to the work of school principals. One of the survey's challenges was to secure adequate participation to meet data quality requirements. Another challenge relates to how well principals understand the questions. Data quality is also affected by response errors on their part. A third challenge concerns electronic data capture, as the capture system produces various types of errors which also diminish the quality of the data. This article presents a brief description of these challenges and provides examples to illustrate the problems observed. The steps taken to resolve the challenges are described, and their impact on data quality is analyzed. Some lessons learned are also presented, in case the survey is ever repeated or a similar survey is conducted.

KEY WORDS: School principals, data quality, response errors, electronic data capture, capture errors

1. INTRODUCTION

Any survey whose collection method is a self-administered questionnaire that is mailed out and returned will present a number of challenges relating to the participation rate and the quality of information collected. Unlike surveys that have computer-assisted data collection, this type of survey does not enjoy the kinds of IT features that improve data quality with edit rules built into the collection process. There is also an additional source of error when data are captured from paper questionnaires. The Survey of Principals (SOP) had to contend with these data quality problems. Section 2 of this article presents a brief overview of the SOP. Section 3 describes the participation problems in greater detail. Sections 4 and 5 cover response errors and electronic data capture problems respectively. Lastly, section 6 discusses the lessons learned from this survey, which will help improve data quality if the survey is repeated or a similar survey is carried out.

2. OVERVIEW OF THE SURVEY

2.1 Objectives and target population

The Survey of Principals was sponsored by the Social Sciences and Humanities Research Council of Canada and was conducted jointly by Statistics Canada and a team of researchers from the faculties of education of the following Canadian universities: Université de Sherbrooke, Université de Montréal, University of Toronto and Simon Fraser University. Statistics Canada was responsible for the sample frame, the sampling plan, sample selection, survey operations (data collection and capture) and data processing (edit, imputation, weighting, calculation of the coefficients of variation). The researchers were responsible for the survey's content.

The survey's main purpose was to collection information from school principals for use in evaluating the impact that various changes observed in education, such as curriculum changes, budget reductions and new policy directives, had had on teaching and the work of principals in Canadian schools. The survey was also designed to gather information from principals about their work, their skills, their professional situations, their daily interaction with students and teachers, and the effect that the changes in the world of education have had on their work in general.

The survey's target population was all principals in Canadian elementary and secondary schools, excluding continuing and adult education schools, trade schools, language and cultural education schools, community education centres, social service centres, distance education schools, virtual schools, schools in Aboriginal

[1] Martin Renaud, Statistics Canada, 16-F R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, martin.renaud@statcan.ca; John Stardom, Statistics Canada, 16-E R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, john.stardom@statcan.ca

communities, and home schools. In all, some 15,500 schools were eligible for the survey across Canada. If a principal worked in more than one school, all the schools he worked in were included in the survey. Hence a principal might have received more than one questionnaire if two or more schools where he worked were selected for the sample. This is acceptable for the survey because a number of questions apply to the principal's work within the school.

2.2 Collection tool

A paper questionnaire was used to collect the desired information for the survey. It was 23 pages long and consisted of six sections on the following subjects: socio-demographic information and school characteristics, perception of change and its impact, duties and responsibilities, social relations in schools, professional integration and development, and projects and educational goals. All questions in every section but the first were scale-based opinion questions. The questionnaire, in English or French depending on the school's language of instruction, was mailed to the principal of each school in the sample. He or she was to complete it and mail it back to Statistics Canada.

2.3 Sampling plan

SOP schools were stratified first by region (Atlantic, Quebec, Ontario, Prairies, British Columbia and territories) and type of school (elementary, secondary and mixed). To ensure that the sample was consistent with the distribution of certain variables within the population, schools were then sorted within each stratum by main language of instruction (English or French), type of funding (private, public or mixed), geography (urban or rural) and size (small, medium-sized or large). Then the sample was selected systematically in each stratum. For further information, see Stardom (2005).

The total sample size of 4,800 schools was determined on the basis of the response rate for a similar survey conducted previously, the required level of quality and the survey's budget constraints. The sample was allocated in two stages: an allocation proportional to the square root of the number of schools by region (except the territories) followed by an allocation proportional to the number of schools by type in each region. For the territories, all schools were surveyed. This allocation was chosen because it substantially improved the smallest strata's coefficients of variation without a significant loss of efficiency for the national and largest strata's coefficients of variation. The population and sample sizes for each stratum are listed in Table 1.

Table 1: Population and sample size for each stratum

	Elementary		Secondary		Mixed		Total	
	Pop.	Sample	Pop.	Sample	Pop.	Sample	Pop.	Sample
Atlantic	689	349	317	161	202	102	1,208	612
Quebec	2,216	716	604	195	139	45	2,959	956
Ontario	4,182	981	1,134	266	231	54	5,547	1,301
Prairies	1,525	455	745	223	1,165	348	3,435	1,026
British Columbia	1,305	518	494	196	175	70	1,974	784
Territories	47	47	20	20	54	54	121	121
Total	9,964	3,066	3,314	1,061	1,966	673	15,244	4,800

2.4 Data collection

Data were collected between late October 2004 and mid-February 2005. Through collection operations, it was determined that about 3% of the sample schools were out of scope. After those schools were removed from the sample, the survey's response rate was about 48%. All questionnaires that were completed and returned to Statistics Canada went through an electronic data capture process. More details about the questionnaire, the participation rate and electronic data capture are provided in the sections below.

3. CHALLENGE 1: PARTICIPATION PROBLEMS

Over the last few years, principals' administrative duties have been expanding, and their budgets have been shrinking following cuts in education funding. Principals are also being asked to take part in an ever-growing number of surveys in the schools. Since they have little time to respond to those surveys, and since completing surveys is not their top priority, it is difficult to secure their cooperation. As the SOP used a mail-out, mail-back, self-administered questionnaire, it is not surprising that the response rate was not as high as if a personal interview had been used, for example. To maximize the response rate, the SOP first contacted all school boards in Canada to inform them of the survey's objectives and request their cooperation. Almost all boards agreed to take part in the survey. Each board then contacted those of its schools which were in the sample and encouraged the principal to participate.

In addition, in an attempt to encourage a higher rate of participation, each sampled school was sent an introductory letter explaining the survey's objectives and importance. Despite this initiative, a number of principals subsequently indicated that they would like to have received more information about the survey.

The principals were given about three weeks to complete the questionnaire and return it to Statistics Canada. If the questionnaire had not been received after three weeks, follow-up was carried out by telephone and facsimile. Before follow-up, the response rate was about 10%. By the end of the collection period, the response rate was 48%, due in part to the follow-up operation.

During collection, some principals indicated that they would prefer to return the questionnaire by fax rather than by mail. Others asked if the questionnaire could be sent to them by e-mail in PDF format. A few principals completed their paper questionnaire, converted it to PDF and returned it by fax. These mail-out and mail-back options had not been anticipated in the initial planning. However, including them in the collection procedures helped improved the final response rate, since one out of eight respondent principals used one of those methods.

In the end, the steps taken to improve the response rate were effective. Since it is not unusual for a survey of school principals to have a response rate in the vicinity of 40%, the SOP's 48% is considered satisfactory. Moreover, extending the follow-up period very likely would have pushed the response rate over 60%.

4. CHALLENGE 2: RESPONSE ERRORS

4.1 Advance testing

Any survey whose collection instrument is a self-administered questionnaire is more susceptible to response error than a survey whose data is collected by computer-assisted interview, for example. The unfortunate result is a larger number of errors due to respondents' misinterpretation of the questions. Accordingly, questionnaire design is very important in minimizing response error.

To make the SOP questionnaire easier to complete and reduce the number of response errors, consultations were held with principals prior to the survey. About 10 principals in each of three cities – Moncton, Montréal and Toronto – were selected to complete the questionnaire. Subsequently, feedback sessions were organized to obtain their comments and suggestions on how to improve the questionnaire. A pilot survey was also conducted with 200 principals chosen at random from the target population. The results of the consultations and the pilot survey were analyzed to identify ways of improving the questionnaire. This led to the reformulation of a number of questions that were considered confusing. The terms used to explain certain concepts were clarified. The scales for the opinion questions were standardized to the same number of response choices. Despite all the improvements made in the questionnaire, response errors were seen during the actual survey. The errors can be divided into two types: generalized errors and localized errors.

4.2 Generalized errors

Most of the SOP's response errors occurred in the school characteristics questions in section 1 of the questionnaire. The socio-demographic questions in section 1 and the opinion questions in sections 2 through 6 were generally answered correctly. Renaud (2005) provides more detail about editing and error correction for the SOP.

Several types of errors were repeated in more than one question. For example, questions in which a total and its components were requested were often poorly answered. The most common error was a total being different from the sum of its components. In some other questions, the sum of the components was supposed to be 100%, but unfortunately this was not always the case. These inconsistencies were resolved by checking the electronic image of the faulty questionnaires. The images, which were available for all completed questionnaires, helped correct errors properly.

In some cases, questions that were supposed to be answered with a quantity were answered with a percentage. This was most prevalent in the case of large schools, as it was easier for the principal to provide a percentage than a specific count. Conversely, questions that asked for a percentage were sometimes answered with a quantity. This was more common for principals of small schools, who found it easier to provide a specific count than a percentage. To correct these errors, the counts were simply converted to percentages, and the percentages to counts.

4.3 Localized errors

Some response errors affected specific questions. On the question in which principals were asked how many staff members hold full-time and part-time positions in the school in various categories, some answered in full-time equivalents. For example, suppose a school has 15 full-time teachers and one part-time teacher working 60%. The correct way to answer the question would be to report 15 full-time teachers and 1 part-time teacher. However, some principals reported 15.6 full-time teachers and 0 part-time teachers. Some even reported 15 full-time teachers and 0.6 part-time teachers. These errors were corrected so that the data were in the proper form. On the same question, some principals shifted their responses down one line. The way the response boxes were laid out on the questionnaire is probably the main reason for this error. In many cases, this resulted in outliers for some response categories. For example, no school has 52 vice-principals; this figure is the number of teachers at the school. This type of error was corrected by moving the responses to the proper categories.

Another question dealt with the approximate percentage of students in the school who were from low-, medium- and high-income families (the category boundaries were specified in the question). The questionnaire provided only two boxes for the percentage in each category. This flaw went unnoticed in the design phase and advance testing. It had not been considered that all the students in a school could be from families in the same income category. This did occur, however, in some of the smaller schools. To remedy this oversight, the electronic images of all questionnaires with just one value of 99%, 10% or 00% were reviewed. Those three values were targeted because a principal might have used 99% to represent 100% or might have written 100% with the "1" or the "0" outside the response box. In cases where this occurred, the values were converted to 100% to correct the errors.

In another question, principals were asked to indicate the percentage of teachers in their school who were in each of five years-of-experience classes. Instead of including only teachers (full time and part time), some principals reported the percentages for different subgroups of school employees (e.g., all full-time and part-time employees, only full-time employees, etc.). These errors were detected using the numbers of employees reported in various categories in a previous question. For example, in a school with 10 teachers, how can 34% of them have more than 20 years' experience? Unfortunately, while the errors were identified, they usually could not be corrected since it was impossible to tell which employees the principal had included in the values reported.

Nevertheless, it was possible to correct most of the response and interpretation errors in section 1. The error rate varied widely from question to question, from as low as 3% to as high as 40%. By the time the errors had been corrected, about 80% of the records had undergone at least one change.

5. CHALLENGE 3: QUALITY OF DATA CAPTURE

Electronic capture with a scanner is much faster than manual capture and extremely effective for scale questions such as those in sections 2 through 6 of the SOP. For alphanumeric questions such as those in section 1 of the SOP, simple edit rules can also be programmed to ensure consistency in the data. On the other hand, the use of a character recognition program is necessary for these questions, which sometimes introduces errors that would not have occurred in manual capture. To capture the data, each questionnaire is fed one page at a time through the scanner, which produces an electronic image of each page. Then, following pre-programmed instructions, the scanner searches for the data for each question at specified coordinates (x,y).

For the SOP, questionnaires returned by fax or in PDF format helped increase the response rate, but they also caused problems in the capture process. The reason is that such questionnaires were in a format different from the one required for scanning. For example, the banner inserted at the top of each fax page moved the questionnaire text down. As a result, electronic capture of such questionnaires produced erroneous data that did not reflect the principals' responses. To remedy this situation, we had to manually capture the data from all questionnaires not returned by mail.

Another problem frequently encountered in electronic capture was caused by responses marked outside the response boxes in the questionnaire. On a number of occasions, principals marked their answer and then changed their minds. So they put a line through the response in the boxes and marked the correct response beside the boxes. In such cases, the scanner was unable to convert the responses into legible text. As a result, the response was simply left blank. In addition, some principals provided responses with decimals, even though the instructions in the questionnaire clearly stated that they were to round their responses to the nearest unit. This automatically produced a capture error, as the scanning software was not programmed to recognize responses with decimals. For example, the scanner would interpret a value of 2.1 as 21. To remedy these two problems, the capture operators had to inspect the image of each page of each questionnaire and manually input all data that were marked outside the response boxes or contained a decimal. In all, 20% of the questionnaires had undergone at least one change by the end of this inspection process.

Two other problems encountered during electronic capture were caused by the scanner. First, since it is a highly sensitive electronic instrument, the scanner has to be calibrated at regular intervals. Because of poor calibration, data that had been properly entered in the questionnaire were sometimes completely missed by the scanner, which meant a missing value for that question. Second, the character recognition software sometimes misinterpreted the principal's response. For example, an "8" might be read as a "0". A third problem found during electronic capture stems from the quality control process. In some instances, principals entered a value outside a response box, but that value was not a response to the question. For example, some principals entered "100%" below the last box in a question to show that their responses added up properly. When the questionnaire images were inspected (to identify cases in which values had been entered outside the response boxes), the values marked below the final box were misinterpreted as outside-the-box responses and were captured incorrectly. Fortunately, these three types of errors were detected and corrected with edit rules. In all, about 4% of the questionnaires were affected by one of these errors.

Despite the problems caused by electronic data capture, in the end they did not affect data quality as they were successfully corrected. In fact, their main impact was to increase processing time and costs. Since almost all of the errors occurred in the alphanumeric questions in section 1, it would be better in the future to capture the data from such questions manually and use electronic capture for scale questions such as the ones in sections 2 through 6, which were virtually error-free.

6. CONCLUSION

In general, the SOP results can be considered very satisfactory. Taking everything into account, even though the survey went very smoothly, several important lessons were learned that will help avoid certain errors if the survey is repeated or a similar survey is conducted. To address the participation problem, it would be a good idea to consider making the questionnaire shorter than 23 pages. In view of the interest expressed by principals, they should be provided with more information about the survey. Further, from the outset, the possibility of allowing respondents to return their questionnaire by fax or in PDF format should be considered. An extended collection period would help

increase the response rate. The advance consultations were very helpful in reducing certain response errors, and they should be retained as well as the intensive follow-up effort during the data collection period. Finally, as a reward for their participation and an incentive to take part in future surveys, the survey results should be shared with all principals who respond.

One way of reducing response errors in the future would be to provide a glossary of the terms and concepts used in the survey. To provide better guidance for respondents, the instructions in the questionnaire should be made clearer and more specific. To prevent confusion, perhaps all the questions could use the same form of response: either quantities or percentages.

To improve the data capture process, allowance should be made in advance for the possibility that some principals will want to return their questionnaires by fax or in PDF format. In addition, for a small-scale survey such as the SOP, it would be more effective to input alphanumeric data by hand. Alternatively, electronic capture could be used for scale questions, in which the respondent simply marks a box.

REFERENCES

Renaud, M. (2005), ‘‘Vérification et imputation des données de l’Enquête auprès des directeurs d’école’’, unpublished report, Ottawa, Canada : Statistics Canada.

Stardom, J. (2005), ‘‘Sample Allocation and Selection for the 2005 Survey of Principals’’, unpublished report, Ottawa, Canada : Statistics Canada.