MEASURING DATA QUALITY WITH LONGITUDINAL DATA

Mark Kinack, Statistics Canada
16-A, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

Key Words: response error, inconsistency

1. Introduction

Questions in all surveys are subject to some degree of response error, the extent of which will depend on a number of contributing and confounding factors. Problems with concepts, question wording and instructions, and the choice of response categories are clearly important as these are the means by which information is communicated between the response source and survey taker. Excessive cognitive burdens on respondents of comprehension, recall, judgment and interpretation can also impact on the quality of information collected. Potential problems also exist with the sources communicating the information; for example, interviewer and proxy effects may be present.

In the case of longitudinal surveys, existing longitudinal data can often be used to quantitatively examine and measure the impact of some of these factors. This paper presents some empirical results on data quality measures obtained from Canadian Labour Force Survey (LFS) longitudinal data. Responses to selected questions on the current LFS questionnaire for selected classes of respondents are examined. Some comparisons are made with results obtained from re-interview data.

In section 2 the LFS is described briefly. Section 3 discusses the longitudinal data available from the LFS. Sections 4 and 5 present two examples of how LFS longitudinal data can be examined. Section 6 contains some concluding remarks.

2. The Canadian Labour Force Survey (LFS)

The LFS is a monthly survey of approximately 62,000 households from which the official measures of employment and unemployment in Canada are obtained. The rotating panel sample design used has selected dwellings remaining in the sample for six consecutive months. Within selected dwellings, all non-institutionalized civilian household members aged 15 and over are asked about their labour force activity during a seven day reference week. Interviews are performed either in-person or by telephone during the one week survey period following the reference week. Proxy reporting often occurs in the survey, since any household member aged 15 or over may provide responses for all other household members.

Until its suspension in March, 1991, the LFS conducted a monthly re-interview program. Each month, except in April and December, a sub-sample of dwellings in the LFS sample was selected, with households re-interviewed by telephone by senior interviewers during the week following survey week. (Questions were modified to allow for the extra week since reference week.) The non-reconciled portion of the re-interview sample was intended as an independent replication of the interview, with data obtained used to estimate response variance. In the reconciled portion, the senior interviewers also conducted independent replications of interviews, but then attempted to obtained true responses by reconciling differences between the two interviews, with the assistance of respondents. The reconciled data obtained were used to estimate response bias.

3. Examining LFS Longitudinal Data

The readily available and accessible data for respondents appearing in the LFS for up to six months makes it possible to easily construct a longitudinal database. The longitudinal data provides information not available in cross-sectional data (i.e. a six month history of responses), that can be examined for evidence of response errors, and hence used to monitor and test certain data quality concerns.

LFS data from two longitudinal periods is examined in this study. The data come from respondents who rotated into the survey in either April 1986 or April 1988. Thus, the respondents were in the LFS between the months of April and September, in 1986 and 1988 respectively. Note that it is possible that respondents may not appear in the survey in all six months. This would occur if, in one or more of the six months, they are nonrespondents for any one of a number of reasons. Only unweighted longitudinal data for respondents with uncorrected original data for all months they are in the survey is considered. The reason for not allowing any corrected or imputed data is to attempt to avoid any longitudinal inconsistencies

that could be introduced during editing and imputation steps.

In order to illustrate the potential use of longitudinal data for examining different types of data quality concerns, examples from two classes of items on the LFS questionnaire will be presented. Subsets of respondents from the two time periods described above are selectively chosen for in-depth analysis.

The first class of items consists of questions for which logical inconsistencies in responses can be identified. For example, questions on durations of events such as the length of time an individual has been looking for work or long-term absences from work would fall into this category. Responses can be compared over time and checked for logical consistency. Inconsistent reporting would indicate the existence of response errors somewhere in the history of the responses.

The second class of items consists of questions requiring coded responses, in which the interviewer bears the task of interpreting responses and selecting appropriate codes. Respondents are not aware of the possible response categories, so that this burden on the interviewer of translating responses to a suitable set is also subject to error. Items included here are reasons for activities such as working part-time or absences from work. Longitudinal data can be used to look at code changes within particular respondents' response histories for recurring code changes that could indicate confusion or ambiguities with response categories.

4. Example 1: Job Search Activity

Information on job search activity is used in determining an individual's labour force status. For those who are not employed, it is one of the criteria applied in distinguishing between the two categories unemployed and not in the labour force. As well, this information is used in estimating duration of unemployment. It is of interest to examine whether the month to month reporting for the items associated with job search activity is consistent.

Three items on the LFS questionnaire pertaining to job search activity (Q56, Q57 and Q58) apply to respondents who fall into one of three categories:

a) had a job but not at work because of a layoff
b) had a new job to start at a definite date in the future
c) did not work nor had a job at which they did not work (excluding permanently unable to work)

Such respondents are asked Q56: "In the past 6 months, has ... looked for work?". If the response is 'Yes', Q57 then follows: "In the past 4 weeks, what has ... done to find work?". Although respondents are asked to list specific job search methods used, responses to this item can be collapsed into the dichotomous pair 'Yes' (for those who have done something to find work) and 'No' (for those who have not). Respondents who have replied that they have done something to find work are asked Q58: "Up to the end of last week, how many weeks has ... been looking for work?".

Thus, depending on the situation a respondent may be asked Q56 (and perhaps Q57 and perhaps Q58) in some months and not in others.

The specific wording of these items makes it possible to detect logical inconsistencies for certain classes of respondents with responses to these items in more than one month. In particular, it is possible to construct a hierarchy of longitudinal consistency checks.

Before doing so, however, note the following about response sequences with transitions in responses for Q56. A 'No' followed by a 'Yes' logically implies job search began since the start of the respondent's history in the survey. In particular, a 'No' followed by a 'Yes' in consecutive months logically implies job search began in the past month. On the other hand, a 'Yes' followed by a 'No' logically implies job search occurred prior to the respondent's history in the survey (i.e. more than 6 months ago).

With this in mind, the hierarchy of longitudinal consistency checks is as follows:

1) check for inconsistent response pattern for Q56

   - a response pattern for Q56 with a response of 'No' followed later by a 'Yes' and then a 'No' again is logically inconsistent

2) consistent response patterns for Q56 can be checked against responses to Q57 and Q58

   i) a transition from 'Yes' to 'No' in Q56 is logically inconsistent if any of the previous months when Q56 is 'Yes' also has a response of 'Yes' for Q57

   ii) a transition from 'No' to 'Yes' in Q56 in consecutive months is logically inconsistent if either:
   a) the latter month has a response of 'No' for Q57, or

b) the latter month has a response of 'Yes' for Q57, but the number of weeks reported in Q58 is greater than the number of weeks between the two surveys

iii) a response pattern for Q56 with a response of 'Yes' followed later by a 'No' and then a 'Yes' again is logically inconsistent if either the 'Yes' to 'No' transition fails the check in i), or the 'No' to 'Yes' transition fails the check in ii)

Respondents who do not fail any of the longitudinal consistency checks above require further examination of their month to month reporting for Q58 in order to detect additional inconsistencies.

Figure 1 displays the results from the longitudinal consistency checks for items Q56, Q57 and Q58 for the longitudinal data from the period April-September 1988. The 3619 respondents correspond to individuals who appeared in the survey in all six months, and had at least two responses to Q56. Clearly, at least two responses to Q56 are needed for the possibility of longitudinal inconsistencies to exist. Note that respondents with exactly two responses to Q56 cannot fail the first consistency check on Q56 alone, but could fail either of the consistency checks on transitions in Q56. Thus, the inconsistency rates observed for Q56 alone could be considered conservative estimates.

As well, 2366 respondents who were not employed in any of the six months and responded 'No' to Q56 in all six months have been separated out. They have been excluded when calculating the inconsistency rates summarized in Table 1. The reason for this exclusion is because it is quite probable that most of these respondents are likely not subject to response errors in Q56 (e.g. those with no attachment to the labour force, such as homemakers, students, the retired, etc.). Since such respondents constitute a large proportion of the number of respondents with at least two responses to Q56, their inclusion would significantly deflate the inconsistency rates observed. This would detract from the important findings obtained with respect to those more likely subject to difficulty with these items.

Figure 1 shows that 1253 respondents had at least two responses to Q56, but not six responses of 'No'. Of these, 171, or 14%, had an inconsistent response pattern for Q56.

Furthermore, 187 (73+25+51+38), or 15%, had inconsistencies associated with transitions in Q56. This gives a total of 358 longitudinally inconsistent records (29%).

A summary of the breakdown by response source (six months non-proxy, six months same proxy and a residual category for other mixed proxy cases) is included in Table 1, which also contains results for longitudinal data from the period April-September 1986.

The results appear to indicate potential concerns regarding the quality of the data for items Q56, Q57 and Q58 for respondents reporting changes in their job search activity. Even a question as seemingly straightforward as Q56 shows surprisingly high reporting inconsistency. Results from the two time periods are very similar. Reasons for the observed results for these items clearly needs to be further investigated.

5. Example 2: Reason for Working Part-time

Respondents identified as having had a job in reference week (whether at work or temporarily absent) are asked a question about their usual hours of work: "How many hours per week does ... usually work at his/her: (Main) Job? Other jobs?". Those whose usual hours total less than 30 are asked: "What is the reason ... usually works less than 30 hours per week?".

Codes: 1 Own illness or disability
2 Personal or family responsibilities
3 Going to school
4 Could only find part-time work
5 Did not want full-time work
6 Full-time work under 30 hours per week
0 Other - Specify in NOTES

Respondents do not see the list of categories associated with the above codes, but rather the interviewer must interpret the reason given, perhaps with further probing, to find the appropriate code. It is an important distinction that this question is intended to determine why the respondent works less than 30 hours per week, and not why the job provides less than 30 hours of work.

Respondents giving any reason which results in a code other than code 6 (which applies to special situations) are classified as part-time workers.

Data from this question is used to measure visible under-employment, by identifying involuntary part-time workers; that is, individuals who desire, but cannot find, full-time work. Hence, it is of interest to examine whether any problems exist with identifying this special group.

In particular, one can focus on codes 4 and 5, which reflect a respondent's preference for full-time or part-time work, given there are no additional constraints limiting the amount of time that can be devoted to work, such as the types of conditions covered by codes 1-3. Code 4 is intended for situations where a respondent would prefer to work 30 or more hours per week, while code 5 reflects a respondent's preference for working less than 30 hours.

Consider those respondents who appear in the survey for all six months and have the same single part-time job in all six months as well. Intuitively one would not expect frequent changes in reason for working less than 30 hours per week, as defined by the seven categories associated with the codes above. A notable exception to this rule would be certain types of combinations of code 3 (going to school) with other codes.

Figure 2.1 contains four tables with results for the pooled data from the two longitudinal periods April-September, 1986 and April-September, 1988.

The first table classifies these respondents by the number of changes in reason for working less than 30 hours per week, over the six months. The results show 55% of respondents reporting at least one change in reason, and 40% reporting at least two changes in reason.

Two other tables display the frequencies of single code occurrences, and the rates of occurrence for each code in response patterns. Code 5 is by far the most common code. Non-proxy response patterns have lower rates of occurrence of code 4 than response patterns involving proxy reporting. Same proxy response patterns tend to hardly ever have code 2.

In each of these three tables, the breakdown by response source suggests the existence of a potentially important proxy effect.

Finally, a table is included that shows the observed combinations of codes that occur in the 381 response patterns with 2 or more codes. Interior entries correspond to the number of respondents with at least one occurrence of each of the row and column codes in their six month response pattern. The most common mixtures of interest are codes 2 and 5, and codes 4 and 5, with 37% (142/381) having at least one code 4 and one code 5.

The data pertaining to codes 4 and 5 is summarized in Table 2. The results show that only 8% of the 689 respondents indicate that in all six months their reason for working less than 30 hours was because they 'could only find part-time work'. A further 32% of the respondents give this reason in some of the six months. The corresponding numbers for the reason 'did not want full-time work' are 26% and 43%. Furthermore, 21% of the respondents had these two reasons at least once each. These numbers suggest possible concern over the proper coding of responses for this question.

Re-interview data can be used as another source to further investigate this issue. Figure 2.2 contains both non-reconciled and reconciled re-interview data from the period January 1986-December 1988, for the same question. The data is cross-classified by coded response in the interview and re-interview. As with the longitudinal data, it can be seen from the non-reconciled data that some problems may exist in distinguishing between codes, particularly between codes 2 and 5, and codes 4 and 5. The reconciled re-interview data does not suggest extensive biases.

6. Concluding Remarks

As seen by results presented in this paper, longitudinal data can sometimes provide useful and important information about data quality not attainable via other vehicles, such as qualitative cognitive research techniques and re-interview analysis. More extensive use of longitudinal data, perhaps in conjunction with re-interview data, could prove to be a valuable commodity in monitoring and testing data quality concerns. Focused studies on specific issues, such as questionnaire revision during survey redesign periods, could use information from quantitative results obtained from actual respondents participating in the real survey. The results from studying longitudinal data can help to detect potential trouble spots and highlight areas to be targeted for further study.

Acknowledgements

I would like to thank Maryanne Webber for valuable guidance and insights associated with this research, and Georges Lemaître for useful comments on an earlier version of this paper.

Reference

Statistics Canada (1990). Methodology of the Canadian Labour Force Survey, Catalogue 71-526.

Figure 1

Longitudinal consistency checks for items Q56, Q57 and Q58 for April 1988 rotate-ins

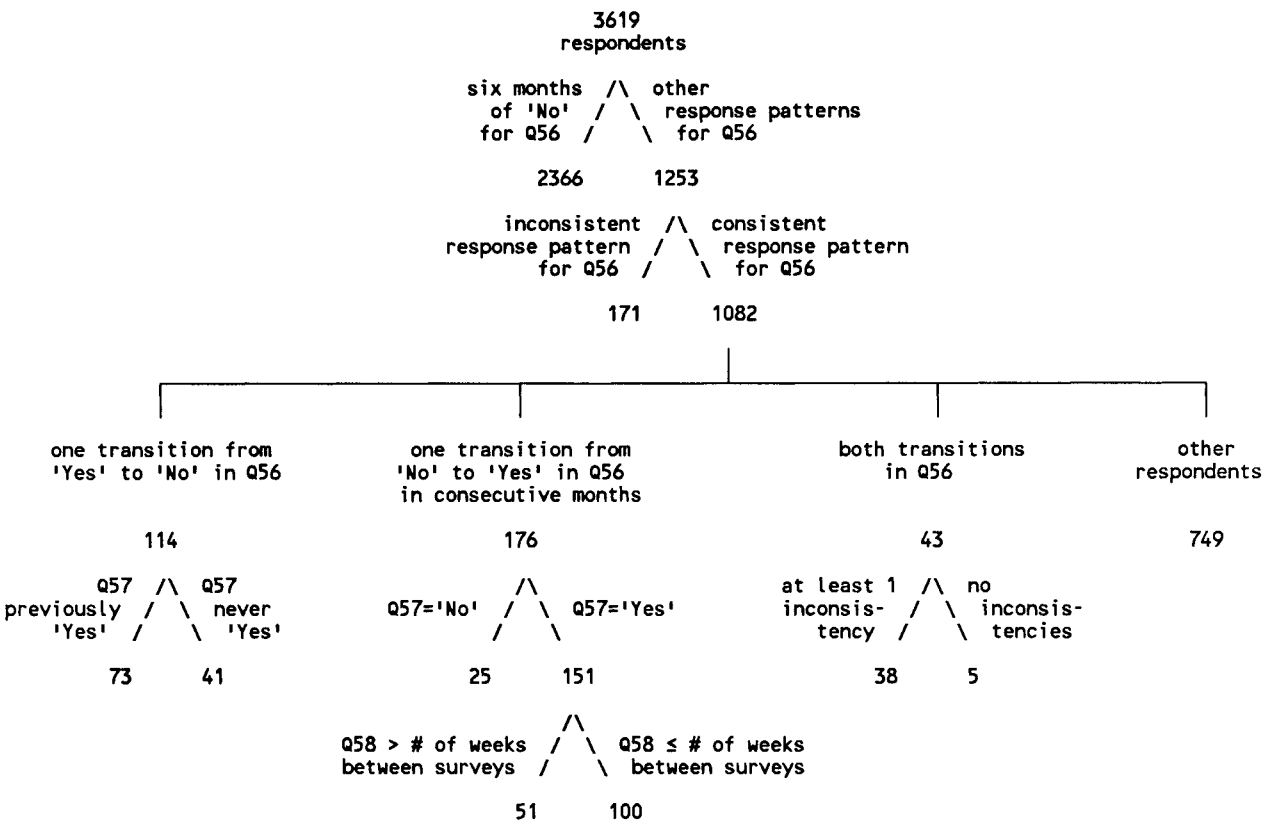

Table 1

Summary of longitudinal consistency checks for items Q56, Q57 and Q58

	respondents	with inconsistent response pattern for Q56	with inconsistencies in transitions in Q56	total inconsistent response patterns
April 1988 rotate-ins				
total	1253	171 (14 %)	187 (15 %)	358 (29 %)
non-proxy	251	25 (10 %)	31 (12 %)	56 (22 %)
same proxy	157	20 (13 %)	26 (17 %)	46 (29 %)
other	845	126 (15 %)	130 (15 %)	256 (30 %)
April 1986 rotate-ins				
total	1340	163 (12 %)	187 (14 %)	350 (26 %)
non-proxy	271	37 (14 %)	37 (14 %)	74 (27 %)
same proxy	193	18 (9 %)	30 (16 %)	48 (25 %)
other	876	108 (12 %)	120 (14 %)	228 (26 %)

Figure 2.1

Pooled April 1986 and April 1988 rotate-ins with the same single part-time job in all six months

Changes in reason for working less than 30 hours

Changes in reason	total	non-proxy	same proxy	other
0	308	121	10	177
1	104	41	6	57
2+	277	68	39	170
	---	---	--	---
	689	230	55	404

Same code in all six months

Code	total	non-proxy	same proxy	other
1	3	0	1	2
2	56	24	0	32
3	3	1	0	2
4	53	17	5	31
5	178	75	4	99
6	14	4	0	10
0	1	0	0	1
	---	---	--	---
	308	121	10	177

At least one month of code

Code	total	non-proxy	same proxy	other
1	12	5	1	6
2	172	83	3	86
3	136	5	36	95
4	274	61	42	171
5	475	176	31	268
6	86	32	2	52
0	11	2	1	8

At least one month of each of row and column codes

	Code					
Code	2	3	4	5	6	0
1	6	0	0	9	5	1
2		4	22	106	18	2
3			97	74	3	2
4				142	24	4
5					54	6
6						2

Table 2

Pooled April 1986 and April 1988 rotate-ins with the same single part-time job in all six months

total respondents	689	
six months of code 4	53	(8%)
mixtures involving code 4	221	(32%)
six months of code 5	178	(26%)
mixtures involving code 5	297	(43%)
mixtures of codes 4 and 5	142	(21%)

Figure 2.2

Pooled January 1986 - December 1988 re-interview data for the question "What is the reason ... usually works less than 30 hours per week?"

Non-reconciled data

interview code	re-interview code 1	2	3	4	5	6	0
1	16	0	1	0	1	1	0
2	0	81	0	12	22	2	0
3	0	2	309	1	2	1	0
4	0	16	4	253	53	16	1
5	1	37	5	44	270	12	2
6	1	5	0	6	9	40	0
0	0	0	1	0	0	1	1

Reconciled data

interview code	re-interview code 1	2	3	4	5	6	0
1	20	0	0	1	1	0	0
2	0	153	0	3	12	3	0
3	0	0	398	4	4	0	0
4	1	7	4	444	27	1	0
5	4	19	5	17	423	3	0
6	1	4	0	11	11	92	1
0	0	1	0	4	2	1	2