WP. 24
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

RECOMMENDED PRACTICES FOR EDITING AND IMPUTATION IN THE EUROPEAN STATISTICAL SYSTEM: THE EDIMBUS PROJECT

Invited Paper

Prepared by O. Luzi (Italy), T. De Waal (Netherlands), B. Hulliger (Switzerland)

I. BACKGROUND

1. National Statistical Institutes (NSIs) need to provide detailed high-quality data on all relevant aspects (economic, social, demographic, etc.) of modern societies. A common tendency at Statistical Agencies is to improve the data quality by standardizing and harmonizing statistical survey processes: to this aim, manuals, guidelines, recommendations and other tools are developed in the different areas of survey processing, such as Quality Guidelines (see for example Statistics Canada, 2003), and Current Best Methods (among others, Statistics Sweden, 2004, and Granquist *et al.* 2002).

2. At European level, given the differences of the National statistical systems, the need for standardising and improving the quality of statistics production in the European Statistical System (ESS) is essential for an effective harmonisation. In the framework of the Leadership Group on Quality (LEG on Quality), specific recommendations on the development of tools such as Current Best Methods (CBM) and Recommended Practices (RPs) were addressed to the European National Statistical Institutes (NSIs) (Eurostat, 2001). In particular, the LEG recommendation number 11 states that "*A set of recommended practices for statistics production should be developed. The work should start by developing recommended practices for a few areas followed by a test of their feasibility in the ESS*".

3. Starting from the work of the LEG on Quality, projects regarding the implementation of the recommendations through the effective cooperation of European Countries were launched by Eurostat. Among others, a project aiming at developing an RPM in the area of editing and imputation in cross sectional business surveys (called EDIMBUS) has been supported. Integrating research on methods for editing and imputation in cross-sectional business surveys carried out at NSIs, as well as sharing knowledge about editing and imputation methods and practices to all potential users in the common environment of the ESS, is highly desirable because of its high impact on data quality and comparability through the standardization of methods and practices in the ESS in this crucial area of data processing.

4. Developing such a handbook requires collecting as much information as possible on currently used approaches and practices for editing and imputation in cross-sectional business

surveys at the different NSIs at European level. To this aim, a state-of-the-art investigation has been planned among the project activities, and is currently being performed, involving all the European Countries as well as some leading Statistical Agencies outside Europe.

5. In the paper the project objectives and its expected outputs are described. The main problems faced until now and the intermediate project's results are illustrated. In particular, the preliminary results of the state-of-the-art investigation are reported.

II. RECOMMENDED PRACTICES FOR CROSS-SECTIONAL BUSINESS SURVEYS: THE EDIMBUS PROJECT

6. Developing a Recommended Practices Manual (RPM) in the area of business statistics has been considered a priority for several reasons.

7. First of all, particularly in business surveys there is a large heterogeneity in practices and methods used for editing and imputation (E&I) at both NSIs and European level. An RPM is expected to have a high impact in terms of harmonisation of practices/methods and hence of comparability of data produced in the ESS.

8. Moreover, it is known that E&I has a strong impact on the quality of final data: since E&I processes, particularly in business surveys, generally consist of an integrated set of complementary solutions, each dealing with a specific data/error problem, higher quality of statistical information can be obtained by using best practices and approaches at each stage of the E&I procedure. For this reason, in the area of business surveys, the availability of a tool, like Recommended Practices, which represent a support for the design and the management of complex E&I strategies is crucial.

9. Furthermore, particularly in business surveys E&I is recognised as one of the most time and resources consuming survey phases. An RPM in this area is expected to contribute to less expensive statistics and to shorten the time from data collection to publication of results.

10. Finally, business surveys are characterised by a considerable lack of documentation concerning E&I methods and practices. In this area there is a strong need for performing systematic studies and for developing common practical and methodological frameworks. The development of an RPM in this area is expected to represent a valuable and useful contribution in this direction.

11. The EDIMBUS[1] project aims at developing a manual of *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys in the ESS*. The project started on January 2006 and will end on June 2007. The project partners are the *Italian National Statistical Institute* (Istat, co-ordinator), the *Centraal Bureau voor de Statistiek* (CBS Netherlands) and the *Swiss Federal Statistical Office* (SFSO Switzerland). The support of other statistically advanced institutions both in ESS and over the world is expected in different stages of the project.

12. The project's activities have been structured as follows. First of all, information about the state-of-the-art in the area of E&I in cross-sectional business surveys is collected based on a literature review and through an investigation, partly by means of a survey, of the current practices and methods adopted in the ESS and in other statistically advanced Institutions. The latter activity

[1] The project is partially funded by the European Community (Represented by the Commission of the European Communities) in the context of the Eurostat Grant Program 2005 in the field of *Theme 10: Quality Management and Evaluation*

aims at ensuring that the RPM takes into account the most advanced E&I solutions (in terms of accuracy, timeliness, costs) currently in use at various Statistical Agencies.

13. Based on the collected information, an initial draft of the RPM will be developed (see next section). It will be revised by statisticians and survey managers within the partner's Agencies and at the other Institutions involved in the state-of-the-art survey or interested in the RPM. The main objective of this evaluation is to verify the Manual's feasibility from both a methodological and an operational point of view in the different NSIs contexts. Furthermore, in order to make the Manual meet as much as possible the needs of potential end-users, this evaluation is expected to allow an improvement of the Manual based on the proposed suggestions and integrations.

14. The final project activities will consist in adjusting the draft manual based on the revision activity's results. The final output will include an operational strategy to be followed in order to disseminate and implement the handbook in the ESS, avoiding that it quickly becomes useless because of lack of further updates.

III. THE RECOMMENDED PRACTICES MANUAL

15. Few experiences at other Countries in developing recommended practices in this area of E&I can be found in literature. Some useful references are Williams *et al*. (2000), who developed a handbook for Statistical Data Editing Design in the area of households surveys with particular focus on those using Computer-Assisted Interviewing, the FCSM methodological report on E&I in federal Statistical Agencies (FCSM, 2002), and the Scarrott's paper (2005), which focuses on selective editing providing a summary of problems, operational suggestions and useful references to approach them.

16. As already mentioned, the handbook produced by the EDIMBUS project should establish general criteria for developing E&I strategies for cross sectional business surveys in the ESS. It should represent a systematic guide for the design, implementation, test and documentation of E&I activities. The Manual will deal with several aspects and issues in the area of E&I, covering both theoretical and practical issues. It will provide quick recommendations on *what should be done* and *how* when building up an E&I strategy in the area of business statistics. Proven good methods and practices for performing the statistical operations will be recommended. The different survey contexts will be taken into account and the main advantages and drawbacks of each method will be highlighted. The expected impact of such a tool is the promotion of good practices with the consequence of harmonising the survey methods, thus improving the quality of the statistics produced. In particular, at NSIs level, the Manual is expected to increase the standardization of E&I processes through the direct support to statisticians and survey managers from the stage of the design to the phase of final documentation of their E&I own strategies.

17. Taking into account the wide range of survey contexts and existing approaches, as well as the high heterogeneity of surveyed phenomena and data problems, the initial Manual template has been defined, and the broad contents of the handbook have been delineated. These results are both considered provisional and will be probably revised based on the following project activities. In fact, open problems mainly relate to the difficulty of identifying the optimal trade-off between the high and complex amount of information to be organized in the Manual, and the need for developing an agile and effective tool. A particularly complex problem which remains still open is how to balance the theoretical and the operational issues in order to ensure the development of effective support for survey managers in terms of completeness, shortness and operational efficacy.

18. In general, it has been decided that the Manual will cover the most relevant topics in the field of statistical data E&I:

- methods and practices for E&I in cross-sectional business surveys (e.g. for outlier detection, for the identification of influential errors, for dealing with systematic and stochastic errors, for treating missing values and so on);
- methods and practices for designing an E&I strategy;
- methods and practices for evaluating the effects of E&I in cross-sectional business surveys;
- methods and practices for documenting E&I activities in cross-sectional business surveys, and for disseminating information to users about E&I and on data quality.

19. The manual sections should deal each with a specific E&I problem, structured following a pre-defined general E&I process flow assumed for the design, testing and documentation of E&I in business surveys. Sections should have as much as possible a standardized structure in order to facilitate the reader in quickly identifying the most relevant information for each treated topic. In order to make the handbook more readable, methodological and technical details about the illustrated methods and practices should be organized in the manual Annexes, together with examples drawn from the state-of-the-art survey (first stage of the project).

20. Each section (E&I problem) should contain a short description of most effective and/or used methods/practices, together with a brief discussion on the data context where each method is either applicable or most effective, and an illustration of advantages and possible drawbacks relating to its use. For each section, recommendations suggesting *what should be done* should be given in the form of a checklist, i.e. a list of items recommending actions which should be performed when dealing with the issue.

21. A very difficult objective is to help survey managers in the overall design of an E&I strategy. While methods and practices for particular problems are more or less known though not always well documented it is relatively rare to find documents about the approach one should follow in designing a strategy.

IV. THE STATE-OF-THE-ART SURVEY

22. In order to develop an RPM which would be effective in terms of balance between theoretical and operational aspects, useful in terms of providing quick support for the design, testing, management and documentation of E&I activities, and complete in terms of being exhaustive with respect to the expectations and needs of end-users, it has been considered fundamental to draw on the knowledge and practical experiences of different NSIs with respect to E&I methods used in cross sectional business surveys.

23. To this aim, a state-of-the-art survey has been planned and is being performed involving the 3 partner’s Institutions (Istat Italy, CBS Netherlands and SFSO Switzerland), all the other member states of the European Union, and some external Agencies considered most advanced in terms of adopted methodologies and practices and/or established strategies for the design, management and documentation of E&I, Statistics Canada, the U.S. Bureau of the Census, and the ABS.

24. Each Country was asked to fill in the pre-defined questionnaire for at least one business survey in one business area: Structural Business Statistics (SBS), Short Term Business Statistics (STS), and Economic Censuses (EC). The aim was not only to collect information about the approaches adopted in the different Countries to deal with some general E&I problems, but also to

have an as much complete as possible overview of the most relevant problems in the different types of business surveys, and to gain knowledge about E&I strategies developed in the different contexts.

25. Based on these requirements, and taking into account the Manual characteristics and structure, a questionnaire has been developed and sent to the involved Institutions. The questionnaire is divided in three sub-sections. The first one asks for general information on the survey (name, periodicity, survey design, and amount of units/variables collected). The second sub-section relates more specifically to the E&I process: respondents are asked to summarize their own overall E&I strategy and to indicate the approaches adopted for the different types of errors they have to manage (systematic errors, influential errors, outliers, not influential random errors, item non-responses). A set of pre-defined approaches is indicated in the questionnaire (for instance, selective editing, macro-editing, deductive imputation, and so on), but the respondent is allowed to indicate additional problems/approaches and comment. For each currently used method, respondents are asked to indicate their main advantages and drawbacks. A specific question relates to the use of registers at the E&I stage at elementary or aggregate level. Information on how the E&I strategy is set up, tested, and updated is required, together with information on how E&I processes are documented. Respondents are asked to attach any manuals or guidelines possibly used at some stage of the E&I design/management. The third part of the questionnaire contains few final, general questions relating to E&I management at the various Agencies.

26. Given the heterogeneity of definitions and concepts adopted in the different survey contexts not only at European, but also at national level, respondents are provided with a glossary of terms (see Appendix 1) in order to comparable responses. One particular problem was that the English notion "editing" often includes changes to the data: since from the point of view of organising the process of E&I this is considered not suitable, in the EDIMBUS project the notion of "editing" is restricted to checks on the data for error identification, excluding all data changes (imputations, re-interview, etc.).

V. SURVEY RESULTS

27. The state-of-the-art investigation, even if restricted to a limited set of Countries and surveys, can be considered a very useful source of information about some general aspects characterising the current E&I activities and approaches in the extremely vast world of business statistics. It can also be considered a way to make European Statistical Agencies conscious about the strong need of working in the direction of: 1) systematize the process of designing, implementing, evaluating and documenting E&I processes in business surveys; 2) standardise E&I approaches, methods and practices; 3) integrate research and experiences in the area of E&I in order to improve the data quality and comparability at National and ESS level.

28. As for the survey conducted at the partners Institutions (*internal survey)*, 22 questionnaires have been filled in. Beside the 3 project partners, 20 Agencies of non-partner Countries (*external survey*) participated to the investigation, filling in at least one questionnaire (41 questionnaires) and/or providing some kind of documentation about their current E&I strategies. In Table 1 the distribution of the 63 questionnaires by type of survey (SBS, STS, EC) is shown. As it can be seen, 37 responses (58.7%) relate to surveys in the SBS area, 25 questionnaires (39.7%) report information for surveys in the field of STS, and one questionnaire relates to an Economic Census (the Italian Census of Agriculture).

29. The following analyses have been performed on the subset of 49 questionnaires (78%) at present entered in the *excel* survey database. Though performed on incomplete data, these analyses are useful for identifying some interesting aspects and highlighting a number of crucial issues. The analysed questionnaires correspond to 28 responses in the SBS area, 21 in the STS one, and the Economic Census.

Table 1: current approaches by type of error

Survey	SBS	STS	EC	Total
External	25	16	-	41
Internal	12	9	1	22
Total	37 (58.7%)	25 (39.7%)	1 (1.6%)	63

30. One of the objectives of the survey was to understand which approaches, and to which extent, are currently used in business surveys to deal with the different types of errors. Table 2 contains the absolute and percentage frequencies of surveys using the various classes of approaches considered in the questionnaire, by type of survey (*All*, *SBS*, *STS*). As it can be seen, a very high percentage of surveys (from 81% to 82%) use *Manual editing/follow-up*. *Macro-editing* is also widely used (68% in SBS, 62% in STS), while *Selective editing* and *Graphical editing* are adopted in a low percentage of cases: 28,6% and 38% of surveys declare to adopt *Selective editing* in SBS and STS respectively, only 14% of surveys declare to use *Graphical editing* in SBS, about 28% in STS. *Deterministic checking rules* for error localization are widely used (75% for SBS, around 67% for STS), while the *Fellegi-Holt principle* is used only in 6 surveys and only in the SBS area (among them, 2 Italian Surveys use the SCIA software (Riccini *et al.*, 1995) to deal with qualitative data). As for imputation, *Model-based approaches* (53% in SBS, 62% in STS) are more used than *Donor-based* techniques (32% in SBS, only 19% in STS). *Deductive imputation* is also largely used (57% in SBS, 43% in STS). Note that 90% of surveys declare to make *Use of other sources and/or historical data* in their E&I procedures in STS; this percentage decreases to 71% for SBS.

Table 2: Frequency of use currently adopted E&I approaches

	ALL*		**SBS****		**STS*****	
Method	*N. surveys*	*%*	*N. surveys*	*%*	*N. surveys*	*%*
Manual review / follow-up	40	81,6	23	82,1	17	80,9
Selective editing	16	32,7	8	28,6	8	38,1
Macro-editing	32	65,3	19	67,9	13	61,9
Graphical editing	10	20,4	4	14,3	6	28,6
Deterministic checking rules	35	71,4	21	75,0	14	66,7
Minimum change error localisation (Fellegi-Holt principle)	6	12,2	6	21,4	0	0,0
Deductive imputation	25	51,0	16	57,1	9	42,9
Use of other sources and/or historical data	39	79,6	20	71,4	19	90,5
Model-based imputation (e.g. regression, ratios, mean)	28	57,1	15	53,6	13	61,9
Hot-deck imputation (e.g. nearest neighbour, random donor)	13	26,5	9	32,1	4	19,0
Robust estimation / re-weighting	14	28,6	9	32,1	5	23,8

* w.r.t. the total of questionnaires (49)
** w.r.t. the total of surveys in the SBS area (28)
*** w.r.t. the total of surveys in the STS area (21)

31. Table 3 contains the detailed frequencies of use of the previous classes of approaches by class of errors. Subdivision in error type appears difficult to respondents, in fact there are some *suspicious* frequencies in the table (e.g. use of the Fellegi-Holt paradigm for systematic errors) generally due to not completely clear or not agreed definitions and concepts in the glossary, but also to the difficulty for respondents to adapt their own E&I flow to the pre-defined flow underlying the table's structure. In particular, most respondents declares that E&I activities in their surveys are not structured by type of error, and that generally the nature of errors is understood during the data E&I process.

Table 3: Frequency of currently adopted E&I approaches by type of error

Method	Systematic errors	Outliers	Influential errors	Random errors	Item non-response	Residual errors
Manual review / follow-up	19	34	25	15	27	19
Selective editing	4	13	8	3	11	3
Macro-editing	5	29	16	3	2	18
Graphical editing	2	8	4	1	0	2
Deterministic checking rules	16	26	16	15	18	7
Minimum change error localisation (Fellegi-Holt principle)	2	1	0	6	5	0
Deductive imputation	9	6	4	8	22	4
Use of other sources and/or historical data	7	27	20	9	24	7
Model-based imputation (e.g. regression, ratios, mean)	1	10	6	5	24	1
Hot-deck imputation (e.g. nearest neighbour, random donor)	2	2	3	8	9	1
Robust estimation / re-weighting	0	6	1	3	4	2

32. Summarizing, Tables 2 and 3 give a first overall picture of the situation w.r.t. the currently adopted solutions, and help to identify some basic problems with definitions and other conceptual aspects used in the questionnaire structure and wording. Further analyses are needed for the project's purposes, supported by the comments and the other information provided by respondents in their questionnaires (e.g. short descriptions of E&I processes and data flow, and all the additional information about methods possibly not included in the questionnaire's table).

33. One aspect not directly connected to the development of the RPM, but very important in the area of E&I relates to the data checking activities performed at the data capturing stage. The increasing interest of Statistical Agencies in Computer Aided Interviews (CAI) is well known in literature. In this respect, the survey results confirms that a high percentage of surveys (60%) use electronic questionnaires as either unique solution for data capturing, or in combination with traditional paper questionnaires. About the same percentage characterises SBS and STS areas (63% and 57%, respectively).

34. It is known that some leading Statistical Agencies (such as Statistics Canada, Statistics Netherlands, Census Bureau), have developed high-level generalized software for E&I of business survey data. In this respect, among the analysed surveys, 19 ones (39%) declare to make use of some type of such tools: 11 surveys in the SBS area (38%), 6 ones in the field of STS (28%). As a matter of fact, in most cases this software consists of *ad-hoc* SAS programs. Actual generalized software currently adopted at the responding Statistical Agencies are Banff (one Canadian and two Italian surveys), Blaise (Slovenia), SLICE (Netherlands) SCIA (used in two Italian surveys to deal with qualitative information), two different specialized tools developed by Eurostat (Poland and Italy). These results could be due to different factors: on one hand, the availability of few

generalised tools and the high costs generally needed to develop them at Statistical Agencies; on the other hand, a certain lack of knowledge of existing software and/or their embedded functions, or the high costs needed to get them, and integrate their methodologies in the existing survey processes.

35. The problem of preliminary tests of E&I procedures is recognised as crucial in all statistical surveys. The question *Do you perform preliminary test of the E&I process in your* survey, was answered negative 50% of times. Concerning the reasons, 50% answered *not enough resources*, 45% *not suitable data available*, and about 33% *lack of time*[2]. This result seems to confirm that there is a lack of preliminary evaluation and tuning of E&I procedures before they are used in survey production processes.

36. One interesting (and problematic) result concerns the question *Could you range the used resources for the E&I process in respect to the workload of the whole survey?* (Note that respondents were let known that E&I activities performed at the data capturing stage were to be excluded from the evaluation). In Tables 4 and 5 the number and the percentage of surveys by amount of resources spent and by type of survey (*All surveys*, *STS*, *SBS*) is provided.

37. It is well known in literature that E&I activities may account for a considerable amount of survey resources (human, budget, time). The results of the state-of-the-art-survey further confirm this fact, but at the same time suggest that in the survey manager's perception the amount of resources spent for E&I is very high. In effect, there seems to be a sort of "overestimation" of the actually used resources, since about 48% of surveys claim to spend more than 50% of their overall resources for data E&I (this percentage increases to around 54% in the SBS area while it is 37% in STS). The highest percentage of answers in the STS area (44%) relate to the range *40%-60% of resources*, while the highest frequency for SBS (43%) correspond to the class *>60% of resources*.

38. These results could be due to several factors:

- *how to measure costs of survey processes is an open problem*. Since it is not easy to define the concept "amount of resources", the unity measure to be used and so on, answers can be interpreted as the result of the survey managers theoretical estimation of the overall E&I workload in their surveys;
- *it is difficult to focus attention on a specific E&I stage*: in most cases the data checking and correction activities are continuously performed on data from the time when they start incoming; furthermore, maybe most of respondents included in their evaluation also preliminary editing activities performed by clerical reviewers on raw data, e.g. those relating to some type of "obvious" errors.

Table 4: Resources spent at the E&I phase by type of surveys

		% of resources							
Surveys		*<30*	*30-40*	*40-50*	*50-60*	*60-70*	*70-80*	*80-90*	*>90*
All	*N. surveys**	7	7	9	5	7	2	4	3
	%	15.9	15.9	20.4	11.4	15.9	4.5	9.1	6.8
STS	*N. surveys**	1	4	5	2	0	0	2	2
	%	6.2	25.0	31.2	12.5	0.0	0.0	12.5	12.5
SBS	*N. surveys*	6	3	4	3	7	2	2	1
	%	21.4	10.7	14.3	10.7	25.0	7.1	7.1	3.6

** 5 missing values*

[2] multiple answers were allowed.

39. The problem of evaluating and documenting E&I activities is generally recognised as central when discussing about standardization of processes and comparability of results. For this reason, the question *Do you document the results of your editing and imputation process?* has been included in the questionnaire. About 82% of the analysed surveys declare to produce some documentation of E&I results: among them, 50% produces *Methodological reports*, 75% *Technical reports*, 40% *Quality Reports to Eurostat*, and 62% computes some type of *Indicators*. These percentages show an underlying attention to the problem of documenting E&I: the still open question relates to the level of standardization of the produced reports and the computed indicators, as well as their completeness in terms of provided information (structure of the E&I strategy, quality of survey data, error profile, documentation of E&I effects on data, and so on).

40. The problem of standardization has also inspired two other questions: *Do you have manuals or guidelines, developed at your institute, describing recommended Edit and Imputation processes or methods in general?* and *Do you use manuals or guidelines, developed at other institutes, describing recommended Edit and Imputation processes or methods in general?* Only 4 Countries (Canada, Finland, Norway and UK) declare to have developed some kind of such tools. Beside them, the other affirmative responses generally refer to Eurostat manuals, UN/ECE documents, or other "non-standard" methodological documentation developed at the Statistical Agencies. From these results it is evident that there is a real need for developing and disseminating standard tools in the specific area of E&I at Statistical Agencies and/or at European level.

41. The aim of the question *In your institution, do you have an established procedure for obtaining approval for the E&I strategy of a survey?* was to investigate the problem of the standardization of at least the high-level organization and control of E&I processes at Statistical Agencies. Only 5 Countries (about 22%) declare to have such procedure. In particular, in all these Countries at least the *Design of the E&I strategy of the survey* has to be submitted for formal approval. Only in 1 case the *Assessment of the E&I strategy* is requested, while in almost all cases *Documentation of E&I strategy* has to be submitted.

42. Based on the collected information on such a wide range of contexts, it is expected to be able to improve and update both the structure and the contents of the RPM. The ultimate goal is to develop a helpful tool that, starting from the knowledge of the real problems, effectively meets the real needs of survey managers.

VI. CONCLUDING REMARKS

43. To establish an RPM for cross-sectional business surveys is an ambitious project because in spite of the long tradition of the E&I discussion it is still difficult to present the manifold approaches in a coherent and concentrated way.

44. The response to the state-of-the-art survey is encouraging and the inputs from the survey will have a major influence on the content. Though not yet complete, the survey results are also useful to have a first picture of the state-of-the-art from a methodological and operational point of view in the business survey area. In particular, it highlights some relevant open problems and areas for further development, among others the need for standards, general guidelines or recommended practices such as the one to be developed in the EDIMBUS project, to support and encourage the progressive standardization of E&I processes from both a methodological and an operational point of view.

45. Hopefully the evaluation of the draft RPM will again collect a lot of helpful feedback. To this aim, much care will be devoted to the revision/evaluation phase of the project, since there is strong awareness that the effectiveness and the relevance of the produced RPM highly depends on its applicability and usefulness in the different real contexts.

REFERENCES

Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995) *The methodology of editing and imputation by qualitative variables implemented in SCIA*. Technical report. ISTAT.

Eurostat (2003). *Definition of Quality in Statistics*. Eurostat Working Group on Assessment of Quality in Statistics, Luxembourg, 2-3 October.

Eurostat (2001). *Summary report from the Leadership Group (LEG) on Quality*. 31 July.

FCSM (2002) *Data Editing in Federal Statistical Agencies*, Statistical Policy Working paper 18. Methodology Reports.

Granquist L., Arvidson G., Elffors C., Norberg A., Lundell L-G (2002). *Guide Till Granskning* (in Swedish). Statistics Sweden.

National Centre for Education Statistics (2001). *Statistical Standards*.

Scarrott C. (2005). *Feasibility Study: A Review of Selective Editing*, Technical report, University of Canterbury, New Zealand (http://www.stats.govt.nz/NR/rdonlyres/4EF0E49F-DF8D-4B46-B9FD-A069D91F5A91/0/AReviewofSelectiveEditing.pdf)

Statistics Canada (2003). *Statistics Canada Quality Guidelines*. Fourth Edition. Ottawa.

Statistics Sweden (2004). *Design your Questions Right.* September 2004.

United Nations (2000). *Glossary of Terms on Statistical Data Editing*. Geneva.

United Nations (2000). *Evaluating the Efficiency of Statistical Data Editing: General Framework.* Geneva.

Williams H., Kennedy E. and Siu-Ming Tam (2000). Statistical Data Editing Design Principles for Households Surveys, With Applications to a Computer Assisted Interviewing (CAI) Environment", paper presented at *The Methodology Advisory Committee* meeting, Australian Bureau of Statistics, 14 July 2000.

Appendix: the questionnaire glossary

Error types

Outliers
An *outlier* is a data value that lies in the tail of the statistical distribution of a set of data values. The intuition is that outliers in the distribution of uncorrected (raw) data are more likely to be incorrect. Examples are data values that lie in the tails of the distributions of ratios of two fields (ratio edits), weighted sums of fields (linear inequality edits), and Mahalanobis distributions (multivariate normal) or outlying points to point clouds of graphs.

Influential error
Influential errors are those errors that have a high impact on aggregate estimates.

Item non-response
Item non-response occurs when a respondent provides some, but not all, of the requested information, or if the reported information is not usable.

Random error
Random errors are errors that exist randomly throughout the data that do not have a tendency to consistently change the estimates in any one direction. They primarily arise due to in-attention by respondents, interviewers and other processing staff during the various phases of the survey cycle.
This looks like a (normal) error with expectation 0. There is another type of random (not systematic) error with expectation *not* 0.

Residual error
Residual errors are errors that are detected during the final stage of editing and imputation, after all records have been processed individually and preliminary estimates of aggregates have been obtained. During this final stage, the validity of this preliminary output is checked. Implausible results at this stage may lead to again checking individual (influential) records for not previously detected errors or errors possibly introduced by the editing and imputation activities.

Systematic error
Errors reported consistently over time and/or between responding units (generally undetectable by editing). A phenomenon caused either by the consistent misunderstanding of a question on the survey questionnaire during the collection of data or by consistent misinterpretation of certain answers in the course of coding. The systematic error does not lead necessarily to validity or consistency errors but always seriously compromises statistical results.

Editing

Editing
Data *editing* is the application of checks that identify missing, invalid or inconsistent entries or that point to data records that are potentially in error.

Macro-editing.
A macro-edit detects individual errors by checks on aggregated data, or checks applied to the whole body of records. The checks are typically based on models, either graphical or numerical formula based, that determine the impact of specific fields in individual records on the aggregate estimates.

Graphical editing
Using graphs to identify anomalies in data. While such graphical methods can employ paper, the more sophisticated use powerful interactive methods that interconnect groups of graphs automatically and retrieve detailed records for manual review and editing.

Selective editing
Editing all records manually is very costly. Selective editing is an approach where only a subset of the records are clerically edited. Ideally this subset is chosen as the records that contain errors that can have substantial impact on publication totals. Detecting these influential errors is therefore an important part of selective editing.

Deterministic checking rules
A deterministic checking rule determines whether data items are incorrect with probability of 1. Example: “If Age=5 and Status=mother then Age=.”, where “Age=.” means that the Age value of 5 is determined to be incorrect.

Minimum change error localisation (Fellegi-Holt principle)
An error localisation method based on the Fellegi-Holt principle. For records that fail some edit constraints, such a method will designate a number of fields in that record as erroneous. These erroneous fields are chosen such that:
1) by changing the values of the erroneous fields it is possible to let the record satisfy all edit constraints;
2) the number of erroneous values is the smallest number for which it is possible to create a record that satisfies all edit constraints by changing erroneous values.

<u>Imputation</u>

Imputation
Imputation is the process used to resolve problems of missing, invalid or inconsistent responses identified during editing. This is done by changing some of the responses or missing values on the record being edited to ensure that a plausible, internally coherent record is created.

Deductive imputation
An imputation rule defined by a logical reasoning or a mathematical relation. Example: it might occur that some items are supposed to add to a total. If only one item in the sum has to be imputed (because it is missing or determined to be in error) then its value is uniquely determined by the values of the other items.

Model based imputation
Use of averages, medians, regression equations, etc. to impute a value.

Hot-deck imputation.
A donor questionnaire is found from the same survey as the questionnaire with the missing item. This donor questionnaire is used to supply the missing value. The "nearest neighbour" search technique is often used to expedite the search for a donor record. In this search technique, the deck of donor questionnaires comes from the same survey and shows similarities to the receiving record, where similarity is based on other data on the questionnaire that correlates to the data being donated.