Coding of survey responses – quality assurance efforts at Statistics Sweden

Svensson, Jörgen
Statistics Sweden, Process Department
Örebro, SE-701 89, Sweden
E-mail: jorgen.svensson@scb.se

Introduction

Coding is in this context defined as follows: to use codes to classify survey objects to different categories according to an established classification or another predefined code frame. This paper is about the endeavors of Statistics Sweden in recent years to assure the quality of the coding process. There have been several reasons for focusing on this process. One reason is that the process of coding survey responses is error-prone. This has been shown by evaluations in 2007 of the occupation coding at Statistics Sweden. Regular quality controls are needed. Another reason is that the coding is time-consuming and thus costly. Rationalization through modern IT support is required. The director-general decided in 2008 that Statistics Sweden shall work towards certification according to the international standard ISO 20252 for market, opinion and social research, see International Organization for Standardization (2006). The standard requires quality assurance and quality control of the coding process. The requirements are quite specific in terms of percentage of records that should have verification coding. How these requirements could be fulfilled at Statistics Sweden has been investigated thoroughly. The overall objective for the broad work described in this paper is to achieve continuous improvements for the coding process.

Quality control of coding

Statistics Sweden has established a standard routine for quality control of the coding process, which is described in the intranet process information system. The main operations are to conduct *independent verification coding* after the original coding and – if the original coding and the verification coding differ – to decide upon the 'correct' code by a *reconciliation* (adjudication) process. The survey manager, in consultation with the process owner for the coding process, then has to stipulate what should be considered *unacceptable error rates*, whereupon measures are to be taken to achieve better quality. This routine is to be implemented for all relevant surveys during 2011.

Automatic coding using a dictionary shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for at least five percent of the automatically coded records in a suitably selected paper questionnaire survey. This quality control should be performed at least once every third year, starting 2011. If the error rate is unacceptable, the dictionary has to be revised. (The amount of automatic coding is relatively small at Statistics Sweden.)

Coding at telephone interviews using a reference file within a CATI system shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for at least five percent of the coded records in the Labour Force Survey for a three months period (restricted to records the first time they appear in this longitudinal survey). This quality control should be performed at least once every third year, starting 2011. If the error rate is unacceptable, the survey manager, in consultation with the process owner, has to decide on what measures should be taken to improve the results.

Computer-assisted manual coding using a comprehensive reference file shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for at least five percent of the coded records in each relevant survey. This quality control is a form of *acceptance sampling* and should be performed continuously, starting 2011. Often the percentage will be higher than five percent, since it must be possible to measure the coding results for each and every coder. If a coder's work contains frequent errors, that coder's work shall be verified and suitable training shall be given. Efforts might be

needed to raise the competence both on the individual level and the group level. Apart from training, the reference files can be improved in order to reduce the error rates in coding.

External coding is a common alternative at Statistics Sweden. Either the data providers perform the coding themselves and send us the coded records, or the data providers code the records according to their own nomenclatures and send us the records, whereupon translation keys are utilized to set the classification codes. For external coding it is not possible to conduct quality control according to the routines above. However, evaluations and quality assurance are recommended for the pertinent surveys.

Sampling for verification coding can be designed in different ways. One easy alternative is simple random sampling. Another alternative is stratified simple random sampling with strata equaling the coders. A third alternative is stratified systematic sampling, which is relevant to use when verification coding should be conducted simultaneously with the ordinary coding within the survey production period.

The quality control has a twofold *purpose*. On one hand, the coding process ought to be improved so that the errors will be less frequent the next survey round. On the other hand, the survey results should (time permitting) be adjusted directly, either through corrections of the erroneous records by using the final, possibly reconciled codes from the sampled verification records or (which is better) through adjustments to the population level using these corrected codes. To start with, the first purpose will be aimed at in most cases.

Deficiencies in the coding results can be shown in two dimensions. First, error rates for *each coder* are calculated directly, using unweighted data on the final codes from the sampled verification records. The result is used for follow-up of coders and groups of coders with frequent errors. Secondly, *gross errors* and *net errors* are calculated in order to show the effects on microdata (registers) and macrodata (statistics), respectively. These errors can be produced both unweighted for the sampled verification records and weighted for the population. An issue here, yet to be resolved in practice, is the definition of an unacceptable error rate. Product-specific features must then be taken into account.

A new IT tool for computer-assisted coding

A *modern IT tool*, named Prisma, for computer-assisted coding has been developed at Statistics Sweden during the period November 2010 – April 2011. The development was necessary, since the IT applications used for the Labour Force Survey and several other surveys have been almost out of date and too dependent on a few persons. Another necessity was to build functionality for verification coding and reconciliation, according to the new standard routine described above. The functionality needed could not be found in any commercial IT tool or in any tool used at statistical offices.

Prisma is programmed in C#.NET and will support coding for all different classifications and (almost) all surveys. The tool provides a user-friendly interface for the coders, which is important for their working environment. Prisma will rationalize the work of the coders by giving instructions and support for coding decisions. When a new record is opened, an automatic search through reference files and suchlike is done. With a simple click on a suitable category, all codes for different classifications within an area, such as occupation, are set automatically. Difficult records can be placed on a waiting list. Scanned images of questionnaires are possible to import. Functionality for handling access rights for original coders, verification coders and 'reconcilers' is also available. A few sampling techniques are supported, to start with. Process data are generated. Configuration for different classifications and surveys is done through a module within Prisma.

The development project has been carried out in close cooperation with the coding staff, which has been testing the software in different early stages. Their wishes and requests have been integrated to a large extent in the specification of requirements for Prisma. Seminars have been arranged to present the IT tool and the standard coding process. In May 2011, the director-general approved Prisma as a standard tool for computer-assisted coding.

The plan now is to *implement* Prisma in almost all relevant surveys during 2011. Among these

approximately ten surveys, the Labour Force Survey is the most important one and also the first one that has been implemented (in April–May 2011). The classifications used in this survey are ISCO for occupations (in two versions), the Swedish socio-economic classification, the Swedish standard industrial classification, sector and county.

Training and other quality assurance measures

Due to the fact that coding of open-ended responses often is subjective in nature, proper *training* is essential. The coders (including some interviewers) at Statistics Sweden are trained on how to use classifications, reference files etc. through awareness of rules regarding what should be included or excluded from a given code. The training is done through a small web based application connected to a database with exercises, which are given in three degrees of difficulty.

The different co-workers, in a wide sense, involved in the coding process have to cooperate in order to assure the quality of the process. To that end, the different *roles* have been pointed out clearly via a decision by the director-general. Here follows some of the roles: coder, interviewer, survey manager, classification owner, process owner (of Process & Analyse) and IT staff responsible for Prisma. The classification owner, for example, will have to revise the classification dictionary if the error rate is unacceptable according to the quality control.

A short *instruction for developing code frames*, when not having standard classifications, has been established. Among other things, it points out how to construct categories and instructions, and how to treat 'other' (catch-all) categories. The survey manager is responsible for producing, documenting and possibly revising the code frame in accordance with the instruction. All classifications and code frames shall be stored in a common repository.

The coding staff is mainly *centralized* to one of the two data collection departments. A full centralization is on its way. The rationale for the centralized approach is to have a strong group of coders that work in a similar fashion with a common IT tool. Dialects among coders in different surveys should be avoided as far as possible, so that consistency and comparability can be achieved. Moreover, 'coder variance' should be minimized by employing many coders for each survey instead of a divided coding group with few coders per survey.

Information on the coding process and required quality assurance measures is given in the intranet process information system. Application will be followed up annually via the directors of the subject matter and data collection departments.

Conclusions and future work

An extensive work has been done throughout the last years in order to improve the process of coding survey responses at Statistics Sweden. The main task now is to implement the adopted methodology and routines as well as Prisma, our new IT tool. Awareness and competence regarding quality assurance of the coding process have to be increased among all parties concerned. Adopted routines have to be evaluated.

In 2012, there will hopefully be a new project for further development of Prisma. The whole specification of requirements could not be met in the recent project. The use of Prisma during 2011 will certainly lead to new functionality requests from the coders. Automatic coding using a dictionary might be included in the next version of Prisma. There is also a need for better connections between Prisma and other standard IT tools at Statistics Sweden. Monitoring of telephone interviews is scheduled to be implemented in 2012 and could lead to a partially different routine for quality control of coding at telephone interviews.

REFERENCES

International Organization for Standardization (2006). Market, opinion and social research – Vocabulary and service requirements (ISO 20252:2006, IDT).