

Behaviour coding, expert panel and interviewer debriefing: the evaluation of the methods on the basis of EU-SILC questionnaire testing results

Marjaana Järvensivu
Statistics Finland

1. Introduction

The EU-SILC (Statistics on Income and Living Conditions) was carried out in Finland for the first time in 2004. The sample unit was a household and interviews were made by telephone. Data collection was integrated into that of the national Income Distribution Survey. Both the EU-SILC questions and questions from the national Income Distribution Survey were included in the questionnaire. A quality report will be made on the data collection, of which one part is laboratory testing. Consequently, a questionnaire testing made by Statistics Finland's SurveyLaboratory was connected to the data collection, aiming to find out the functioning of the EU-SILC questionnaire and the quality of the data collected.

Three different methods were selected for the testing: questionnaire appraisal made by an expert panel, behaviour coding made from interview tapes and written interviewer debriefing. The testing was made at the same time as the actual data collection, from January to May 2004. The questionnaire appraisal was made right at the beginning of January, the behaviour coding from February onwards as the field interviewers were recording the interviews, and the deadline for the interviewer debriefing was the end of May when the fieldwork ended.

The testing enabled evaluation of the quality of the collected data in addition to the assessment of the methods used by the laboratory. Because we used behaviour coding in the testing for the first time, it was interesting to examine what kinds of results it produced compared with methods more familiar to SurveyLaboratory.

In order to examine what types of problems we can find out with the methods, their results should be somehow moulded into the same form. One way of condensing the testing results is to group them according to the factor causing the problem. Such analysis mode has been used previously in methodological comparisons by Presser and Blair (1994), Rothgeb, Willis and Forsyth (2001) and Forsyth, Rothgeb and Willis (2004). The framework used by Forsyth et al. is appended to this paper (see Appendix 1). I applied a similar analysis mode to the SurveyLaboratory methods.

2. Implementation of the testing

2.1. Behaviour coding

In testing the EU-SILC questionnaire, the SurveyLaboratory employed the method of behaviour coding for first time. Nine field interviewers recorded a total of 41 interviews during the data collection in spring 2004. On the basis of the recorded tapes, the speech acts of the interviewer and respondent in the interview situation were coded and the results have been presented in a separate testing report (Kallio 2004a). The testing report is based on coding made by one person. The interviews have already been coded also by another person in order to verify coding reliability but the verification will be done later.

2.2. Expert panel

Four researchers familiar with using the questionnaire appraisal system took part in the questionnaire appraisal of the expert panel. The panel examined the questionnaire from the viewpoint of an imaginary household. The members of the panel took notes on the questions they found difficult and they were discussed in the group meeting. Based on the discussion I summarised the results into a testing report (Lehtinen 2004).

2.3. Written interviewer debriefing

Interviewer debriefing was gathered from 20 field interviewers in writing. It was asked in the questionnaire to provide question-specific feedback on the problems arising in the interviews. The results given in the debriefing form have been collected together and reported in a separate testing report (Kallio 2004b).

Due to the high number of questions the whole questionnaire could not be reviewed in the expert panel's questionnaire appraisal, but some of the questions had to be excluded from the testing. Instead, almost the entire questionnaire was handled in the behaviour coding and interviewer debriefing. In this methodological comparison the data examined are confined to those parts of questionnaire and questions that have been examined with all the methods.

For the methodological comparison all three testing reports were inspected question by question. I coded the reported problems using the classification coding scheme of Rothgeb, Willis and Forsyth (2001) and Forsyth, Rothgeb and Willis (2004). There were some problems that seemed to accumulate on the category *other* so I added to the classification a few categories I considered necessary. I named them: problems with proxies, the place of the question on the questionnaire, the question unnecessary for some respondent group. In other respects the classification follows the four-phase model of the question-answering process. I made the coding on the basis of the testing reports from question-specific problem descriptions. For each testing report questions were given a code according to the problem it involved. An individual question got as many codes as the problems found in it. After studying and coding each testing report codes were combined into one data. I will next examine the results of the methodological comparison.

3. Results

The comparison comprised a total of 175 questions from the EU-SILC questionnaire. In all, 515 problems were coded with different methods for the questionnaire. In the expert panel 157 problems were found, in the behaviour coding 173 and on the basis of the interviewer debriefing 185 problems. The method-specific averages were: the expert panel 0.9 problems per question, the behaviour coding 1.0 and the interviewer debriefing 1.1. All methods were found to involve a high number of problems, but the interviewer debriefing was the most effective in that respect.

The number of questions assessed unproblematic with all methods was 17 (9.7%) and such questions that were appraised as problematic with all methods numbered 83 (47.4%). Thus nearly one half of the questions were identified as problematic with all the testing methods.

3.1. What kinds of problems were detected?

It is more interesting to examine the types of problems identified with different methods than the number of problems. For individual codes the results accumulated on four codes (*difficult*

for interviewer to administer, undefined/vague term, high detail required or information unavailable, problems with proxy answering). In all, 53.4 per cent of all problems were recorded on these codes.

In addition to the frequencies describing the use of individual codes I studied the problem types on a less detailed level of the classification, which describes the phases of the question-answering process (*comprehension and communication, retrieve from memory, judgement and evaluation, response selection and other*).

Examined by the method, Table 1 shows that clearly the most problems belonging to the category *comprehension and communication* were identified with each method. In the behaviour coding the next highest number of problems were placed in the *retrieve from memory* category. In the expert panel and interviewer debriefing the second highest number of problems concerned the category *other*, but percentages in the category *retrieve from memory* were also high. The high percentages of the category *other* in all methods are explained by the large number of problems with proxy answering found in all the methods. Problems of the *judgement and evaluation* and *response selection* type were present considerably less compared with other sections. In the main, the results are analogous with the study of Rothgeb et al. (2001).

Table 1. Distribution of problems (in %) according to the question-answering process division in different methods

	Expert panel	Behaviour coding	Interviewer debriefing
Comprehension and communication	52.2	49.1	49.7
Retrieve from memory	10.2	23.7	16.8
Judgement and evaluation	6.4	6.9	7.6
Response selection	2.5	4.0	5.4
Other	28.7	16.2	20.5
Total	100	100	100

3.2. Uniformity of the results between different methods

The problems in the questions were classified after the coding so that an unproblematic question received the value 0, a question with one problem the value 1 and a question with more than one problem got the code 2.

In the examination of these “problem indicators” it was found that the problematicity of the questions was not quite unanimous between the methods (Table 2). For example, 34.9 per cent of the questions were completely unproblematic in the expert panel, 27.4 per cent in the behaviour coding and 24.6 per cent of the questions in the interviewer debriefing. Thus, the expert panel found less problematic questions than the other methods. The interviewer debriefing, on the other hand, revealed the most questions with several problems (25.1 per cent against 21.1 and 21.7 percent).

Table 2. Problematicity of questions with different methods (in %)

	Expert panel	Behaviour coding	Interviewer debriefing
No problems	34.9	27.4	24.6
Some problem	44.0	50.9	50.3
Several problems	21.1	21.7	25.1
Total	100	100	100

In addition to the above examination of percentual distributions, I used cross tables to find out to what extent the problematicity assessments of the questions matched between the methods. In other words, I examined if the different methods assessed the same questions as having no problems, some problems and several problems.

The results of the behaviour coding confirmed the results of the expert panel to some extent (Table 3). The small frequencies of the left bottom cell and the right top cell ($n=4$ and $n=8$) of the table indicate the uniformity of the results. Questions identified as unproblematic with one method and very problematic with another were few. However, the frequencies of the questions found to have some problem with one method and no problems with another method were considerably higher.

The results of the behaviour coding and interviewer debriefing are somewhat more clearly analogous (Table 4). Consensus was highest in the questions where some problem had been identified with both methods ($n=53$). In cases where several problems had been found in the question with one method and none with another method the frequencies were low ($n=3$ and $n=4$).

In the results of the expert panel and interviewer debriefing there were more questions assessed as unproblematic with both methods than between the other methods (Table 5). The uniformity of the results is also visible as low frequencies in the same conflicting places as in the cross tables discussed above ($n=6$ and $n=4$). In contrast, there is more dispersion for the questions considered problematic with both methods. In addition, there are many questions that were regarded unproblematic by the expert panel but which according to the interviewer debriefing had some problem.

Table 3. Comparison of problematicity between behaviour coding and expert panel (frequency of questions)

	Behaviour coding			Total <i>n</i>
	No problems	Some problem	Several problems	
Expert panel				
No problems	23	30	8	61
Some problem	21	41	15	77
Several problems	4	18	15	37
Total <i>n</i>	48	89	38	175

Table 4. Comparison of problematicity between behaviour coding and interviewer debriefing (frequency of questions)

Interviewer debriefing	Behaviour coding			Total <i>n</i>
	No problems	Some problem	Several problems	
No problems	26	14	3	43
Some problem	18	53	17	88
Several problems	4	22	18	44
Total <i>n</i>	48	89	38	175

Table 5. Comparison of problematicity between expert panel and interviewer debriefing (frequency of questions)

Interviewer debriefing	Expert panel			Total <i>n</i>
	No problems	Some problem	Several problems	
No problems	28	11	4	43
Some problem	27	40	21	88
Several problems	6	26	12	44
Total <i>n</i>	61	77	37	175

4. Discussion

4.1. The model of the question-answering process as the framework of the problems identified

The classification of the problems was based on the model of the question-answering process, but the classification also included a specific place for coding the problems outside the question-answering model.

Most of the problems identified with the method can be placed in the classification according to the question-answering model. Clearly the most of the problems found with all the methods were the type of *comprehension and communication* (52.2 per cent in the expert panel, 49.1 per cent in the behaviour coding and 49.7 per cent in the interviewer debriefing).

In the classification the codes describing interviewers' problems were placed under the heading *communication and comprehension*. The conventional cognitive model of the question-answering process does not include the interviewer's role as the presenter of questions, but only focuses on describing the respondent's processes. In my view this has been a great deficiency earlier, so the solution of Forsyth et al. (2001) seemed to be a natural and workable extension to this one-sided model.

The inadequacy of the question-answering process model in taking account of all the problems in the questionnaire is illustrated by the high number of problems recorded under *other*. Of all identified problems 21.6 per cent had been entered in this category (in the expert panel 28.7 per cent, in the behaviour coding 16.2 per cent and in the interviewer debriefing 20.6 per cent).

Problems were especially placed under *problems with proxy answering*. The EU-SILC survey is interested in the situation of the household and therefore in the interview very much information is asked from one household member about the other household members. It would be an ideal situation if all household members were present to answer to the questions concerning themselves, but very seldom the situation is like that. In addition, data collection as a telephone interview lowers the possibility for changing of the respondent in the middle of the interview, even if other household members were available during the interview. In the expert panel the problems with proxy answering were already anticipated (10.2 per cent of the expert panel problems) and the behaviour coding and interviewer debriefing further supported this assumption (11 and 17.3 per cent).

4.2. All three methods have their own pros and cons

In testing the EU-SILC questionnaire, Statistics Finland's SurveyLaboratory used the behaviour coding method for the first time. Behaviour coding was laborious to implement, but the results were promising. It proved very efficient in detecting problems related to memory retrieval and interviewers' problems.

The most problems were found by interviewer debriefing, but all of the methods were efficient in this respect. Clearly most of the problems that were found related to the first phase of the question-answering process (*communication and comprehension*). All of the methods were efficient in detecting these kinds of problems. However, the methods were different from each other in that they identified different questions problematic. It seems that behaviour coding and interviewer debriefing gave the most uniform results.

There were lots of questions with high cognitive burden for the respondent in the EU-SILC questionnaire. Therefore it was interesting to examine how the different testing methods could detect problems related to memory retrieval. Based on the results it seems that the expert panel is not very efficient in detecting these kinds of problems. Behaviour coding seems to perform better than the expert panel, and interviewer debriefing slightly better than behaviour coding.

The comparison of the methods in this study is based on descriptions of problematic questions in three independent testing reports. The descriptions are summaries made by the report writers and are therefore interpretations of the original testing data. Additionally, the coding processes in the reports were performed by only one person.

In assessing uniformity between the methods we need to remember that the actual problem behind a code may be different with different methods even if a question has been coded with the same "problem indicator". In order to study the uniformity of the actual problems, we would need to approach the testing data in some other way than in this study.

Based on this evaluation we can say that a certain level of uniformity does exist between the different methods used by Statistics Finland's SurveyLaboratory, but each of the methods has its own pros and cons. In other words, the methods cannot fully replace each other and it is always beneficial to apply more than one method simultaneously.

References

- Forsyth, B., Rothgeb, J.M. & Willis, G.B. 2004. Does Pretesting Make a Difference? An Experimental Test. In Presser, Rothgeb, Couper, Lessler, Martin, Martin & Singer (ed.) *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons, 525-546.
- Kallio, M. 2004a. Tulo- ja elinolotutkimus 2003 - käyttäytymiskoodauksen tuloksia. (Income and Living Conditions Survey 2003 - Results of behaviour coding; in Finnish only). *Unpublished testing report*. Statistics Finland.
- Kallio, M. 2004b. Haastattelijapalaute Tulo- ja elinolotutkimuksen 2003 lomakkeesta. (Interviewer debriefing on the Income and Living Conditions Survey 2003 questionnaire; in Finnish only) *Unpublished testing report*. Statistics Finland.
- Lehtinen, M. 2004. Lomakearviointi Tulo- ja elinolotutkimuksen lomakkeesta. (Questionnaire appraisal on the Income and Living Conditions Survey questionnaire; in Finnish only). *Unpublished testing report*. Statistics Finland.
- Presser, S. & Blair, J. 1994. Survey Pretesting: Do Different Methods Produce Different Results? In P. Marsden (ed.) *Sociological Methodology*, volume 24. The American Sociological Association, Washington DC. 73-104.
- Rothgeb, J., Willis, G. & Forsyth, B. 2001. Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9, 2001.

Appendix 1: Classification Coding Scheme (modified from Forsyth, Rothgeb & Willis, 2004)

COMPREHENSION AND COMMUNICATION

Interviewer Difficulties

1. Inaccurate instructions (move to wrong place; skip error)
2. Complicated instructions
3. Difficult for interviewer to administer

Question Content

4. Vague topic/unclear Q
5. Complex topic
6. Topic carried over from earlier question
7. Undefined term(s)/vague term

Question Structure

8. Transition needed
9. Unclear respondent instruction
10. Question too long
11. Complex or awkward syntax
12. Erroneous assumption
13. Several questions

Reference Period

14. Reference period carried over from earlier question
15. Undefined reference period
16. Unanchored or rolling reference period

RETRIEVE FROM MEMORY

17. Shortage of memory cues
18. High detail required or information unavailable
19. Long recall period or long reference period

JUDGEMENT AND EVALUATION

20. Complex estimation, difficult mental arithmetic required, (Guessing or heuristic estimation may be likely)
21. Potentially sensitive or desirable bias

RESPONSE SELECTION

Response Terminology

22. Undefined term(s)
23. Vague term(s)

Response Units

24. Responses use wrong or mismatching units
25. Unclear to R what response options are

Response Structure

26. Overlapping categories
27. Missing response categories

OTHER

28. Problems with proxy answering
29. Question order
30. Question not applicable to some respondent group
31. Something else