Indicators with respect to process model for editing in Statistic Finland

Pauli Ollila[1], Outi Ahti-Miettinen[2] and Saara Oinonen[3]

[1]Statistics Finland, e-mail: pauli.ollila@stat.fi
[2]Statistics Finland, e-mail: outi.ahti-miettinen@stat.fi
[3]Statistics Finland, e-mail: saara.oinonen@stat.fi

Abstract

Indicators developed for editing models have two main functions. On one hand, they are used to control error identification and correction actions. On the other hand, indicators can analyse effect of the editing actions on the quality of data at the different stages of the editing model and estimate the overall quality of the final data. We divide indicators into three groups: Indicators of raw data, indicators that relate to the error identification and indicators that relate to error correction. In this paper, we discuss what kind of indicators we need and in which stages of the editing model they should be computed. We also make an overview of the demands of the ESS standard of quality reporting in editing and outline recommendations for what indicators to use.

Keywords: Editing; Imputation; Indicator; Process model

1 Introduction

Statistics Finland has carried out an editing project, whose main task was to survey editing and imputation practices at Statistics Finland and produce Editing Model for Statistics Finland. Statistical editing refers to activities by which statistical data are checked to detect of missing, invalid or inconsistent values. In some definitions editing includes also error correction. Imputation implies that missing or erroneous (e.g. edit failures) values for variables are replaced with imputed values, which have to be as correct as possible in regard to the true but unknown values. Imputation methods vary considerably depending on the type of data set, its scope and the type of missingness of data. In practice, editing and imputation are carried out in subsequent phases. Statistical editing is needed at each phase, starting from the planning of a data collection up to the formation of a data file, data processing and analysis. (Statistics Finland, 2007)

Editing model for Statistics Finland contains three main phases. The first phase of the model contains planning of editing process, descriptive analysis of data and error diagnostics. The second phase is the editing process in which the error identification and corrections are made. The last phase of the model evaluates the quality of both the editing and imputation process and the processed data. The editing model should be part of every statistical production process. By creating systematic process model for editing, Statistics Finland expects a clear improvement in efficiency of statistic process, but also improvement in quality and transparency of data. A big part of achieving the transparency of data for users and systematic quality control is to creating a list of indicator to be published.

In the Statistics Finland's editing model, actions for editing and imputation of statistical data are presented in process form. Indicators are involved in the editing model in two ways: On the one hand, they control error identification and correction actions and their effects on the data; on the other hand, they evaluate the development of the quality of data in different stages of editing process and overall quality of the final data.

We divide indicators for statistical data editing into three groups. In section 2 in this paper, we discuss indicators related to descriptive analysis of raw data. Indicators that relate to error identification are presented in section 3. Moreover, indicators that relate to the error correction are discussed in section 4. In section 5, we make an overview of the demands of the ESS standard of quality reporting in editing and outline recommendations for what indicators to use.

Indicators presented in this paper are collected from several sources: EUREDIT-project (EUREDIT / Ray Chambers, 2004), EDIMBUS-project (EDIMBUS, 2007), Eurostat standards for quality reports (EUROSTAT, 2009a), Quality Guidelines for Official Statistics by Statistics Finland (Statistics Finland, 2007) and functional report of BANFF-program by Statistics Canada (Statistics Canada, 2007).

2 Indicators of raw data

Indicators that describe raw data give information about errors on data, their effects on results and variables' or subgroups' significance on results. This initial editing and imputation is in the beginning of editing process. It may include observing the data in varying ways: tabulations, statistics calculations, distributional evaluation and data visualisation. Indicators describing the raw data are calculated in this phase.

Let Y be obtained data matrix (n x p), where n is number of observations and p is number of variables. For every variable y_j its observation y_{ij} ($i = 1, ..., n; j = 1, ..., p$) may have a value or it can be missing. The response indicator r_{ij} for observation y_{ij} has value of 1 if unit i has value on variable j. It has value of 0 when observation y_{ij} is empty. If value is missing due to structural reason, it should be marked. To identify structural missing values, we have factor b_{ij} that has value of 0 when y_{ij} is missing due to structural reason and 1 otherwise. Some indicators presented below are referring to auxiliary variable x. Variables x may originate from obtained data or from other sources. Survey weight for observation i is denoted by w_i.

First group of indicators we present include **indicators for missingness of data.** From two types of missingness (unit and item nonresponse) only item nonresponse is considered. Response rate (1), weighted response rate (2) for variable y_j and weighted response rate for variable y_j proportioned with auxiliary variable x (3) are defined in following table 1.

Table 1: Basic measures for missingness of data.

$\frac{\sum_i r_{ij}}{n}$ (1)	$\frac{\sum_i w_i r_{ij}}{\sum_i w_i}$ (2)	$\frac{\sum_i w_i x_i r_{ij}}{\sum_i w_i x_i}$ (3)

Weighted response rate (2) evaluates proportion of responses in variable j when whole population is inspected. This should be noted especially when data is stratified or quota sampling is used with deviant proportions or when calibration of weights are used. Weighted response rate for variable y_j proportioned with auxiliary variable x (3) gives information about effect of nonresponse to aggregates of variable y_j when x is correlated with y_j. Indicators for response rates, such as item nonresponse rate (4) and full response rate(5), are defined in table 2.

Table 2 : Indicators for response

$\frac{\sum_i(1-\prod_j r_{ij})}{n}$	(4)	$\frac{\sum_i(\prod_j r_{ij})}{n}$	(5)

Proportion of missing values in all variables (6) and average proportion of missing values (7) are presented in table 3.

Table 3: Indicators for measuring the proportion of missing values

$\frac{\sum_j(1-r_{ij})}{p}$	(6)	$\frac{\sum_i\sum_j(1-r_{ij})}{np}$	(7)

Proportion of missing values in all variables (6) is an unit-specific indicator of response. Proportion of missing values is usually reasonable to calculate in relation to some subset of variables $s < p$.

The effect of missing values on data can be evaluated with auxiliary variable x, which is available for all observations. In some cases it may be necessary to produce item nonresponse adjusted weights w_{ij}^* for variable j, where $\sum_i w_{ij}^* r_{ij} = \sum_i w_i$. Then it is possible to use indicators that evaluate the effect of missing values with auxiliary variable as presented in table 4.

Table 4 : Indicators evaluating the effect of missing values

$\frac{\sum_i w_{ij}^* r_{ij} x_i}{\sum_i w_i x_i}$.	(8)	$\frac{\sum_i w_{ij}^* r_{ij} x_i - \sum_i w_i x_i}{\sum_i w_i x_i}$.	(9)

Ratio of item nonresponse estimated and survey weight estimated totals of *x* (8) and proportion of variation of item nonresponse estimated and survey weight estimated totals of *x* (9) are defined above. If x and y are correlated, that ratio estimates the proportion of change that item nonresponse has on the total of variable y_j.

Next we will discuss on **indicators for impact of observation**. Variables that have skew distributions may have values, which have a large impact on results. Survey weights need also be taken into account when assessing the impact of observations. These indicators are more relevant when calculated from certain subgroup rather than from whole data, since the impact is easier to notice. It is also useful to specify essential subgroup by contents which significance on result can be calculated. Indicators for significance calculation are presented in table 5:

Table 5 : Significance indicators

$\frac{y_{ij}}{\sum_i y_{ij}}$	(10)	$\frac{\sum_{i\in q} y_{ij}}{\sum_i y_{ij}}$	(11)	$\frac{w_i}{\sum_i w_i}$	(12)	$\frac{w_i y_{ij}}{\sum_i w_i y_{ij}}$	(13)

Significance of each individual observation y_{ij} in sum of variable y_j (10) and similarly significance of each observation y_{ij} subgroup q (11) are typical indicators for significance examination. Significance of each weight w_i in sum of weights (12) identifies units that may have a large impact on results through survey weights. Significance of each weighted observation $w_i y_{ij}$ in estimate of total variable y_j (13) reveals the true impact of observation to total estimate. There are also some other indicators related to observation and

weight significance. It is possible to calculate total estimate from subgroup that has one observation unit removed. This describes the effect the removed unit has on the estimate of total. These calculations require computing adjusted survey weights w_i^- with one unit removed accordingly.

Estimators sensitivity to unit *i*

$$c\left(\hat{\theta} - \hat{\theta}_{(i)}\right) \tag{14}$$

describes the change in estimate $\hat{\theta}$ when observation *i* is omitted. Term *c* standardizes the result and according to the EDIMBUS report, it can be formed as a mean of estimators on removed units $i: = \sum_i \hat{\theta}_{(i)}/n$. These estimates can be examined often through so-called sensitivity curve and it is linked with outlier evaluation.

Indicators presented above are mainly simple functions proportioned to total estimates. Significance evaluation has been crucial part of editing in recent years. In terms of selective editing there is several score functions available and most of them include reference value. These score functions will not be presented in this paper.

3 Indicators related to identifying of errors

Indicators concerning error recognition have two aims: 1) they describe the amount of errors in variables and observations, 2) they describe the effectiveness of error identification procedures. Indicators describing the efficiency of error detection are not discussed in this paper. Defects on data are evaluated in data studies and editing process phase on editing model. Error identification phase includes actions that are designed to identify errors so that they can be individualized on variables and observations.

The most common practise to notice error is to use an edit rule, which flags observation to be either error or error suspicion. Some error identification actions may include data processing with functions or modelling. Some errors are detected from results of macro editing. Visual examination on both observation and result levels is useful. It is essential for indicator calculations that identified errors and error rules used to identification can be tracked by flagged observations.

Error identification indicator f_{ij} for unit *i* on variable *j* has value of 1 if observation is detected to be erroneous on error identification process. Otherwise, it has value of 0. It is also possible to add denotation *l* for the parameter to identify method used on error detection f_{ijl}.

3.1 Indicators for error identification on different levels

Table 6 : Indicators for error identification: Variable level

$\frac{\sum_i f_{ij(l)}}{n}$	(15)	$\sum_i \sum_l f_{ijl}$	(16)

Indicators for error identification are needed in different levels of data. **On variable level**, error degree of variable y_j on data (by identification method *l*) (15), presented in table 6, measures the error identification rate of variable y_j. Adding the information of error identification method *l* describes the sensitivity of the method proportioned to all errors on variable. It is important to notice that error identification may include false alarms, so great number of identified errors may not imply a good error identification method. On the opposite, one sole error may be significant if the magnitude of error is exceptional. This is why it is

sometimes useful to calculate the amount of error identifications on variable over all identification methods (16) for comparison. It also controls the operations of edit rules.

Indicator for error identification on observation level: Observation *i* is erroneous by edit rule *l* if at least one parameter f_{ij} $(j = 1, \ldots, p)$ has value of 1. Then error parameter e_{il} for observation *i* combined with edit rule *l* has value of 1 and otherwise value of 0. Variable amount p can be replaced with subset $(s < p)$ that are included on error identification. On the observation level of the data the proportion of error occurrence on all variables (by error identification method *l*)

$$\frac{\sum_j f_{ij(l)}}{p} \tag{17}$$

measures overall quality of observation.

Table 7 : Indicators for error identification: Data level

$\frac{\sum_i e_{il}}{n}$ (18)	$\frac{\sum_i (1 - \prod_l (1 - e_{ij}))}{n}$ (19)	$\frac{\sum_i w_i y_{ij} f_{ij}}{\sum_i w_i y_{ij}}$ (20)

Standard indicator for error identification **on data level,** presented in table 7, is error identification rate for edit *l* (18). If there are several edit rules in use it is possible to define general rate of error detection for all error detection methods (19). Proportion on error detection (20) describes the ratio of erroneous variables from total estimate.

4 Indicators related to correction actions

Indicators associated with error correction describe 1) quality of data after error correction; 2) amount of error correction in variables/units/data; 3) effect of error correction on results. After error identification and nonresponse assessment, we have gained information on to which observations and variables we should focus actions on. Actions might include inquiring the correct value from respondent, searching for right value from other sources, cold-deck imputation, imputation from other data source or imputation based on statistical methods. Let the inserted or imputed value be denoted as $\hat{y}_{ij}$. After different corrective actions we get the final data. It is quite common, that the edits on data have not been recorded in any ways and the only method to evaluate the changes is usually to compare the raw and final data.

Indicators for error correction can describe the final data and its flaws and edits. They remark on the broadness of edit actions, the amount of edits caused by error identification and influence of error identification to edit actions. In some cases, they show impact of faultiness on estimates. For calculating such indicators, it is essential to tag the information needed on observational level during the editing process.

4.1 Indicators measuring the proportion of missing values on data after error correction

The item response indicator $\hat{r}_{ij}$ takes the value of 1 if observation *i* on variable *j* has value *after E&I-actions*. The value can be the original y_{ij} or corrected $\hat{y}_{ij}$. Correspondingly $\hat{r}_{ij}$ takes value of 0 when y_{ij} is missing. Survey weight for observation *i* is denoted as w_i. Some standard indicators are presented in table 8.

Table 8 : Response and inconsistency rates

$\frac{\sum_i \hat{r}_{ij}}{n}$ (21)	$\frac{\sum_i w_i \hat{r}_{ij}}{\sum_i w_i}$ (22)	$\frac{\sum_i f_{ij} \hat{r}_{ij}}{n}$ (23)

The item response rate after E&I-actions (21) and weighted item response rate after E&I-actions (22) evaluates remaining missingness of variable *j* at population level. As with raw data, it is important also with final data to calculate weighted indicators if the data is stratified or quota sampling is used with deviant proportions or when calibration of weights is used. Usually it is recommended to calculate response rates with and without weights. The rate of inconsistent data (23) describes how much the variable *j* has values as a result of error detection focused on the variable *j*.

Table 9 : Observation specific indicators for missingness after imputation

$\frac{\sum_j (1 - \hat{r}_{ij})}{p}$ (24)	$\frac{\sum_i f_{ij} \hat{r}_{ij}}{p}$ (25)

As with the raw data, the proportion of missingness after imputation (24), presented in table 9, might be more informative to calculate in relation to some sensible subset *s* instead of all variables. Inconsistency proportion (25), presented in table 9, describes the portion of variables that has values as a result of error detection. The variable set of p can be substituted with subset of variables *s* that are included in error detection method.

Table 10 : Data specific indicators for missingness after imputation

$\frac{\sum_i (1 - \prod_j \hat{r}_i)}{n}$ (26)	$\frac{\sum_i (\prod_j \hat{r}_i)}{n}$ (27)	$\frac{\sum_i \sum_j (1 - \prod_j \hat{r}_{ij})}{np}$ (28)

Proportion of item nonresponse after imputation (26) and proportion of full response after imputation (27), presented in table 10, are exclusive classes, similarly to the situation with raw data, but both forms can be useful depending on situation. The remains of non-structural item nonresponse after correction actions indicate that it is not possible to form full response by current standards or data management system is enabled to allow some absence on observations. Mean proportion of missingness (28) describes how much missing values are included on observations on average.

4.2 Indicators describing error correction actions

Correction of an error is a modification of data and it is represented with variable $I(\hat{y}_{ij} \neq y_{ij})$, which takes value of 1 when value of raw data differs from value that is corrected and value of 0 otherwise. Parameter b_{ij} for structural missingness on raw data is included in some indicators and it's equivalent on corrected data is defined as $\hat{b}_{ij}$. It is highlighted in EDIMBUS-report that these indicators need to be calculated with and without weights. We start introducing **indicators for error correction actions on variable level** in table 11.

Table 11 : Indicators for error correction actions on variable level

$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n}$ (29)	$\frac{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij})}{\sum_i w_i}$. (30)	$\frac{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij})\hat{y}_{ij}}{\sum_i w_i \hat{y}_{ij}}$ (31)

Edit rate (29) describes the quantity of edited values on specific variable. Also weighted edit rate (30) is defined. If we are only interested in imputed edits, terms are correspondingly imputation rate and weighted imputation rate. Edit ratio (31) describes the effect of edits proportioned to total estimate. Modification rate (32) is defined as

$$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})\hat{b}_{ij}}{n}. \quad (32)$$

Term $(r_{ij}b_{ij})\hat{b}_{ij}$ takes value of 1 only when variable has value given on observation$(r_{ij} = 1)$, it have not been structurally missing on raw data $(b_{ij} = 1)$ and it is not structurally missing on corrected data either $(\hat{b}_{ij} = 1)$. Similar indicators are net edit rate with term $(1 - r_{ij}b_{ij})\hat{b}_{ij}$, which takes account cases where response is formed for missing value, and cancellation rate with term $(r_{ij}b_{ij})(1 - \hat{b}_{ij})$, which takes account only removals of exiting values. These indicators can also be calculated with weights w_i.

If there is demand for indicators for specific correction method m, it is possible to calculate indicators above with a method specific parameter I_m as is presented in table 12.

Table 12 : Edit proportions and overall edit rate

$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})I_m}{\sum_i I(\hat{y}_{ij} \neq y_{ij})}$ (33)	$\frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})}{p}$. (34)	$\frac{1}{p}\sum_j \left(\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n}\right)$. (35)

Proportion of correction method m from all correction methods used for variable y_j (33) can also be calculated with the weights w_i. Indicators presented above have also their corresponding versions in observation level, describing the change within the observation from different starting points, for example edit proportion in observation level (34). Modification proportion, net edit proportion and cancellation proportion can be calculated by using terms $(r_{ij}b_{ij})\hat{b}_{ij}$, $(1 - r_{ij}b_{ij})\hat{b}_{ij}$ and $(r_{ij}b_{ij})(1 - \hat{b}_{ij})$ correspondingly. Set of variables p can also be substituted with variable subset s. There are also similar indicators for describing edits on data level, for example overall edit rate (35) presented in table 12. Overall modification rate, overall net edit rate and overall cancellation rate can be calculated correspondingly. In addition, weighted versions are available by adding weight parameter.

4.3 Indicators for implications of error identification

Not every error identification results in an error correction. On next table 13, we present some standard indicators for implications of error detection.

Table 13 : Indicators for implications of error detection

$\frac{\sum_i f_{ij} I(\hat{y}_{ij} \neq y_{ij})}{\sum_i f_{ij}}$ (36)	$\frac{\sum_i e_{ij} I(\hat{y}_{ij} \neq y_{ij}\|l)}{\sum_i e_{ij}}$ (37)	$\frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i}$ (38)	$\frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i y_{ij}}$ (39)

It is possible to calculate edit rate proportioned to errors detected for variable *j* (36). It describes amount of corrections resulted from error detection proportioned to all errors included in variable. Such edit rate is also possible to define for specific detecting method *l* respectively, assuming that there is certainty that an edit resulted only from inspected method. For this reason, it is necessary to define conditional indicator $I(\hat{y}_{ij} \neq y_{ij}|l)$, which takes value of 1 only when value modification has resulted from error detection method *l*. Now we can define rate for error corrections in variable *j* caused by error detection method *l* (37) and it is possible to extend for data level.

There are also defined indicators that describe the impact of corrections and error quantity on data. Error corrections have effect on results and this effect can be evaluated with indicators that are based on value difference $\hat{y}_{ij} - y_{ij}$ or estimate difference $\hat{\theta}(\hat{y}_1, \dots, \hat{y}_i, \dots, \hat{y}_n) - \hat{\theta}(y_1, \dots, y_i, \dots, y_n)$. If there is information on real values y_{ij}^* it is possible to evaluate defects of error correction with similar indicators by replacing the raw data values y_{ij} with real values y_{ij}^*. Some common variable level indicators for measuring edit impact are weighted average edit impact (38) and weighted relative average edit impact (39) presented in table 12. Weighted α-relative edit impact

$$\frac{\left(\frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i\, y_{ij}}\right)^{1/\alpha}}{\sum_i w_i y_{ij} / \sum_i w_i} \tag{40}$$

is average correction impact modified with α proportioned to variables weighted mean $\bar{y} = \frac{\sum_i w_i y_{ij}}{\sum_i w_i}$. With α it is possible to regulate examination and when $\alpha = 1$ indicator results in weighted average impact of edits proportioned by variable *j*. Same examination can be done for correction defects.

Total impact of edits can be calculated as difference of estimates

$$\hat{\theta}(\hat{y}_1, \dots, \hat{y}_i, \dots, \hat{y}_n) - \hat{\theta}(y_1, \dots, y_i, \dots, y_n)\,. \tag{41}$$

Sensitivity of the parameter estimate $\hat{\theta}$ for edits is denoted as

$$c\left(\hat{\theta}(\hat{y}_1, \dots, \hat{y}_i, \dots, \hat{y}_n) - \hat{\theta}(\hat{y}_1, \dots, \hat{y}_{i-1}, y_i^*, \hat{y}_{i+1}, \dots, \hat{y}_n)\right) \tag{42}$$

where *c* is a suitable standardization constant (usually *c* = 1). It describes the change on estimate that is based on corrected values when one corrected value is replaced by real value. Edit error rate

$$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij}) I(\hat{y}_{ij} \neq y_{ij}^*)}{\sum_i I(\hat{y}_{ij} \neq y_{ij})} \tag{43}$$

describes amount of false edits proportioned to all edits.

If all edits under examination are done by imputation, we can refer to indicators above as *weighted average imputation impact* (38), *weighted relative average imputation impact* (39) etc.

5 Recommendations for use of indicators

Quality standards of Eurostat: Eurostat has determined standards for quality reports for European Statistical System. The guidebooks of standards describe different dimensions of quality and in this paper, we will refer to the following demensions: measurement errors, nonresponse errors, processing errors in micro-data and imputation.

A measurement error is the discrepancy between the observed value of a variable provided by the survey respondent and its underlying true value. Eurostat reports discuss the reasons for measurement errors, how they are formed and what methods exists to investigate measurement errors. Error identification offers information on certain and suspected errors, and indicator for *error identification rate for edit rule l* (18) is mentioned. It is also recommended to use this indicator to examine crucial subgroups. For bias caused by measurement errors, they offer different evaluation studies that are mainly related to assessing the questionnaire and the interview situation. One way to estimate bias is to calculate results from original data and final data and compare them.

Item nonresponse on variable is crucial of nonresponse errors according to Eurostat report. Indicator for *item response rate* (1) is mentioned and optional indicator for *weighted item response rate* (2) is offered. It is important to specify the essential variables for which the response rates are calculated and consideration must be used to decide whether to use weighted or unweighted indicators. Examination in relevant subsets of units is also highly recommended. The effect of nonresponse can be tested with values that are available for all responded units by comparing full response estimate to the estimate that notices nonresponse. Indicators for nonresponse estimates are mentioned for example *Ratio of item nonresponse estimated and survey weight estimated totals of* x (8).

Processing errors in micro-data imply on data recording, editing and sometimes on data treatment and imputation. By Eurostat report, it is essential to explain the extent and impact of processing errors if they are significant. Calculating results from original and corrected data and comparing the results is mentioned as a simple testing method. This provides the total net effect of editing. The amount of imputation, or generally error correction, can be evaluated by calculating *edit rate* (35).

Recommended and considered indicators for different purposes: Concepts used in indicator calculations are not always unambiguous. In next section we will discuss on different situations and problems related to these concepts. ***Target variables:*** Statistic producer must decide the group of variables from which the indicators are calculated. Not all indicators can be calculated for categorical variables. Some variables are too insignificant for detailed examination. Therefore, it is essential to define certain target variables of which indicators are calculated. ***Subgroups:*** It is possible that indicators perform better on subgroup level than in complete data. Knowledge of contents and overall experience are needed to define such subgroups. ***Raw data:*** There are several indicators targeted for raw data examinations, but it is not always obvious what is considered as raw data. For the basis of the editing model, raw data is defined as data that includes all needed material combined. It is possible that initial error check has been done to some parts of data when it has been received and before the raw data has been formed. In some statistics editing is greatly emphasized on data reception phase. All these initial edit actions must be regarded when indicators from raw data are calculated.

Structurally missing values are problematic with indicators that describe item nonresponse. In some cases, this can be taken into account with separate parameter. If structurally missing values exists in raw data, they must be defined and their effect on indicators should be eliminated. ***Error identification:*** Different methods for error identification are categorized in different phases in editing model. For all cases, it is not possible to define exactly the method of error identification due to unsystematic detection. There are indicators that are used to identify error detection methods but they are mainly designated for edit rules. Sometimes error identification method is not necessarily important to notice. Crucial identification methods should be selected based on criteria of the editing model and indicators should be targeted on those identification methods. In multiple error cases, it is not apparent whether to tag all errors on the value or just the error that caused disqualification.

Error correction: Main problem with error corrections is what corrections are included in indicator calculations. Original value can be changed unambiguously because of logical inspection or inquiry of the correct value. These cases might not be necessarily involved in indicator calculations. When error correction

indicators are concerned usually only imputation is mentioned. Generally, only significant error correction methods are included in indicator calculations. There might be problems with individualizing correction methods and with serial corrections. Some balance adjusting edits also modify values that are not erroneous.

6 Discussion

In this paper we have collected indicator in respect to editing process of official statistics. We have classified these indicators in three groups according to their functions: Indicators of raw data, indicators that relate to the error identification and indicators that relate to error correction. These groups relate to three different phases of editing model. The number of indicators presented in this paper is quite substantial considering how diverse statistics production processes are. Not all indicators are suitable for every type of statistics. Hence, consideration on which indicators to be applied in each process is essential. Many questions related to choices of subgroups or variables from which indicators are calculated need solid substance knowledge. Some indicators presented are important but suitable only for specific situations. There for, it is not possible to define a detailed list of indicators to be published with all statistics. Some standard indicators for editing process should always be computed. Indicators measuring missingness in data (presented in section 2.1) are valuable tools for statistics production process by describing the coverage of data in each stage of process. Then, if any editing actions are done, the user of the final product should have access to information on edit rates of the data (presented in section 4.2). Altogether, indicators are essential part of editing process as they provide information on quality of data, results and editing process in general.

References

EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys. Project report. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

EUREDIT / Ray Chambers (2004). Evaluation Criteria for Statistical Editing and Imputation, Project report.

EUROSTAT (2009a). ESS Standard for Quality Reports, Luxemburg.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

EUROSTAT (2009b). ESS Handbook for Quality Reports, Luxemburg.
http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/EHQR_FINAL.pdf

OLLILA, P. (2012). Raaka-aineiston,editoinnin ja imputoinnin indikaattorit (luonnos). Unpublished memorandum. Statistics Finland.

OLLILA, P. and ROUHUVIRTA, H (2012). Editoinnin prosessimalli (luonnos). Unpublished memorandum. Statistics Finland.

STATISTICS CANADA (2007). Functional Description of the Banff System for Edit and Imputation, Ottawa.

STATISTICS FINLAND (2007). Quality Guidelines for Official Statistics 2nd Revised Edition, Helsinki.
http://www.tilastokeskus.fi/meta/qg_2ed_en.pdf