CODING OF VERBAL INFORMATION

Lars Lyberg, National Central Bureau of Statistics, Sweden

1 CODING

As the term 'coding' has several meanings in various contexts, we give a short review of its use here.

Coding is a major operation of such statistical studies as, for instance, a census of population. It is assumed that every element E_1, E_2, ..., E_N in the population belongs to one and only one of, say k, categories. Usually written information about the element is obtained on schedules. For the purpose of data processing such written information must practically always be converted to numbers ("codes"). This act of converting is called coding although a better word might be 'classification'.

2 THE ERROR PROBLEM

There is ample evidence that the coding operation may be rather susceptible to errors: elements are not assigned into proper categories. As a consequence, there is need for control. The error rate is in fact substantial in many statistical studies. Gross errors of 10-25% when coding multi-digit difficult variables such as occupation and industry are not unusual. The solutions to the error problem have so far mainly consisted of methods for intense training and education of clerks along with the use of more or less efficient verification systems.

However, there are new approaches which will be touched upon in this paper. One example is automatic coding. Despite large gross error rates the net error rate could sometimes be very small. In the 1970 U S census coding of industry and occupation, gross error rates of 9 and 13 percent respectively were estimated. In Jabine and Tepping (1973) it is shown that this error rate results only in a relatively small contribution to the total mean square error for the two variables. Obviously the effect of coders and their error is small in some studies but in others the effects could be alarming. In the U S studies the small effects were obtained in a quality controlled material. Many surveys have no such program and if they have it could be a rather inefficient one. But the problem becomes acute having the forthcoming era of data bases in mind. Suppose we want to study subpopulations such as "people in retail trade". A gross error rate of 10% could be a very serious drawback in this situation. The coding errors result in over- and undercoverage.

3 SOME STUDIES OF ERROR RATES AT THE NATIONAL CENTRAL BUREAU OF STATISTICS, SWEDEN (SCB)

3.1 CODING IN LABOR FORCE SURVEYS

One early study described in Olofsson (1965) treats the "variability in occupation and industry data in Labor Force Surveys". There it is shown that coding errors are seriously affecting the estimates of changes such as the flow between different occupation and industry categories. The main result of the study was that only 40% of the changes in major occupation categories were real changes. The corresponding estimate for industry was 46%. The rest was due to coding errors. As a consequence an exaggerated picture of the mobility in the labor market is created. In fact the coding errors lead to overestimations of 100-200% for some categories.

3.2 CODING ERRORS IN THE 1965 SWEDISH CENSUS OF POPULATION

An evaluation study of coding errors was carried out in connection with the 1965 Swedish census of population. The study is described in Lyberg and Dalenius (1968). The modest prime objective of the study was to illuminate, in a concrete fashion, the performance of the dependent verification used in the census. As a by-product an evaluation of the coding was obtained. Here some selected results are given.

From a population of census material comprising about 70 percent of the 1965 population a two-stage sample of verified census schedules was selected. The population was partitioned into four strata subsequentally resulting in four subsamples. The evaluation study contained the following four variables:

(1) Relationship to head of household
(2) Employment
(3) Occupational status
(4) Industry

The codes used for the variables 1-3 were one-digit-codes; the code used for 'industry' was a three-digit-code.

The samples were coded by a team of three experimental coders. Each coded independently of the others. After that the codes were matched and three cases could occur. First, all three coders could agree; we call that case 3-0. Secondly two could agree but not the third; we call that case 2-1. Finally no two coders agree; we call that case 1-1-1. Apparantly in the first and second cases we are able to define a majority code. We used that code as an evaluation code. When the third case appeared we let a 'super-expert' decide an evaluation code.

Let us give the results for the three-digit variable (4) (industry). Table 1 a-b. A comparison between dependent and independent verification: the majority code M_4 is compared to P_4 and V_4. P_4 means production coder and V_4 means verifier.

Table 1 a

Experimental coder combinations	V_4 agrees with 3	2	1	0	Super expert cases	Total
3-0	451	-	-	24	-	475
2-1	-	44	23	6	-	73
1-1-1	-	-	-	-	5	5
						553

Table 1 b

Experimental coder combinations	P_4 agrees with 3	2	1	0	Super expert cases	Total
3-0	427	-	-	48	-	475
2-1	-	41	24	8	-	73
1-1-1	-	-	-	-	5	5
						553

This is a difficult variable to code. The experimental coders agree only in 475 of the 553 cases (86 percent). The error rate for the production coder is 80/548 or 14,6% and the associate figure for the verifier is 53/548 or 9,7%. As could be seen from the tables the dependent verification system reduces the error rate but the reduction is rather modest. In fact the tables illustrate the well known experience that dependent verification is rather ineffective. Especially the reduction is very small among the 2-1 cases. A possible explanation is that those cases are hard to code and that the coder has a tendency to let an assigned code remain unchanged. But even when we are dealing with the 3-0 cases only a 50% reduction in error rate is registered.

The code for multi-digit variables is often built on the principle of chineese boxes; i.e. the first digit stands for a major classification, the second digit for a classification within this major group etc. This is the case for the industry variable. Usually an error on the first digit is more serious than an error on the second and so on. We have studied the distribution of errors on the different digits for the industry varaible. Let us consider the deviations in the tables above.

Table 2 a-b. Frequency of deviations between M_4 and P_4, and M_4 and V_4 on first, second, and third digit level.

Table 2 a

	P_4			
Digit	First digit	Second digit	Third digit	Total
Deviation cases	41	17	22	80

Table 2 b

	V_4			
Digit	First digit	Second digit	Third digit	Total
Deviation cases	28	9	16	53

As could be seen from the tables most errors are serious; i.e. the error occurs already on the first digit (major group classification).

3.3 CODING ERRORS IN THE 1970 SWEDISH CENSUS OF POPULATION

In the 1970 census of population some improvements concerning the coding quality control program were carried out. For instance, about one third of the schedules was controlled by means of independent verification. However, one third was controlled by dependent verification and for the rest the quality measures were only estimated. So there was a need for an evaluation study. The primary goal for this study was to estimate the coding error rate after verification. A nationwide sample of 7 000 individuals was selected. The population was separated in three different strata reflecting the fact that three different control programs had been used.

Stratum 1: Dependent verification on a 100 percent basis

Stratum 2: Independent verification on a 10 percent sampling basis using an acceptance sampling plan

Stratum 3: Independent verification on a 10 percent sampling basis without using an acceptance sampling plan.

A pool of expert coders was used to generate a set of 'true' evaluation codes for each schedule in the sample. These codes were compared with the production codes after verification and this led to estimates of error rates for the different variables on economic activity. These variables were

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Kind of employment
(7) Way of travel to place of work
(8) Amount of hours at work

Variable (3) was a three-digit one and variable (5) was a four-digit one. The rest were one-digit ones.

In table 3 estimates of error rates for these variables are given.

Table 3 Estimated error frequency (%)

Variable	Percent error rate Stratum 1	2	3	Total population
(1)	4.5	3.8	5.1	4.3
(2)	4.4	5.3	4.0	4.7
(3)	12.6	12.7	16.5	13.5
(4)	4.2	3.1	3.8	3.7
(5)	8.8	9.9	11.6	9.9
(6)	9.5	10.7	5.4	8.9
(7)	11.0	11.3	12.6	11.5
(8)	4.0	4.2	5.4	4.4

The table shows that the multi-digit variables are difficult to code but even the one-digit variables are erronously classified to a relatively large extent. One reason could be that the coding situation is too complex for one coder, i.e. each coder has too many variables to manage. The errors on occupation and industry have the same pattern as has been shown in earlier studies. Most errors occur already on major group classification. Thus a coding error on these variables is often a serious error.

We also calculated the within expert coder variability WV defined as

$$WV = \frac{x}{n}$$

where n is the number of coded individuals in the experiment and where x is the number of unequally coded individuals in two independent trials.

For the five experts in the expert pool the following results were obtained.

Table 4 Within expert coder variability (%)

Variable	Expert A	B	C	D	E
(1)	0.7	1.2	2.4	1.1	0.8
(2)	1.2	2.1	3.0	1.5	1.8
(3)	8.0	10.6	10.9	9.2	7.1
(4)	2.4	0.9	1.8	1.1	1.9
(5)	3.7	8.8	11.6	6.9	5.4
(6)	0.8	2.7	6.0	1.4	2.9
(7)	1.3	1.5	2.1	1.8	2.5
(8)	1.6	3.2	3.9	2.7	2.1

The variability is substantial although these coders have worked for several years with this kind of coding.

4 ALTERNATIVE APPROACHES TO ERROR CONTROL

The control of coding operations could be carried out in many different ways. Some approaches are

- evaluation of coding results
- training and education of clerks
- the use of verification systems
- improving dictionaries and clerk manuals
- using automatic coding.

A total coding quality control system involves more than one of these approaches.

Evaluation of classification results is the basis for dimensioning the quality control efforts. We have already given examples of different evaluation studies. The results of such studies give hints concerning the size of the necessary quality control program.

Evaluation systems are based upon the existence of 'true' codes which are generated by means of more skilled clerks or expert coders. These true codes are compared to those assigned by the production coders and an estimate of production coding gross error rate could be calculated. Evaluation studies are, for instance, found in Fasteau et al (1962), Fasteau et al (1964), Minton (1969), Jabine and Tepping (1973) and U S Bureau of the Census (1972).

The training and education of clerks is indeed valuable since the error rate curve often decreases with time. If it is possible to 'cut' error rates at the beginning of a coding operation one will probably get a more acceptable average outgoing quality.

The literature covering this field is not especially extensive. However, the subject is discussed in Minton (1969) and in Dalenius and Frank (1968). In the latter the idea about using master sets is presented. A master set is a set of elements for which the correct classification is known. Such a set could be used during the training period and as a device for controlling the production process.

The use of verification systems is important to keep up the aimed at quality level. However, the systems could sometimes be rather inefficient, i.e. errors of type I and type II could occur.

The impact of these errors on single sampling plans is discussed in Minton (1972). The flow of coders between total and sampling controls is another problem. The flow must be regulated by means of some prespecified criterion. In Cook (1961) a special point system is given, where each coder receives a point for each erroneous coding. In Minton (1970) some other decision rules for administrative applications of quality control are discussed.

There are two main schemes for verification of coding. These are called dependent and independent verification. Dependent means that the verifier has access to the code assigned by the production coder. Independent means that the verifier has no such access and that the decision upon outgoing code must be based on different rules such as majority or modal rules. Within these schemes several realistic sub-schemes could be defined. The schemes could be used on a total or on a sampling basis. We have seen that dependent

verification could be rather ineffective. Many errors are not corrected. On the other hand the superior independent systems are more costly. Dependent and independent verification is dealt with in Lyberg (1967), Lyberg (1969) and Minton (1969).

Obviously many of the coding errors do not depend on the ability of clerks. Often the dictionaries and the clerk manuals are insufficient and cause a great variability in the coding process.

It is possible to use automatic coding in order to master the variability problem and to speed up the whole operation. Verbal discriptions of the variable under consideration are fed into a computer, a built-in dictionary is consulted and codes are assigned by the computer.

5 AUTOMATIC CODING

Automatic coding might be a complement to manual coding. The method has its strength in speeding up the entire operation but it could also be an instrument for reducing the coding variability. The method is described in O'Reagan (1972) and the main components are the following.

The verbal information for an element is transferred to a punchcard or a magnetic tape. Then the information is fed into a computer where a dictionary is stored. The information is matched against the descriptions in the dictionary. If match occurs the element is coded. Otherwise the element is sorted out and coded manually. The system for automatic coding must also contain continuous evaluation.

5.1 THE COMPUTER-STORED DICTIONARY

The dictionary should replace the coding instructions used in manual coding. Thus the construction of such a dictionary is very important. The construction work could be done manually in simple applications, but when dealing with multidigit variables we must have support from the computer. There are several steps in this work, for instance:

A Choise of a basic material
B Sampling a basic file from the basic material
C Expert coding of the basic file
D Establish inclusion criterias
E Construction of preliminary dictionary
F Testing and making complementary additions and reductions in the dictionary.

The basic material should ideally consist of the material to be coded. If you want to apply automatic coding in the 1980 census the dictionary should be based on descriptions actually obtained in the census. Unfortunately time is not on your side. Most of the basic material must be collected from earlier applications of the same survey. It is also possible to get basic material from pilot studies and from other surveys where the same variable is under study. However, those latter possibilities might be hazardous.

In fact it is very important that the basic material is up to date. In the Swedish experiments with automatic coding on census material the basic material consisted of schedules from the 1965 censuses. On the basis of that material independent 1970 and 1965 census material concerning industry and occupation have been coded automatically. We found that the coding of the 1965 material was more successful than the coding of the 1970 material. The probable reason for that is a change in the population during these five years. Changes can be structural, i.e. entry and exit of industry and occupation categories occur. It is also possible that the reporting pattern has changed during such a long period of time. One example could be the following: In the 1965 census of population cleaners described their occupation as "cleaner". In the 1970 census a new term, "local keeper", was used by some cleaners. The new term was not even invented in 1965 and as a consequence it was not represented in the basic material. The result was that the dictionary based on the 1965 census material could not code the 1970 census individuals describing their occupation as "local keeper".

Considering the coding error experience shown above in this paper the expert coding of the basic file ought to be verified. For instance, a sequential independent scheme with two experts (and a third when necessary) could be used. The descriptions of the expert coded basic file are of different kinds. We have descriptions with high or low frequencies which point at specific codes. We have variations of these (including abbreviations, spelling errors and so forth) and we have descriptions with high or low frequencies which do not point at specific codes. When we are constructing our dictionary we are interested in covering the first two of these categories. We want to keep the last one out of the dictionary.

The dictionary could be constructed by man or by computer. Presumably a combination of the two is the most efficient approach. In most of our experiments at the Swedish National Central Bureau of Statistics (SCB) the dictionaries have been constructed manually. However, we now have a program working for computerized construction.

The following is a brief description of the manual construction phase.

The expert coded file is first sorted according to code number (list no 1) and after that alfabetically (list no 2). These two lists are the material for the dictionary construction. List no 1 is used to get some hints about the structure of the verbal descriptions sorted under a certain code.

We now choose a frequency limit for classification of "high frequency" descriptions. Then high frequency descriptions are stored in the DA-dictionary (Direct Access), which is scanned first in automatic coding. After that we start looking for discriminating word strings to deal with the variants.

These word strings are stored in a subdictionary called CM (Central Memory). This dictionary is

scanned if the DA-dictionary fails to code a certain description.

By means of list no 2 we check whether the descriptions stored in the dictionaries are unique or not. This check leads to reducing the dictionaries since only unique or "almost unique" descriptions are permitted.

The word strings in the CM-dictionary, which are expensive to look for, should be common to several descriptions or be parts of special highly frequent descriptions.

We have to control that those word strings which are included in the CM-dictionary do not fit the DA-descriptions for other codes. Besides they must be unique in the sense that the same word string does not show up more than once in the CM-dictionary. Unfortunately such controls can not be carried out until a first version of the dictionary is available for each code.

Parts of this job could be carried out by a computer. Such efforts have been shown in O'Reagan (1972) and in Corbett (1972). At the SCB our computerized system contains two programs. One program, LEXSRT, abbreviates the incoming descriptions. After that the descriptions are sorted and the frequency of descriptions with the same code is computed. This file is now used as an input to another program, DALEX, with a couple of sub-routines, CMLEX and CMLIST. DALEX puts the descriptions in the DA-dictionary except for descriptions with low frequency (this value could easily by changed) and for identical descriptions with different codes. In fact we allow "almost unique" cases. We buy coding degree to the price of a hopefully small computer coding error. DALEX calls the sub-routines CMLEX and CMLIST. CMLEX creates an abbreviated description (a six letter word string consisting of the first six letters of the DA-description) and puts it in the CM-dictionary. If the word string is not unique then a new six letter word string is created starting with letter number two in the DA-description. Then the program tries again. At most six such word strings are created. After that the program gives up. CMLIST removes the unusable word strings from the CM-dictionary.

5.2 MATCHING AND CODING

The general matching problem is that exact matchings can be obtained only for a fraction of the verbal descriptions to be coded. We are saved by the fact that for most variables a relatively small number of DA-descriptions is enough to code a relatively large part of the descriptions. For the variants we use the CM-dictionary. Earlier we have used special matching rules. For instance we used a method based on Spearman's rank correlation coefficient. The method worked but the costs were prohibitive.

For automatic coding with the dictionaries described above we use the program AUTKOD. As an input the file with descriptions to be coded is used. Each such description is abbreviated according to the same rules applied when constructing the dictionary. Then the program checks whether the description exists in the DA-dictionary. If so the code is assigned. If not the first six letter word string of the description is matched with the CM-dictionary. If match occurs a code is assigned. If not a new word string is created according to the same rules applied when constructing the CM-dictionary. If match has not occured after six such trials the description is rejected to manual coding.

5.3 SOME EXPERIMENTS AT THE SCB

At the SCB we have carried out automatic coding of the industry variable. The descriptions come from censuses and Labor Force Surveys. This coding has not been especially successful.

Table 5 Automatic coding of industry

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Manual	1965 census	50	80
2	Manual	Labor Force 1974	65	69
3	Computerized	1970 census	61	83

Perhaps one can accept the low coding degree but the errors are too frequent. One reason is that the descriptions are rather long for this variable. On the other hand we have not been working with the dictionary that much.

We have been more successful with the occupation variable.

Table 6 Automatic coding of occupation

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Manual	1965 census	62	95
2	Manual	1970 census	66	92
3	Manual	1970 census	74	84
4	Manual	1970 census	80	90
5	Manual	Labor Force 1974	81	81
6	Computerized	1970 census	69	87

For census coding we have an acceptable dictionary. The low quality on Labor Force coding is explained by the fact that a translation of the census dictionary was used. Now we have a dictionary based on Labor Force descriptions but it has not yet been tested. The less successful result of the computerized dictionary is explained by the fact that it is still "untouched by human hands". Obviously it is a good raw material for further work.

We have also tried to code goods in the Family Expenditure Survey. The results are good.

Table 7 Automatic coding of goods

Experiment	Kind of dictionary	Kind of data	Coding degree (%)	Quality (% correct coding)
1	Computerized	Family Expenditure Survey 1969	78	93
2	" -	" -	80	93
3	" -	"-	82	96

The results are so promising that automatic coding will be used in the 1978 Family Expenditure Survey.

5.4 GENERAL CONSIDERATIONS

Automatic coding have to be cheaper than manual to be considered. The automatic coding itself is cheap but the punching and the manual coding of the rejects is not. So far we have not been able to calculate costs with enough precision in our experiments. The laboratory differs from reality. However, we are now going to predict the costs for an automatic system in the 1978 Family Expenditure Survey. Manual coding of the whole survey will cost 1,4 million crowns. Automatic coding of the whole survey will cost .07 million. The extra punching of rejected verbal descriptions will cost .2 million. Thus we have quite a margin for manual coding of the 20% rejected and the extra punching of these.

6 REFERENCES

Corbett, J P (1972): Encoding from Free Word Descriptions. U S Bureau of the Census, Draft.

Dalenius, T and Lyberg, L (1968): An Experimental Comparison of Dependent and Independent Verification of Coding, Memo.

Dalenius, T and Frank, O (1968): Control of Classification, Review of the International Statistical Institute, Vol 36:3

Fasteau, H, Ingram, J and Mills, R (1962): Study of the Reliability of Coding of Census Returns, American Statistical Association Proceedings, Social Statistics Section.

Fasteau, H, Ingram, J and Minton, G (1964): Control of Quality of Coding in the 1960 Censuses. Journal of the American Statistical Association, Vol 59, No 305, pp 120-132.

Jabine, T B and Tepping, B J (1973): Controlling the Quality of Occupation and Industry Data, Invited paper to the 1973 ISI meeting.

Lyberg, L (1969): On the Formation of Coding Teams in the Case of Independent Verification under Cost Considerations, Forskningsprojektet "Fel i undersökningar", rapport nr 18, Stockholms universitet.

Lyberg, L (1967): Beroende och oberoende kontroll av kodning, Forskningsprojektet "Fel i undersökningar", rapport nr 4, Stockholms universitet (In Swedish).

Minton, G (1969): Inspection and Correction Error in Data Processing, Journal of The American Statistical Association, pp 1256-1275.

Minton, G (1970): Some Decision Rules for Administrative Applications of Quality Control, Journal of Quality Technology, pp 86-98.

Minton, G (1972): Verification Error in Single Sampling Inspection Plans for Processing Survey Data, Journal of the American Statistical Association, pp 46-54.

Olofsson, P O (1965): PM beträffande variabiliteten i näringsgrens- och yrkesangivelser vid arbetskraftsundersökningar, SCB/UI. (In Swedish).

O'Reagan, R T (1972): Computer-assigned Codes from Verbal Responses. Communications of the ACM, No 6.

U S Bureau of the Census (1972): Coding Performance in the 1970 Census, U S Government Printing Office, Washington, D.C.