AN INDUSTRY STUDY OF CODER VARIABILITY

Martin Collins, SCPR Survey Methods Centre & City University Business School

1. INTRODUCTION

In this paper I will describe an experiment conducted recently in the U.K. as part of a planned programme of research into coding - the process of converting verbatim comments into analysable (usually numeric) code. Within the programme, the experiment represents only an initial step. It is important, however, in demonstrating the willingness of a number of research agencies to join together to conduct (and finance) basic research into their methodology.

Background to the Experiment

Survey researchers have long been aware that sampling variability is only the best known of many potential sources of error in their data. In particular, they have concerned themselves with the influence of the interviewer on the quality of survey data. As a result, quite elaborate systems have developed to control recruitment, training and work quality in interviewing. Other aspects of survey methods have received less attention.

One comparatively neglected area is that of coding, where a respondent's verbatim comments are summarised in an analysable form. This is a common task, especially in ad hoc survey research. The researcher may be unwilling to 'close' a question by specifying the available response categories because he seeks spontaneity (eg. in a question of awareness) or a measure of salience (in a question of motivation) or - more often - because he feels unable to forecast the categories that will be needed. Then, an 'open' question will be used (eg. 'How did you first come to see this periodical?' 'What did you like about this product?'). The respondent is not guided as to the dimension or the units in which a response should be given. The interviewer is instructed to record the response verbatim, usually after probing for depth and/or to remove ambiguities. The list of eligible response categories (the coding frame) is drawn up after the event, by reference to the answers received, and verbatim comments are coded to indicate the category or categories into which they fall.

Again, attention has tended to focus on the way in which interviewers handle such open questions: do they probe uniformly, do they avoid prompting the respondent into certain clear response categories, do they provide a faithful record of what was said? The coding process has been regarded as a mechanical stage, approached fairly casually and subject to quality controls that may well not be adequate. Attention was recently drawn to the fallibility of the process by two papers (Kalton & Stowell, 1979; Collins & Kalton, 1980). These illustrated, in the context of social surveys, the extent to which the judgements made by coders in summarising verbatim comments could affect the precision of survey estimates. As a result, the Market Research Society set up a Study Group to investigate the activity of coding. The first requirement was to establish the 'state of the art'. Were the results derived from social surveys equally true of market research? Or were questions in market surveys less complex and hence more amenable to reliable coding?

The Study Design

Seven survey organisations were invited and agreed to take part in an industry study. Five were market research agencies (BJM, BMRB, NOP, PAS and RSGB); the other two were the government social survey organisation (OPCS) and an independent social research institute (SCPR). Each organisation selected three open questions from recent project(s) and, for each question, selected 100 verbatim responses in a haphazard way. These responses were then coded independently by each of six of the organisation's coding staff, using the coding frame prepared for the project concerned. Normal supervisory procedures were not used, the object being to establish the extent to which judgements could vary - the variation that such procedures are designed (rather informally) to control.

2. MAIN RESULTS

The data have been analysed, following the methods of Kalton & Stowell, to investigate four aspects of reliability:-

(i) Overall reliability of a coding frame: a summary measure of the reliability of the frame adopted for each question in capturing the content of the set of responses.

(ii) Reliability of individual codes: an index of agreement between coders in their use of each separate code in the frame, taking into account both haphazard errors and any consistent coder bias.

(iii) Coder bias: for each code, a test for the existence of correlated coder variance - a tendency for some coder(s) to consistently over or under-utilise the code, compared with the rest of the panel.

(iv) Coder agreement: for each question, a summary measure of the extent to which each coder agreed with other members of the same panel in applying the coding frame.

Overall Reliability

Early studies of coder reliability were largely concerned with situations where each response is to be summarised by a single code. Then, agreement between two coders is easy to define. This situation is comparatively rare in market surveys and all 21 experimental questions allowed for multi-coding. That is, a coder was free to use more than one code to indicate the response categories covered by a single verbatim answer. In such cases it can be argued that a measure of reliability that depends on total agreement between two coders in their treatment of a given response is too severe. The measure should take into account the fact that data will normally be analysed one code category at a time: the number of responses involving a given code, regardless of whether or not other codes were also used. Only in the unusual case of analysis being in terms of specific combinations of codes will total agreement be a relevant indicator of precision.

Table 1 Index of overall reliability (K*) at each question, by organisation

% Reliability	Organisation						
	A	B	C	D	E	F	G
Question 1	74	73	80	78	74	66	58
" 2	83	78	74	75	73	62	62
" 3	83	87	73	71	77	64	53
Average	80	79	76	75	75	64	58

Table 2 Comparison with two recent experiments

	% Reliability:	
	Average	Range
Current study	72	53 to 87 (21 items)
Railway noise[1]	72	63 to 80 (6 items)
Industrial tribunals[2]	70	62 to 78 (10 items)

[1]Kalton & Stowell (1979)

[2]Collins & Kalton (1980)

We therefore followed Kalton & Stowell in adopting as a summary measure of the overall reliability of a coding frame an estimate of the reliability of the typical coding. This is a weighted average of the reliability indices of the individual codes making up the frame, the weighting reflecting the incidence of use of each code:-

$$K^* = \frac{\left[\Sigma q_j\ (1-q_j)\ K_j\right]}{\left[\Sigma q_j\ (1-q_j)\right]}$$

- where q_j is the incidence of code j, the proportion of codings to which it was applied;
- K_j is the reliability index of code j, the form of which is given below: a measure of the agreement between two coders in applying the code, less an allowance for the possibility of chance agreement.

The average value of this index of overall reliability across the 21 questions was 72%. For individual questions, the index ranged from 87% down to 53%, a distribution that is shown in Table 1.

The table shows substantial differences between the seven organisations taking part and a good deal of consistency between the three results for each organisation. (Variation between organisations accounts for just over 80% of the total variation in the results.) This pattern reflects variations in the surveys used in the experiment, in terms of subject matter and complexity of responses, ranging from simple tests of consumer acceptance through to surveys of complex social issues. The highest index of 87% occurred for a question about the bad points of a product in a consumer trial, where a third of the respondents answered "Nothing" and most other responses contained obvious keyword pointers to product attributes. The lowest index of 53%, in contrast, was for a question asking why a respondent found a magazine useful (or not useful) in his work. Answers to the question 'Why do you say that?' tend to be enormously varied (not least because they are given in order to explain a range of answers to the previous question) and are often vague.

The important finding from Table 1 is, of course, that unreliability occurs across a range of organisations and questions. Even the "best" set of results includes an index of 74%. As Kalton & Stowell show, this implies a loss of precision equivalent to a 26% cut in the effective sample size, equivalent in turn to a sample design effect of 1.35.

Table 2 shows that the overall average reliability index in this study - 72% - is in line with the social survey results presented by Collins & Kalton. The range of values obtained is rather greater in this study (even allowing for the larger number of questions studied), reflecting the range of different projects covered.

Reliability of individual codes

The index adopted to measure the reliability of an individual code is a measure of the extent of agreement between pairs of coders in their use of that code, adjusted downwards to allow for the possibility of chance agreements:-

$$K_j = \frac{(Q_j - q_j)}{(1-q_j)}$$

- where Q_j is the proportion of occasions on which coder B applied code j, given that coder A did so, averaged over all possible pairs of coders;
- q_j, as before, is the incidence of code j.

Table 3 shows the distribution of the values of this index found for the 200-plus code categories in the experiment which were used in at least 5% of the relevant codings. The individual codes covered virtually the full range of reliability from 100% down to 6%. Only a very small proportion (about 1-in-7) had reliability indices of 90% or higher, but two-thirds had indices that we might consider acceptable - in the absence of supervision or check-coding - of 70% or higher. At the other extreme, 18 codes (8% of the total) were very unreliably applied. These cases were spread over six of the seven organisations and 12 of the 21 questions. Most of them were residual response categories - 10 'Other answer' codes and 5 codes covering 'Unspecified answers'. Only three cases involved more specific response categories.

Table 3 Index of reliability for 221 individual code categories

	Number of codes	%	Cumulative %
Reliability index:			
90% or higher	31	14	14
80-89%	64	29	43
70-79%	47	21	64
60-69%	31	14	78
50-59%	20	9	87
40-49%	10	5	92
30-39%	9	4	96
20-29%	4	2	98
10-19%	3	1	99
Under 10%	2	1	100
Total	221	100	-

Table 4 Comparison between two experiments

	Current Experiment	Industrial Tribunals
Number of code categories	221	121
Codes with reliability of:	%	%
80% or higher	43	30
50-79%	44	56
Under 50%	13	14

Table 5 Significance test for coder bias in 221 code categories

	Number of codes	%	Cumulative %
Probability of Q-value			
p≥5% ('not significant')	135	61	61
5%≥p 1%	32	14	75
1%>p≥0.1%	26	12	87
p<1%	28	13	100
Total	221	100	-

Again, Table 4 shows that the results of the current experiment are broadly similar to those reported by Collins & Kalton for their Industrial Tribunals study.

Coder Bias

The reliability index measures the total error arising in the use of a particular code. Part of this will reflect the haphazard 'as if at random' mistakes made by coders. It may be a significant addition to variability but, provided 'sampling' errors are properly calculated from the resultant data, the effect will be included within total measured imprecision. More worrying is the possibility of coder bias - or correlated coder variance - where some coder(s) in a group tend consistently to over or under-utilise a particular code category. Haphazard errors - simple coder variance - will reduce the precision of survey data and probably attenuate relationships in the data. Bias will threaten the validity of the data and may create spurious relationships.

The existence of correlated coder variance in the current study was tested using Cochran's Q (Cochran, 1950):-

$$Q = \frac{L(L-1)\,\Sigma\,(T_1-\bar{T})}{(L\Sigma u_i-\Sigma u_i{}^2)}$$

- where L is the number of coders;
- T_1 is the number of times coder 1 used the given code, and $\bar{T}$ is the arithmetic mean of the T_1;
- u_i is the number of coders using the given code in coding the i th response.

On this basis, significant coder bias was found for a large proportion of the 221 separate code categories analysed, as shown in Table 5. While most of the code categories with low reliability indices showed significant coder bias, the bias was not confined to those codes. Of the 142 codes with reliability indices of 70% or higher, as many as 34 (24%) showed bias significant at the 5% level. Even apparently acceptable reliability levels may conceal significant correlated variance, indicative of consistent differences between coders in their

Table 6 Coder agreement with other members of the same panel

	Number of coders	%
A_1 (% agreements - average agreements) =		
High (+3 to +5)	10	23
Average (+2 to -2)	23	53
Low (-3 to -4)	8	19
Very low (-9)	2	5
Total	43	100

interpretation of responses and the coding frame used to summarise them.

Examination of the code categories that performed badly at either or both tests points to four broad problem areas. Two of these concern the residual categories mentioned above: 'Other answers' and 'Unspecified answers'. These categories will usually be of little substantive interest to the researcher, who will anyway be striving to minimise their incidence. The other two areas are more important:

- code categories that subsume too many slightly different response areas;
- pairs or sets of categories that attempt to differentiate between different facets of the same basic concept, especially when the same 'keyword' occurs in more than one category.

Unfortunately, for many questions there will be little safe ground between these two problem areas. Some problems may be overcome in data analysis. Thus, a coding frame that errs towards difficult differentiation may be collapsed at the analysis stage into fewer categories. This will improve precision to the extent that coder errors involve judgements within a broader category. Evidence provided by Kalton & Stowell is encouraging: at two experimental questions they found that data produced by coding to a detailed frame followed by collapsing at the analysis stage were more reliable than either the uncollapsed data or data produced by coding directly to the less detailed frame. But these results require verification over a broader range of trials, and, particularly, for multi-coded data.

Coder Agreement

The overall pattern of unreliability could have arisen from occasional errors made by all the coders in a group or from errors made by one or two 'rogue' coders. In assessing the contribution made by each coder, we can calculate how often each coder agrees with each other member of the same panel, involving 1500 comparisons in each case (3 questions x 100 responses x 5 other coders). In this instance, it is justifiable to adopt the more stringent definition of total agreement: that two coders should apply the same code or combination of codes to a given response.

In order to take into account general differences between organisations in the reliability of the code frames they used in the experiment, we can then express each coder's percentage of agreements as a deviation from the average for his/her organisation:

$$A_1 = P_{i1} - P_i$$

- where P_{i1} is the percentage of coder 1's comparisons with other coders in the same panel where agreement was found;
- P_i is the average of those percentages for the coders of organisation i.

Most of the values of this index of coder agreement were close to zero as shown in Table 6. Only two individual coders stood out as tending to disagree with their fellow panel members. Even in these two cases the patterns were not dramatic. For one of the two, 48% of comparisons with other coders yielded agreement, compared with 57% of all comparisons for that panel; for the other, the figures were an individual 37% against an overall 46%. The overall conclusion must be that unreliability occurred not because one or a few individuals worked quite differently from the rest but because all or most coders contributed to a general pattern of subjective disagreement.

3.SOME IMPLICATIONS

It has to be remembered that this experiment was conducted under artificial conditions. Responses were collected and coding frames developed in the normal way; the coders were members of existing trained panels; but there was no supervision. Each coder worked in isolation, undoubtedly increasing the risk of disagreement. Nevertheless, there clearly is cause for concern. The proportion of comparisons between coders revealing disagreement - 49% is well in excess of the proportion of codings that would normally be checked, let alone changed, under most supervision systems.

Should we ask open questions?

An immediate reaction could be to reject open questions as sources of survey data. Schuman and Presser (1979) provide a rare example of a direct comparison between open and closed questions. They show that the two approaches can yield quite different response distributions and suggest that the data obtained from properly developed closed questions may be the more valid. But a superficially derived closed question is unlikely to produce valid responses. It is likely to miss some dimensions of response and to bias responses towards the arbitrarily pre-set options.

We could reduce the number of open questions asked. The same question may have been asked before or the responses may be predictable on some other external basis. Unfortunately, the results of our experiment show that the questions that yield the simplest response patterns, those that would most easily be closed, are the questions where coding is anyway comparatively reliable. Conversely, the questions most liable to coder variability are those that will be most resistant to closure within normal budget constraints.

An important feature of Schuman and Presser's work is they compared the open question and the 'true' closed question, where both the dimensions and the units of response are explicitly stated to the respondent, either in the reading of the question or in an associated aide such as a showcard. The comparison was not between the open question and what is sometimes loosely referred to as the 'pre-coded' question. This is often an open question from the respondent's point of view, in that the available response categories are not made explicit. Closure is applied only to the interviewer's task: instead of recording an answer verbatim, he or she will be asked to fit the answer into pre-coded categories. This is coding in the field and is unlikely to be even as reliable as coding in the office.

Apart from the difficulty of finding an adequate closed alternative, there are other common reasons for a preference for an open question. We may wish to measure knowledge or awareness, without 'giving away the answer'; or to establish the salient reasons for some activity or belief without putting ideas into the mind of the respondent. Another reason is that given by Montgomery and Crittenden (1977) that open questions "are a source of subtle and often valuable information about reality from the point of view of the respondent". Even if we have our doubts about this argument, the verbatim response does offer a less constrained and summarised response that may, as Schuman and Presser point out, be of value to a future re-user of our survey data.

Should we summary code the responses?

If it is not realistic to prohibit the use of open questions, should we cast doubt on the subsequent process of summary coding? It reflects the basic compromise between a desire for depth of information and the need to measure all respondents on the same dimensions in the pursuit of aggregation and comparability. It is this need for standardisation that gives rise to the problems shown in our experimental results.

It is now realistic to key and store verbatim responses in a computer, in an easily legible and accessible form. This will be essential if such responses are to be kept for posterity and will facilitate a more flexible approach to the use of data obtained from open questions. We can avoid the finality of the summary coding process and consider repeated or interrogatory approaches to the data. But it is not realistic to think that alternative, more reliable, approaches can be used in all cases. The need for summary coding before analysis will persist for the foreseeable future.

Do we need a new coding process?

The more-or-less standard approach to summary coding is that a relatively small panel of coders is briefed for the task. Each codes a large number of questionnaires, raising queries with colleagues or a supervisor. The work is partially checked, but rarely in any scientific way corresponding to the quality control systems seen in industry.

When correlated coder variance occurs, its effect on the precision of a sample estimate will be roughly proportional to the average number of codings performed by each coder. This implies that the panel of coders should be as large as possible, consistent with the need to recruit, train, brief and oversee the panel. But existing procedures are still liable to be relatively inefficient in statistical terms. Judgmental differences and other errors will persist. They could be reduced by attention to the coding frame (discussed more fully below), training and briefing. They could be allowed for by the adoption of double - or treble - independent coding of each response, coupled with a reconciliation process. And they could be spotted by the adoption of more rigorous batch-checking procedures in quality control.

Double or treble-coding, where two or three independent assessments of each response are made and the final coding is determined by reconciliation or - less satisfactorily - by accepting the modal value, is attractive in terms of data quality. It is, however, expensive and time-consuming and is unlikely to be generally adopted. In the case of double-coding, up to one-half of the responses may call for reconciliation. Even then, a substantial number of potential disagreements could escape the system - cases where the two coders agree but neither records the 'correct' answer as defined by the researcher. Improved quality control seems a far more promising development.

4.PRE-TESTING A CODING FRAME

It would seem that the majority of ad hoc surveys common in market research, will continue to use open questions and that the responses will continue to be summary coded, albeit with improved quality control procedures. The key to precision must then lie in the construction, communication and application of the coding frame used to summarise the data.

Montgomery & Crittenden have suggested an improved method of constructing coding frames, basically a method involving double independent coding and reconciliation. But their approach is best suited to smaller surveys and to crucial questions, particularly diagnostic or classifying questions where the requirement is that each respondent should be assigned to the single most appropriate class. A more generally applicable concept would be a formal system of pre-testing the adequacy of a coding frame before the commitment is made to full data processing.

Most routine coding operations start with the extraction of up to 100 verbatim answers and the preparation of a draft coding frame. This frame is then assessed in some totally subjective way before being applied to the full data set. In order to obtain a more objective assessment, we can select a further 100 responses and key them into computer storage. Three or more coders can then independently use a terminal to examine the responses and input their codings. The researcher and, perhaps, the client can also code the 100 responses. The input codings are compared and output is produced showing the reliability of the coding frame and each of its componont categories. This will highlight problems with the frame calling for re-design or special precautions in both briefing and quality control.

This approach does not represent a startling advance. But it makes use of the computer terminal to overcome many of the practical problems that would otherwise arise. And it provides an objective measure of the quality of this particular aspect of survey design. (The same approach could, of course, be used post hoc to measure and declare the quality of research data, a notably rare occurrence in respect of non-sampling errors.

The approach was applied to a recent survey conducted by the SCPR Survey Research Centre. The survey included a crucial open question to local authority tenants who had built up rent arrears: 'How did you come to get behind with the rent the last time it happened?'(An attempt to close the question had proved in a pre-test to be excessively constraining and the list of response categories was felt to be too long for the use of prompting through a showcard - one of the less obvious practical reasons for the persistence of open questions)

The draft coding frame developed for this question (allowing multi-coding of complex responses) consisted of 31 categories, examples being:

HEALTH

1/1 Illness. Any mention of illness, accident or hardship, mental or physical of household member. Include nervous breakdown, agoraphobia, failing memory, addiction. Include mention of people being in hospital, off work sick, on sickness benefit.

CHANGED DOMESTIC CIRCUMSTANCES

1/2 New Baby

1/3 Marital breakup: Divorce, separation, split up with cohabitee.

LOW/LOWERED INCOME FROM WORK

1/4 Unemployment of Household Member: any mention.

1/5 Lowered Wages: Drop in wages not involving actual loss of job (shorter hours, on strike, temporary lay-off, no overtime, off work sick, etc.)

1/6 Low Wages: Household member in low paid job, low income from work (long-term position, not a result of drop in wages as 6).

In the pre-test, 100 responses were independently coded by a coding supervisor, 2 other coders, a researcher and the client. Table 7 shows the results for the first few codes. The example is less than perfect in that reliability proved to be comparatively high! The level of overall reliability, 79%, was high by any standards but especially for a question of this complexity. Of the 31 individual code categories, 9 had reliability indices below 70%, but only 2 of these were categories used in 5% or more of codings. Finally, these two categories were both residual 'Other answer' categories.

Nevertheless, this systematic assessment of the draft coding frame was found useful by the researcher. Several detailed changes were made before full scale coding began. Even with a fairly unfamiliar exercise, the cost of the pre-test was not great. The total cost was about £200 (say $400), evenly divided between the cost of the researcher's involvement and clerical/production costs. This cost can be reduced by about one-quarter and the risk-reduction involved will clearly justify the expense.

5.FURTHER RESEARCH

It is clear that the process of summary coding deserves and will receive further attention. Some priorities for further research can be recognised:

Coding frames

- How best can we reconcile the researcher's views based on research objectives with the views of coding staff based on greater exposure to the material to be coded? Specifically, who should prepare the coding frame?
- How should coding frames be communicated to coding staff; what is the value of illustrative examples?
- How valuable are hierarchial coding frames; do the Kalton & Stowell findings on more or less detailed frames generalise?

Procedures

- Can we recognise and test aptitude for coding staff?
- What supervision and query-raising procedures should be used?
- What procedures should be adopted to control and measure quality?

Broader Issues

- Assessment of the comparative values of open, closed and intermediate question forms.
- Assessment of alternatives to summary coding, especially in the context of computer development.

ACKNOWLEDGEMENTS

This experiment was organised by the MRS Coding Study Group and financed by the participating agencies. Analysis was conducted at the SCPR Survey Methods Centre, funded by the SSRC. The results were first presented at the 1981 MRS Annual Conference by Collins & O'Brien. The example of pre-testing a coding frame was provided by Douglas Wood of SCPR.

REFERENCES

Cochran,W.G.(1950): The comparison of percentages in matched samples. Biometrika,37,256-266.

Collins,M. & Kalton,G.(1980): Coding verbatim answers to open questions. Journal of the Market Research Society, 22,4,239-247.

CollinsM. & O'Brien,J.(1981): How reliable is the coding process? 24th Annual Conference, Brighton; Market REsearch Society.

Kalton,G. & Stowell, R.(1979): A study of coder variability & Applied Statistics, 28,3,276-289.

Montgomery,A.C.& Crittenden,K.S.(1977):Improving coding reliability for open ended questions. Public Opinion Quarterly, 41,2,235-243.

Schuman,H. & Presser,S.(1979): The open and closed question. American Sociological Review, 44,692-712.

Table 7 Results of pre-testing a coding frame

Code	Total incidence (%)		Incidence of use by:				Kj (%)
		Supervisor	Coder 1	Coder 2	Researcher	Client	
1/1	11	13	11	12	11	10	92
2	3	2	4	2	3	3	82
3	7	7	7	8	7	5	82
4	14	15	12	16	14	14	88
5	6	7	5	6	8	5	83
6	2	2	2	2	1	2	89