

# CODING OF RESPONDENT BEHAVIOUR BY INTERVIEWERS TO TEST QUESTIONNAIRE WORDING

M. J. Burgess and D. Paton, Statistics Canada

D. Paton, 16th floor, R. H. Coats Building, Statistics Canada, Ottawa, Canada, K1A 0T6

## Abstract

**KEY WORDS:** Behaviour Coding, Questionnaire Testing, Computer Assisted Telephone Interviewing.

Behaviour coding to test questionnaire wording traditionally involves analyzing the behaviours of both the interviewers and respondents using recorded interviews. This analysis is usually based on a small number of respondents due to the cost of analyzing the recorded interviews and the results are not available quickly due to the time required to review the recorded interviews. This paper describes a method for coding respondent behaviour which does not require the recording of interviews. This method was implemented and tested at Statistics Canada in a Computer Assisted Telephone Interviewing (CATI) environment where it made possible a detailed and timely analysis based on all of the respondents in the survey pretest. This method has the disadvantage that it only allows analysis of the behaviour of the respondents and not that of the interviewers. The challenge was to develop a method for interviewer coding of respondent behaviour that would not negatively affect the interviews, and to find strategies for overcoming the limitations of the method. The paper describes and comments on the success of these strategies. It is concluded that the advantages of this method in terms of cost, coverage and timeliness will often outweigh its disadvantages, especially for long and complicated questionnaires and when pretests include large numbers of interviews.

## 1 Introduction

Behaviour coding is a technique used to evaluate questionnaire wording, interviewer performance, and the interview process as a whole (Dijkstra and Van der Zouwen, 1982; Sykes and Morton-Williams, 1987; Oksenberg, Cannell, and Kalton, 1991). It allows for a rigorous analysis of the behaviour of both the interviewers and the respondents since their behaviours are classified and coded by trained coders listening to live or taped interviews. While this methodology (**interview behaviour coding**) provides a powerful tool for testing it is not without its disadvantages, the most important of which is that it is labour intensive; the coders must listen to each interview

thus essentially doubling the number of person-hours spent per interview. As a result the analysis of behaviour codes is usually based on a small number of interviews (60 in the study of Oksenberg, Cannell, and Kalton (1991) and 89 and 68 in the studies cited by Sykes and Morton-Williams (1987)).

This paper describes a method for coding **respondent behaviour** which does not involve a third person listening to the interview. This method (**respondent behaviour coding**) thus avoids some of the disadvantages of the **interview behaviour coding** method, but has the limitation that it only allows analysis of the behaviour of the respondents and not that of the interviewers. The advantages of this method in terms of cost, coverage and timeliness will often outweigh this limitation. This method was developed and implemented during the design of a Computer Assisted Telephone Interviewing (CATI) survey at Statistics Canada. The CATI technology made the implementation of this method particularly efficient.

## 2 The Survey

In 1991 Statistics Canada began developing a new survey on violence against women. There were many dimensions of the survey topic that were (and continue to be) of interest including: perception of personal safety and measures taken to enhance it; sexual harassment; experiences of sexual and physical assault by strangers, dates/boyfriends, husbands or common-law partners, and other known men; power and control and emotional abuse in relationships with husbands or partners; spouse abuse by the respondent's father or father-in-law; a number of demographic items to be used as covariates in tabulations and analysis; and for women who report being a victim of violence: the impact of the violence, who they turned to for help, and involvement and satisfaction with the criminal justice system following the incident. The measures of sexual and physical assault needed to be consistent with *Canadian Criminal Code* definitions that are not well known to most people. Statistics Canada had conducted surveys on criminal victimization previously, but detailed questions about violence against women had not been included so many of the questions were new and

untested. The draft questionnaire contained over 300 questions, of which many were new questions, many were questions on sensitive topics, and many involved complicated definitions.

There were a number of concerns about the questionnaire and the survey. The most important ones were: i) that responding to the survey would endanger the well-being (psychological and/or physical) of some of the respondents, ii) that respondents would find the questions too sensitive and would not disclose their experiences, iii) that respondents would find the interview to be too long and not complete the questionnaire, iv) that respondents would have difficulty understanding the questions, v) that the characteristics of the ideal interviewer for this survey might be different from those for another type of survey, and vi) that the CATI system could handle such a long and complex questionnaire.

The first concern was addressed through the survey's procedures (having a toll-free line that respondents could use to call back if needed for any reason, having victim services information available, helping interviewer become sensitive to difficult situations). It was hoped that the second, third, and fourth concerns could be addressed through the design of the questionnaire and that some insight would be gained into the fifth and sixth through the implementation and analysis of the pretest.

### 3 The Pretest

The main pretest for this survey was designed to provide feedback on all of the concerns mentioned above and so it had to test the survey procedures, the questionnaire, the interviewers, and the CATI system. It was expected that only about 10% of the pretest respondents would be asked an important block of questions. To ensure that adequate numbers of respondents would be asked these questions, the target sample size was about 1,000 completed interviews.

The methods that had been used by the survey development team to pretest other social survey questionnaires were the observation of field operations and monitoring of interviews by the team, the debriefing of interviewers and other field staff, and the analysis of the captured data. These methods are described in DeMaio (1983) and Converse and Presser (1986). While these methods had proven to be effective they are largely subjective and there was a general feeling that it would be desirable to supplement them with something that would add a more objective measure to the evaluation. The

methods described by Dijkstra and Van der Zouwen (1982), Sykes and Morton-Williams (1987), and Oksenberg, Cannell, and Kalton (1991) would have provided such objective measures. However, there were problems with implementing such procedures in the pretest: if interviews were tape recorded for subsequent coding, the respondents would have had to be informed of the taping and there was a concern about the effect of this on non-response and on the quality of response given the subject matter of the survey; and the budget did not allow for the coding of interviews by behaviour coders.

It was decided to try to use the interviewers to code the behaviour of the respondents during the interview. It was realized that this would only be feasible if a classification system and a capture procedure could be developed that would be so easy to apply that the normal flow of the interview was only minimally affected. It was also realized that the CATI software being used would allow the interviewers to simply capture codes that described the behaviour of the respondents. Hence, it seemed to be possible to analyze the behaviour of all the respondents at minimal cost and without the problems associated with taping the interviews and without the delays needed for coding taped interviews.

## 4 Respondent Behaviour Coding

The set of respondent behaviour codes that was used was designed to be easy to use by the interviewers and to address some of the specific concerns of the survey. It was felt that it might be too difficult for the interviewers to record codes that related to the inadequacy of the responses received as they would be busy probing for an adequate response. The set used addressed the issues of respondent's difficulties with the questions due to their complexity or length, of the length of the interview and of the sensitivity of the subject matter.

A simple coding system for 5 categories of respondent behaviour of interest was used. During the pretest, whenever the respondent either: i) asked that the question be repeated, ii) asked for the clarification of the question or any part of the question, iii) interrupted the interviewer while the question is being asked, iv) asks how much more time the interview would take, or v) seemed to be uncomfortable with the question or providing the response, the interviewers were to record the corresponding code. More than one code could be captured per question.

The CATI software (USBC) used for this survey

allowed comments to be recorded by the interviewers. These comments were on a separate file from the data, but they were identified by the question on the screen at the time they were entered. To make a comment, the interviewer had to press a function key, type the comment, and finish with two slashes and enter. The interviewers were trained in this procedure and applied it during the pretest.

The questionnaire for this test was long and complicated: there were more than 300 questions. During the test our target population was Canadian women aged 18 to 65 living in the ten provinces. The pretest was conducted using random digit dialing with samples in two geographic areas; in one area interviews were only conducted in English and in the other only in French. The total sample size was 1748 and complete responses were obtained from 940 respondents: 644 English and 296 French. There were 11 interviewers, classified into four groups: those with crisis-line or social work experience but no interviewing experience, experienced interviewers from an employment agency, experienced Statistics Canada interviewers, and experienced Statistics Canada CATI interviewers.

After data collection, since the behaviour codes were in the same file as any other notes made by the interviewers, the codes were extracted from the notes file using a short manual procedure (a couple of hours). Frequency distributions were then obtained to determine the questions that had been assigned the most codes.

While it was expected that the coding of respondent behaviour would produce some useful data, it was realized that there were some problems with the method: interviewer effect, both through interviewers' varying abilities to ask the questions clearly and their varying propensities to code borderline respondent behaviour; information about the adequacy of the responses was not obtained; information about the behaviour of the interviewers was not obtained; and while the location of potential problems might be indicated, the underlying reason for the problem would not be.

## 5 Results

The final sample after non-response of the pretest consisted of 940 respondents (the sample before non-response consisted of 1748 potential respondents). On average, 117.2 questions were answered and 2.86 codes were used per questionnaire. The use of the five codes is summarized in Table 1.

The rare use of the "T" codes (see Table 1 for the

Respondent Behaviour	Code	Average Number of Codes per Questionnaire
All Codes		2.86
Asks for Repetition	R	.54
Asks for Clarification	C	.76
Interrupts Interviewer	I	1.53
Asks about Time	T	.03
Respondent seemed Uncomfortable	U	.003

**Table 1.** Use of respondent behaviour codes.

code definitions) was in agreement with the observation that few partial interviews were obtained and that respondents, once convinced to participate in the survey saw the survey as important and wanted to complete the interview (When the main survey was in the field, respondents often called to make appointments to ensure that they could complete their interviews.) The "U" code was very rarely used, in fact only by one interviewer. It is not clear whether respondents were comfortable with the questions, whether they were effective at suppressing that discomfort or whether the interviewers were not very good at noticing and coding discomfort when it was displayed. Nevertheless, the rare use of these two codes did diminish somewhat the concerns about the length of the interview and about the sensitivity of the questions.

Table 2 presents numbers of codes assigned by type for the twenty questions with the most codes. A pattern was noticed with the codes "C", "R", and "I": questions with many "C" codes also had many "R" codes and questions with many "I" codes tended to not have many "C" or "R" codes. The most common problems were with questions with large numbers of "I" codes.

When the "interruption" questions were examined as a group they were found to be long questions that in several cases would for many respondents be asked following a very similar long question. This led to the concern that respondents were answering the question before they had absorbed the difference

Question Number	Behaviour Code			
	All	C	I	R
F8	263	7	245	9
F5	183	3	176	4
B3	131	7	119	4
F19	130	2	126	2
A12	112	11	97	4
G18	108	7	101	0
F2	102	7	91	4
R3-1	85	47	2	36
C19-1	59	41	0	18
A20	55	39	5	11
A24	42	26	14	2
A2	41	13	3	24
C32	39	3	33	3
C23	38	19	1	17
D5	38	24	0	14
E15	33	6	21	6
J15-1	31	1	29	0
K2	31	5	24	1
K3	31	17	1	13
R4-1	30	11	0	18

**Table 2.** Number of respondent behaviour codes assigned by question.

between it and the previous one.

The most commonly coded questions following this pattern were in a series of three questions that were long and complicated with only slight differences (F2, F5 and F8). The first of these was:

Since the time you were 16 years old, have you ever been forced into a sexual activity with a **male stranger** when you did not want to because he held you down, threatened you, slapped you or hurt you in some other way?

For the second and third questions in this series, **male stranger** was replaced with **date or boyfriend** and **male acquaintance**. While there were a number of questions separating these, if the response was "No" to the first, the next question asked was the second in the series, and similarly with the second and third, so that in fact the most common sequence of questions asked and responses given was:

Since the time you were 16 years old, have you ever been forced into a sexual activity with a male stranger when you did not want to because he held you

down, threatened you, slapped you or hurt you in some other way?

*No.*

Since the time you were 16 years old, have you ever been forced into a sexual activity with a date or boyfriend when you did not want to because he held you down, threatened you, slapped you or hurt you in some other way?

*No.*

Since the time you were 16 years old, have you ever been forced into a sexual activity with a male acquaintance when you did not want to because he held you down, threatened you, slapped you or hurt you in some other way? By acquaintance we mean a neighbour, friend, teacher, doctor, someone at work, clergy, lawyer, relative, or any other person you know.

*No.*

The obvious problem with this sequence of questions was noticed while monitoring the interviews, but respondent behaviour coding gave an objective measure of the extent of the problem. Almost 19% of respondents interrupted the interviewer during the second question in the series while 26% did so during the third.

This section of the questionnaire was revised extensively following the pretest. The time reference was moved to the introduction to the section and reiterated in the third of these questions. The questions were changed so that they began with the part that changed, the description of the offender. New questions were introduced for other reasons that incidentally had the effect of breaking up the repetitive sequence. With the final questions the above sequence of questions and answers became:

It is important to hear from women themselves if we are to understand the very serious problem of male violence against women. I'm interested in whether any of the following has happened to you since the age of 16. Your responses are important whether or not you have had any of these experiences.

Has a MALE STRANGER ever forced you or attempted to force you into any SEXUAL activity by threatening you, holding you down or hurting you in some way?

*No.*

Has a MALE STRANGER ever touched you against your will in any sexual way, such as unwanted touching, grabbing, kissing or fondling?

*No.*

Has a DATE OR BOYFRIEND ever forced you or attempted to force you into any SEXUAL activity by threatening you, holding you down or hurting you in some way?

*No.*

Since you were 16, has a MAN YOU KNOW ever forced you or attempted to force you into any SEXUAL activity by threatening you, holding you down or hurting you in some way?

*No.*

Since you were 16, has a MAN YOU KNOW ever touched you against your will in any sexual way, such as unwanted touching, grabbing, kissing or fondling?

*No.*

The “clarification/repetition” questions could be generally divided into two groups. One group consisted of open ended questions designed to collect information during the pretest but not intended for the main survey and the other of some questions from the classification section about work and income for which one might expect some reasonable demands for more precision. The questions in the first group were dropped from the main survey, while those in the last remained.

Logistic regressions were used to examine the factors affecting the probability that a questionnaire would have one or more codes assigned. The model included language of interview (two levels), age of respondent (four levels), education of respondent (three levels), whether or not the respondent reported a victimization, type of interviewer (four levels), and week of the interview (six levels). None of the respondent characteristics proved significant at a 90% level, while there were highly significant effects from language of interview and type of interviewer. There was no consistent pattern of changes in use of the codes over time. At least one code (of any type) was used in 73% of the interviews, more often with French interviews (85%) and less often by the group of inexperienced interviewers (51%).

It was noted that one interviewer was an outlier, with long interviews and far more extensive use of

the behaviour codes than the others (18.9 per interview vs the overall average of 2.86). This analysis was repeated with the data from this interviewer excluded and the results were similar. Again, none of the respondent characteristics proved significant at a 90% level, while there were highly significant effects from language of interview and type of interviewer. A code was used in 71% of the interviews, more often with French interviews (85%) and less often by the group of inexperienced interviewers (31%). Further analysis was performed with the data from this interviewer excluded (899 questionnaires).

A similar analysis was made of the use of the codes for clarification and repetition. These results differed somewhat: language, education and type of interviewer all had effects significant at the 90% level (in fact at the 95% level). These codes were used in 45% of the interviews, less often with French interviews (32%), less often with respondents with at least some post-secondary education (40%), and less often by the least experienced interviewers (22%).

When this analysis was repeated for the use of the codes for interruption the results differed again: this time, it was language, age and type of interviewer that had effects significant at the 90% level. This code was used in 55% of the interviews, more often with French interviews (73%), less often with the youngest respondents (18-24) (48%), and less often by the least experienced interviewers (21%). (The age effect was only there in the french sample.)

The differences between the results for the two interview languages could come from a variety of sources. The french interviews were all conducted after the english ones so that the interviewers had had more time to become familiar with the questionnaire and the subject matter, and with the use of the CATI system. In addition, no matter how careful the attention to the correspondence between the French and English versions of the questionnaire, they could not have all of the same connotations. Finally there remains the possibility of cultural differences in the reaction to interviews.

## 6 Assessing the Method

The method did prove to be easy to apply. Team members monitoring the interviews noticed that when interviewers captured one of these codes there was a perceptible delay of 2-3 seconds. This added perhaps 10 seconds on average to the length of the interviews, which averaged over 30 minutes. It is estimated that the behaviour coding added less than .5% to the cost of interviewing for the pretest. The coding added

to the interviewers' workload but during debriefings they did not indicate that this caused any problems.

## 7 Conclusions

The coding of respondent behaviour performed during this pretest proved to be inexpensive, easy for the interviewers to apply and was useful during the redesign of the questionnaire. The analysis of behaviour codes was only one of the sources of information used during the redesign; insight gained during the monitoring of interviews by the questionnaire development team and verbal feedback received from the interviewers informally and during formal debriefing sessions was essential. In general, while the analysis of the behavioural coding did not point to any problems that had not been noted by the other more subjective methods, it did provide some objective indication of the extent of the problems. This more objective evidence was extremely useful in directing the team away from problems whose perceived importance was not supported by the behaviour codes and so towards the most serious problems with the questionnaire. The importance of the other methods of gaining insight was in trying to identify the nature of the problems once the behaviour codes had been used to identify their location.

The two primary inadequacies of this methodology as implemented for this pretest for analyzing interview behaviour are the lack of data concerning interviewer behaviour and the lack of data concerning the adequacy of the respondents' answers. A behaviour code for "inadequate answer, probing required" could be easily added to the set used in this study. It was not used here because of the concern that interviewers would have more difficulty using it than the others, but since the interviewers experienced little difficulty with the set of codes used, such a code will be tested on future applications of the respondent behaviour coding methodology at Statistics Canada. This methodology only collects information concerning respondent behaviour, so supplementary methods must be used if information about interviewer behaviour is also needed. Careful monitoring of interviews is essential to identifying the nature of the problems located by respondent behaviour coding and can also provide information on interviewer behaviour.

Our general conclusion after using this method is that it provides, at minimal cost, objective information to supplement the more subjective information derived from monitoring interviews and debriefing

interviewers. We intend to use it when possible during future pretests and to attempt to refine the set of behaviours coded.

## 8 References

- Converse, J. M. and Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Quantitative Applications in the Social Sciences No. 63. Beverley Hills, CA: Sage Publications.
- DeMaio, T. (ed.) (1983). *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington.
- Dijkstra, N. and Van der Zouwen, J. (eds) (1982). *Response Behaviour in the Survey Interview*. Academic Press, London.
- Oksenberg, L., Cannell, C., and Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics*, 7, 349-365.
- Sykes, W. and Morton-Williams, J. (1987). Evaluating survey questions. *Journal of Official Statistics*, 3, 191-207.