An analysis of editing strategies for mixed-mode establishment surveys

Mark P. J. van der Loo

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussionpaper (08004)

Statistics Netherlands

Voorburg/Heerlen, 2008

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/'06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/'04–2005/'06	= crop year, financial year, etc. 2003/'04 to 2005/'06 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg

second half of 2008:
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

Information
Telephone .. +31 88 570 70 70
Telefax .. +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
http://www.cbs.nl

ISSN: 1572-0314

6008308004 X-10

An analysis of editing strategies for mixed-mode establishment surveys

Editing practices in mixed mode Computer Assisted Self Interviewing (CASI) and Paper And Pencil (PAP) business surveys are investigated. A comparison between policies at Statistics Netherlands and other national statistical institutes is made. The use of CASI with built-in edits is addressed in the context of questionnaire design and respondent's reactions, the relation with post-processing of data and the comparability of CASI with PAP data. Directions for future research and survey planning are indicated in the conclusions and recommendations section.

Keywords: Mixed mode, editing, electronic surveys, quality, quality measures

Contents

1 Introduction

The introduction of electronic survey methods opens up the opportunity to have (part of) the data checked and edited by the respondent. Furthermore, using electronic questionnaires, one can store audit trails which gives insight into the behaviour of individual respondents, and can possibly be of use in post-processing of data. In many cases, business surveys will be done in mixed mode, such that the respondent can choose between completing a paper or electronic form. This report will comment on the consequences for the editing process when such a mixed-mode design is employed.

This report is part of a wider research programme which intents to clarify Mixed Mode effects throughout the statistical process. The program itself is motivated by Statistics Netherlands' policy of emphasizing electronic data gathering methods.

1.1 Electronic surveys

With electronic surveys, we here mean a (sampling) survey where a respondent fills in a questionnaire using a computer. Automated forms of registration such as XBRL (eXtendable Business Reporting Language) do not belong to this category. Questionnaires which are offered in electronic form are called Computerized Self-Administered Questionnaire (CSAQ) or Computer Assisted Self Interviewing (CASI). Here, the term CASI will be employed, since it seems more common in the literature.

Two forms of CASI are distinguished, based on the way the questionnaire is offered to the respondent.

- **Online survey** Also: Computer Assisted Web Interview (CAWI). The respondent completes a questionnaire using a browser via a (secured) internet connection.
- **Off-line survey** The respondent installs some dedicated software or a file on his computer which can be used to complete the questionnaire. The file or program can be offered to the respondent by download (internet, ftp), by e-mail, or by mail as CD-ROM or Diskette. Completed questionnaires can be returned by (secure) upload, (encrypted) electronic mail or by ordinary mail, either by sending a disk or by printing a completed form.

1.2 Scope of this report

In this report editing strategies for mixed mode business surveys are treated. Results from literature on editing in mixed-mode CASI/PAP (PAP = Paper

And Pencil) survey designs will be summarized and compared with practises at Statistics Netherlands. In Section 2, questions and experiences regarding questionnaire design and built-in edits will be addressed. In Section 3 the consequences for post-processing of data in a mixed-mode setting will be discussed by a systematic discussion of two edit procedures used at Statistics Netherlands. A small mathematical analysis of possible edit failures in PAP/CASI questionnaires is included. Section 4 is dedicated to a discussion on data quality and comparability of quality between modes and questionnaires. Every area covered in the sections leads to specific remaining questions and conclusions, which will be summarized in Section 5.

The objective is to identify possibilities and pitfalls when moving from PAP to a mix of PAP and CASI practice at Statistics Netherlands. The relation between the current situation and the wish to use mixed-mode survey approaches in business surveys is discussed. A discussion on data quality and comparable data quality metrics for different modes is also included.

2 Editing at the source

By checking data at the time it is entered by the respondents, and asking respondents to fix errors, statisticians hope to improve the data quality at record level. By warning the respondent when erratic, inconsistent, implausible or no data is entered, it might be possible to improve the validity, consistency, accuracy and completeness of entered data. Over the last two decades or so, statistical institutes have gained experience with electronic survey methods and the possibilities of building edits into the questionnaire.

In the last years, efforts have been made to create generic (web-based) applications for sample surveys in several institutes (Zeila 2005; Kurkowski 2005; Koller 2005), and issues concerning "moving editing strategies to the source in establishment surveys" have been addressed by several authors (Anderson *et al.* 2003; Arbuez *et al.* 2005; Laroche 2005; Cohen 2003).

As a consequence, questionnaire builders are confronted with design problems mostly regarding questions of building in as many edits as possible while keeping the questionnaire as user-friendly as possible. A general trend that can be observed however, [See *e.g.* Nichols *et al.* (2005)] is that as more experience is gained, the number as well as the complexity of built-in edits increases.

In this chapter built-in edits are defined and discussed (Sect. 2.1). Next we report on experiences of foreign statistical institutes (Sect. 2.2) found in literature and compare them with policies at Statistics Netherlands (Sect. 2.3). A discussion is given in Sect. 2.4.

2.1 Built-in edits

An edit is a rule which determines whether a record contains invalid or suspect data. An overview of the types of edits which have been built into business surveys by several statistical institutes is given below. They are classified according to which aspect of record quality they are aimed to improve.

2.1.1 Validity

These are edits which verify that the entered data lie within a certain range of values. Validity checks are amongst the most commonly implemented checks and can often be forced from the user, for example by accepting only numeric keystrokes when a numeric field is selected. Edits that have been reported to be built into business surveys include alphanumeric checks (Weir 2003; Nichols *et al.* 2005; Koller 2005), data format checks (Nichols *et al.* 2005), numerical range checks (Weir 2003; Nichols *et al.* 2005), and length restrictions (Koller 2005). Although it does not seem to be mentioned in literature, offering a list (*e.g. via* a drop-down menu) of possible values for categorical data can also be considered a range check.

2.1.2 Completeness

Completeness edits check that a certain field or number of fields are not left empty by the respondents. They have been built into many electronic business surveys, see for instance Anderson *et al.* (2003), Nichols *et al.* (2005) or Weir (2003). Completeness checks can also be seen as validity checks, since they are built-in when item nonresponse is considered invalid.

2.1.3 Consistency

Consistency edits check internal consistency of a record. Examples include conditional null-checks which define when a field may be (non-)empty (Weir 2003), common elements checks that require that a filled-in value is copied to other fields (Weir 2003), conditional edits (Anderson *et al.* 2003) and rounding tests (Anderson *et al.* 2003). A lot of business surveys include some form of automated summation or balance edits, see *e.g.* Beckler (2005), Anderson *et al.* (2003), Nichols *et al.* (2005), or Weir (2003).

2.1.4 Accuracy

The accuracy of the entered data -or how close it relates to reality- is hard to determine. However, in a statistical sense, the data can be checked for plausibility (Karr *et al.* 2005), by comparing with historical data (longitudinal edits), comparing with external standards (outlier detection) or by performing ratio checks. One example where current data are checked against historical data of the same respondent is the U.S. Census Bureau (USCB) Manufacturer's, Shipments, Inventories and Orders Survey (M3) (Anderson *et al.* 2003). The same electronic questionnaire also contains some ratio edits such as detection of unusual sales to employment ratios. Another example is the annual Survey of Industrial Research and Development of the USCB (Anderson *et al.* 2003), where as many as 23 longitudinal edits are build in which warn for extreme rise or fall in values (budgets *etc.*) with respect to the previous year.

2.2 Built-in edits as part of questionnaire design

When developing electronic questionnaires with built-in edits, questionnaire designers have to decide on the number, complexity and presentation of edit failures. When an edit is failed it can either be reported to the respondent immediately after data was entered, or later when (part of) the questionnaire is completed. Correcting the edit failure can be compulsory or optional for the respondent. Sometimes it is possible for the respondent to leave a comment or comment code without changing their input.

In a recent survey of electronic data editing practices in the US Census Bureau and the US Bureau of Labour Statistics (USBLS), Anderson *et al.* (2003) note that the number of built-in edits per questionnaire field can nowadays be as high as 2.43 in certain CASI's. Anderson *et al.* (2003), as well as Nichols *et al.* (2005) conclude that a seemingly high number of built-in edits does not necessarily lead to higher unit nonresponse, provided that the questionnaire is well designed. Although quantitative data is not often reported, several authors claim that the willingness of respondents to correct edit failures is rather high (Nichols *et al.* 2005; Anderson *et al.* 2003).

Cohen (2003) reports that in the 1997 web version of the monthly USBLS Current Employment Status survey about 40 percent of the web samples failed at least one edit check. In 88 percent of all edit failures, the respondent corrected the data during the same session. About 3 percent of the web reports still fail one edit each month after submitting. In the 1997 version only a few edits concerning validity and consistency were included. In a newer version, several plausibility edits were included, such as longitudinal edits (checking data against historical values) and range checks. The fraction of web reports failing one edit after submitting increased from 3 to 7 percent. The fraction failing records where comment codes were entered in stead of corrections increased from 6 to 14 percent.

In a usability study of the Winter heating, Fuels, and Telephone Survey by Weir (2005), respondents had the option of checking their data against a longitudinal edit at any time during data entry. It is reported that more than 99 percent of the respondents use this option. Most of them (about 87 percent point) do so after completing all the data. The other respondents check it several times during completion. In at least 56.7 percent of the cases where the edit was failed, respondents corrected the data one way or another. Finally, Weir (2005) also notes that "*the overall reaction to the new edit [...] was very positive*".

As noted above, the presentation of edit failure to the respondent, and the choices that are offered are important factors that determine the willingness of the respondent to review and/or correct the data. Based on a comprehensive study on the use of edits in electronic questionnaires in the USCB and USBLS, Anderson *et al.* (2003) and Nichols *et al.* (2005) have formulated a number of guidelines for user interface design. A complete discussion is beyond the scope of this report, but for future reference the guidelines are given below in Table 1. For a detailed discussion, the reader is referred to the mentioned references.

Table 1. Guidelines for building edits into electronic questionnaires by Nichols et al. (2005) and Anderson et al. (2003).

1. Minimize edit failures through good design.
2. Perform edit checks immediately unless for missing data or performing an inter-item [consistency] edit. Defer activating those checks. Run them either immediately before the respondent finishes the form or after all the items in the inter-item edit have been entered.
3. Research edit checks before implementing an edit that might be too strict.
4. Avoid violating user expectations for single-purpose functions. (For example: do not mix editing and submitting under one button).
5. Allow edit failure reports to be run iteratively, as expected by respondents.
6. Allow for easy navigation between an edit failure list and the associate items.
7. Clearly identify which edit failure goes with which item.
8. Include a location, a description of the problem and an action to take in the edit message.
9. Avoid jargon, be polite, use good grammar, be brief, use active voice, and use words that match the terminology used in the question.
10. Prior to implementation, cognitively test any edit messages offering a solution.
11. Do not automatically erase data that fail an edit check.
12. Inform respondents about the policy for submitting data with unresolved edit failures.
13. Give the respondent as much control as possible.
14. Use newly created icons with caution since they do not have universal meanings. Use standard icons only for expected purposes.

2.3 Built-in edits at Statistics Netherlands

Since about two decades, a variety of software to retrieve data from respondents electronically has been developed at Statistics Netherlands. Offline variants include the BLAISE-based Electronic Data Reporter (EDR) and CBS-QUEST, CBS-IRIS, and the now obsolete EDI/EDISENT program. Online questionnaires can be generated by QUAT or BLAISE. All these programs contain some form of built-in data checks.

Whenever possible, new electronic business surveys are developed using the CBS-QUAT (QUestionnaire Application Tool) program. QUAT was developed at Statistics Netherlands as part of the PRODONNA (2005) project and is operational since February 2005. Amongst other things, QUAT is a software tool designed to build and store questionnaires in a generic way. Questionnaires built with QUAT can be exported directly to PDF (Portable Document Format), or a web page. It is also possible to generate stand-alone questionnaire applications by exporting questionnaires to BLAISE script. The BLAISE language offers the possibility to built in hard edits (BLAISE: CHECK) and soft edits (BLAISE: SIGNAL). In the future, the majority of business surveys [about 95% (PRODONNA 2005)] should be in QUAT format. At this moment about 10-15 questionnaires have been designed in, or transferred to QUAT (Mol and Groen 2006; Spaans 2007). In a feasibility study, Vonck and van der Vegt (2007) note that there are at least fifty more candidate statistics which have not been implemented in QUAT yet, for various (technical) reasons.

For the electronic versions of questionnaires QUAT offers the possibility to have the data checked during completion, and to present messages to the user. Possible edits include hard edits such as automated summation or checking for inconsistencies, as well as soft edits such as ratio edits. It is possible to build routed questionnaires, but edits based on longitudinal data are not available (yet). Upon edit failure, the respondent gets a pop-up screen, and depending on the design, the respondent can change the data, leave the data unchanged and/or type in a message. It is also possible to force format edits, for instance by ignoring all keystrokes except the numbers for numeric fields.

Two recent examples of questionnaires which are built using QUAT are the Structural Business Survey (SBS)[1] and the New Establishments Survey (NES)[2]. The former is a large annual survey with more than a hundred variables, which is implemented as an offline downloadable electronic questionnaire. The latter is a survey, held quarterly under new businesses, intended to establish economic activity, consisting of only a few variables. The New Establishments Survey was implemented as an online survey. Both surveys are implemented as mixed-mode surveys, although the policy of Statistics Netherlands is “not to send a paper questionnaire unless it cannot be avoided” (Göttgens and Baart 2006). Respondents receive a login-code and URL by mail for a download page or a questionnaire page. Only with the second reminder a paper form is send.

In the New Establishments Survey (Spaans 2006), this yields a very high electronic *versus* paper response ratio: about 87% of the response is reported electronically via the web. The total response is comparable to fully paper questionnaires, with a slightly higher overall response coming from the paper questionnaires. Since routing is built in, as well as some consistency and range

Table 2. Edits in some electronic questionnaires at Statistics Netherlands. Listed are edits in the yearly Structural Business Survey (Productie Statistiek PS), the montly Short-Term Business Survey (STS) (Kortetermijn statistiek, KS), the quarterly Vacancy Count (Vacaturetelling, VT) and International Trade (Internationale Handel, IH).

Statistic	fields	edits	edits/field	method
SBS (PS) [1]	134	range(132), format(132), sum (18)	2.1	CBS-QUEST
STS (KS)	1	range(1), format(1)	2	web/html
VC (VT)	9	range(4), format(9), completeness(4)	1.9	web/html
IT (IH)	29	range, routing	N/A[2]	IRIS

[1] Based on "Utiliteitsbouw"

[2] Not Available because of routing.

edits, the electronic records are of "*high completion quality* " (Spaans 2006). No audit trails are recorded, so it is not possible to investigate the respondents reaction to edit failures. Spaans (2006) notes that there are differences in item nonresponse between the paper and electronic questionnaires, which are most likely due to design issues in the paper questionnaire.

The recently redesigned Structural Business Survey, which includes the electronic version of the questionnaire, was tested in the spring of 2006 (Giesen and Vis 2006; Giesen and Hak 2006; Vis 2005) in a pilot study. A total of 7800 respondents, divided over five strata, were asked to participate. A preliminary evaluation, held before the pilot was finished showed a total nonresponse rate of 52.2%. About 4% of the respondents sent in a paper form, which they had to request at their own initiative. The survey encompasses economic data such as turnover, number of employees, number of Full-Time Equivalents (FTE), costs, *etc.* The electronic questionnaire had a summation functionality, plus 14 other edit checks. These edit checks included only consistency and completeness checks, such as "final date of the financial year must be later than the starting date" and "turnover has to be entered". It is noted by Giesen and Vis (2006) that item nonresponse is in some cases higher in the electronic questionnaire, probably due to some design issues in the electronic version. Audit trails were recorded in the pilot studied, which were investigated by Lammers (2006). As the appearance of error messages was not recorded, studying respondent behaviour to edits is difficult. However, the results indicate that audit trails might be used to classify the respondents into three types: fast, "serious" (slow) and paper-minded respondents. The latter respondents print and complete the questionnaire on paper before filling in the electronic version.

Both in the Pilot SBS and the NES, there was no clear indication that reporting *via* CASI happens faster than with PAP (Giesen and Vis 2006). However, Spaans (2006) notes that the first CASI response came in very quickly. About 7% of the response was retrieved at the day the respondents received the letter. The conclusion that electronic response comes in faster than paper response can also be drawn from the data presented by Hoekstra (2007), who investigated the response time for the monthly Short-Term Business Survey[3]. Although weak, these results provide evidence that when moving to electronic surveys, response times for small questionnaires (*i.e.* with a small number of variables) shorten more than the response times for large surveys.

In Table 2 we list four business questionnaires from Statistics Netherlands with the number and type of built-in edits, and the method of data-entry. The figures for the Structural Business Survey correspond to the questionnaire which is operational now, not the one used in the pilot. In general, entries in numerical fields are limited to numbers and (sometimes) minus-signs and there is a maximum number of characters which can be entered. For example, turnover numbers in the SBS have a maximum length of eight characters including minus signs, which yields a range edit forcing the numbers to be integers between $-9,999,999$ and $99,999,999$. Thus, a length check yields an interval check which is asymmetric about zero. Furthermore, dates are forced into the format $\mathsf{dd}-\mathsf{mm}-\mathsf{yyyy}$, and sums are computed automatically. Some fields, such as number of FTE, are checked for negative values: the minus sign key is ignored by the program. Thus, in general there are two checks for every numerical field: range and type.

In the quarterly Vacancy Count, there are four numerical fields and five fields for contact information. The latter fields are only limited in the number of characters, hence the lower number of edits per field.

The International Trade survey is somewhat different. The questionnaire contains a "table question" with 29 variables. Each row is used to report one shipment, and the number of rows completed by the respondent is variable. Since some routing is built in, and some values can be copied automatically, the number of edits/field cannot be unambiguously determined. Instead, it is noted that all categorical data is entered by selection from a list, and range and type edits are implemented for numerical variables. However, categorical data need not be error free when it arrives at the CBS, since respondents may export records from their own bookkeeping system, which can contain erroneous product codes, for example.

2.4 Discussion

Most of the built-in checks at Statistics Netherlands are fairly simple one-variable checks regarding ranges or type (format) of input. The only exception found here are automated summation and copying of variables. However, even these fairly simple input checks are important since it is known that these edits are frequently failed on paper questionnaires [See *e.g.* Giesen and Hak (2006)].

A survey of international literature indicates that respondents are usually willing to correct edit failures, provided that correcting can be done in a user friendly way.

In terms of the quality aspects mentioned in paragraphs 2.1.1 to 2.1.4, the built-in edits address validity, completeness and consistency, but not accuracy. All the built-in edits are basically copied from edits which are also checked during the post-editing process.

The number of built-in edits per variable in electronic questionnaires at Statistics Netherlands is comparable with numbers found in literature such as Anderson *et al.* (2003): about two. However, during the creation of this report it appeared that statistics employees sometimes have the impression that virtually no edits have been built in. The built-in edits are fairly straightforward and may thus not always be perceived as edits.

The documentation of post-processing of data in complex statistics such as the Structural Business Survey or International Trade is mostly available. Documentation of electronic questionnaire software is usually limited to user manuals. Edits which are built into the electronic questionnaires are currently not part of the process documentation while it does influence the data format and quality retrieved by Statistics Netherlands.

The decision about which edits to build in are made by the questionnaire builders and the statistics department. Usability testing is done mostly by colleagues, which yields a danger of over-editing. For example, by limiting what respondents can fill into a questionnaire, the questionnaire can become too strict to cover what a respondent wants to answer. Pilot studies can serve as a tool to locate these cases.

There is no general system or guideline stating what edits to build in for certain variables. There is a danger that different edits are used for similar variables across different questionnaires, which in turn can compromise comparability of variables over various questionnaires. For example, suppose that turnover can be in the interval $[-10^6, 10^7]$ in one questionnaire, and in the interval $[-10^5, 10^6]$ in another one. Then the second questionnaire will more often yield an error than the first one, presuming similar respondents. Since edits depend on the context of the questionnaire, it may be valid to use different ranges for different questionnaires. One should consider however, what this means for comparability.

Electronic questionnaires with built-in edits are developed independently of the editing processes which take place after data retrieval. Since editing at the source is a form of editing, there is conceptual advantage of coordinating, or at least documenting the two forms of editing together.

Audit trails are quite commonly used in usability studies of electronic questionnaires, both at Statistic Netherlands as in other statistical institutes. Using audit trails as an indicator of data quality is not fully explored yet.

3 Post processing

In this chapter two examples of post processing of mixed-mode surveys at statistics Netherlands will be discussed. Also, a more quantitative view on mixed mode data editing will be developed by studying the set (space) of possible records and a discussion of plausibility indices.

3.1 Two examples

3.1.1 Structural Business Survey

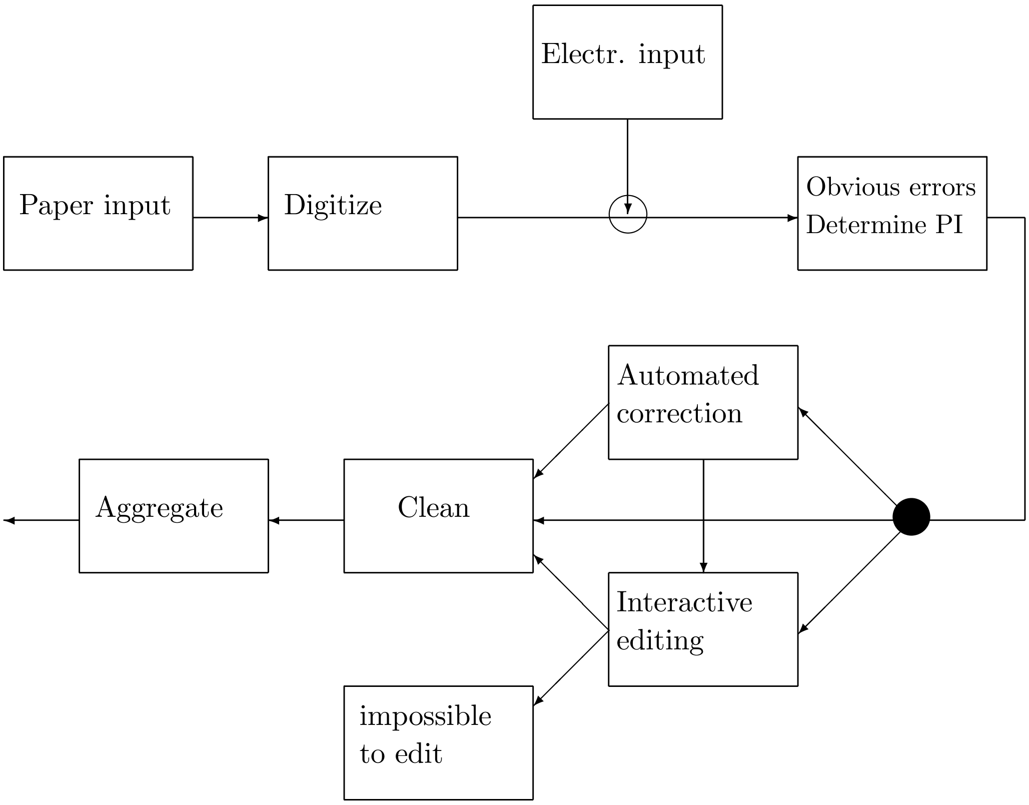

Figure 1. Simplified flow diagram of the micro editing proces for the Structural Bussiness Survey at Statistics Netherlands. Electronic and digitised paper input are put together and stored in a single file before processing. See Lurken (2005) for a more extensive overview.

In Figure 1 a flow diagram for the micro-editing process used in the Structural Business Survey is shown. The data arriving at statistics Netherlands is processed in the following steps: Paper forms are scanned and digitised, and electronic forms come in via upload. The records are marked either with the code of the data-entry employee for paper forms or with the codeword VLSTORE for uploaded forms. All records are stored in a database system which is also equipped with functionality to keep track of the logistics involving the questionnaires. In the next step, the data are exported and automatically checked and corrected record-by-record for obvious errors. These corrections include making invalid negative values positive, removal of double entries, correcting unity measure errors (such as factors of 1000 in currency), and computing totals which have been left empty. Also, empty specifications are given a start value for automated processing in following steps. After this, a Plausibility Index is determined and used to send records either to interactive (manual) or automated editing. The automated editing process is done using the CBS program Cherrie Pi (Lurken 2005). Records which cannot be imputed automatically are send to the interactive editors.

This editing process is an example of a bottom-up approach to editing. That is, records are checked for internal consistency and are corrected record-by-record until all records are processed, after which the data can be aggregated and analysed further.

3.1.2 International Trade

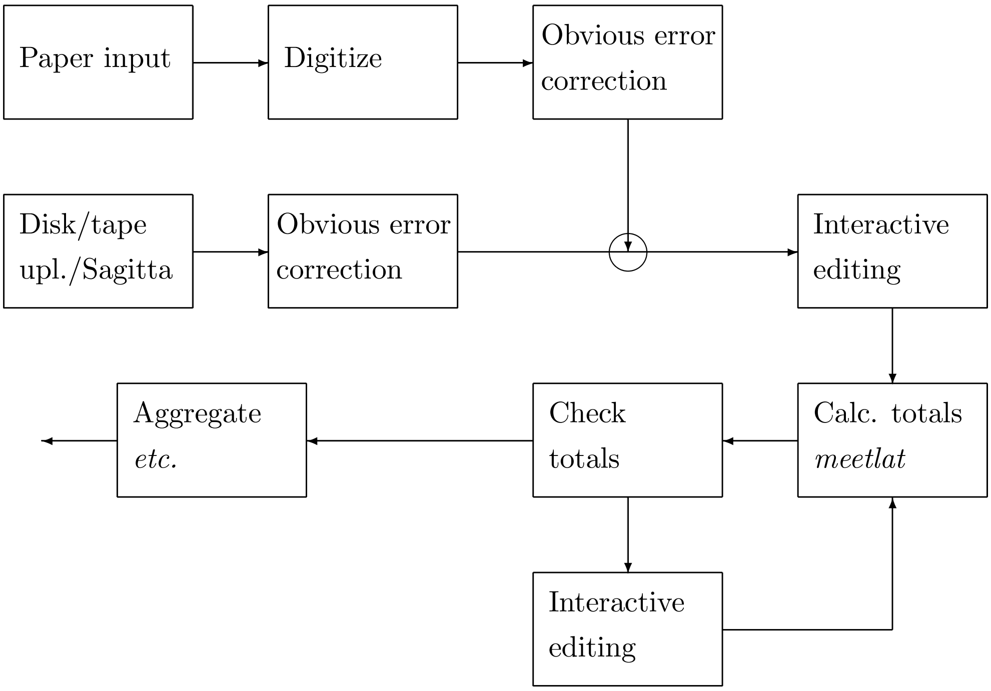

Figure 2. Partial flow diagram of the editing process for the "International Trade" statistic at Statistics Netherlands. A thorough description can be found in Booleman et al. (1997) and Jongen (1997).

In Figure 2 a simplified flow diagram for the editing process for the International Trade survey is shown. In the figure, two input streams are distinguished: an electronic and a paper stream. In reality, the electronic stream is composed of a stream of customs registration data [Sagitta, see *e.g.* Ensinck] and a stream of respondent data in the form of uploads, disks and tapes. Respondents can use CBS-IRIS to fill out the questionnaires or export data from their own bookkeeping system it into a predefined format.

The input streams have been documented in Thijssen (2002) and Ensinck. The incoming data streams are checked for processability and obvious simple errors are corrected automatically. It is worth noting that not all processability checks for electronic data are similar to the ones which may occur in the paper questionnaires. For example, the end-of-line characters for uploaded (encrypted) ASCII files must be checked and converted, since different (operating) systems use different formats. Apple and Unix systems use the Line Feed character (LF), while in Windows CR LF (CR= Carriage Return) is used. Another specific check is to check the length of records in an uploaded file.

All records are then converted to a single data format and stored in the same file. After interactively correcting obvious errors which could not be repaired automatically, the central editing step called *meetlat* (ruler) is invoked. In this step the data is viewed as a matrix with (type of goods)×(group of countries) labeling columns and (sets of) respondents labeling rows. Row and column totals are computed, and compared with expected values, based on historical data. When outliers are found, an editor can zoom in on a cell to find and repair suspect values using specially designed software (Booleman *et al.* 1997).

This editing process is an example of a top-down approach to editing where aggregated data serves as an indicator of suspect values. Editing and error detection on the micro-level is only done after aggregated values seem suspect. It must be noted however, that a limited amount of micro editing (of obvious errors) is performed in the process before the top-down approach is applied.

3.2 Mathematics of data editing

In order to gain insight in the nature of the difference in editing CASI data and PAP data, in this subsection the properties of the sets of records coming from PAP or CASI are analysed. The example introduced here will also be used in Section 4.

Mathematically, any questionnaire with say, m variables gives rise to a so-called *code space* $\mathcal{A}$. The code space is the set of all possible records that can be delivered by a respondent. It can be written as as follows:

$$\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \dots \mathcal{A}_m. \tag{1}$$

Here, each $\mathcal{A}_i$ stands for the possible values that a respondent can enter in the ith field of the questionnaire. The $\mathcal{A}_i$ are called *domains* and $\mathcal{A}$ is a *cartesian product* of domains. The elements of $\mathcal{A}$ are the records $\mathbf{a} = (a_1, a_2, \dots a_m)$. A record is an ordered list of filled in values. In general, these values can be strings, numbers, “No Answer” (denoted NA), booleans, categories, and so on. We will call records coming from some code space $\mathcal{A}$, $\mathcal{A}$-records.

For example, in the statistic “Construction Works under Development (CWD)” [*Bouwobjecten in Voorbereiding, BIV*, see van der Loo and Pannekoek (2007)] two of the requested variables are building budget b and building surface s. The part of code space associated with these variables can be written as $\mathcal{A} = \mathcal{B}\times\mathcal{S}$, where $\mathcal{B}$ and $\mathcal{S}$ are the sets of values that can be filled in by the respondent. In principle, on a paper form, the respondent can fill in anything, so the sets are given by

$$\mathcal{B} = \mathcal{S} = \{\mathsf{NA}\} \cup \mathbb{R}, \tag{2}$$

where $\mathbb{R}$ is the set of real numbers, and for simplicity, NA here means both “no answer” and anything which cannot be read as a real number, such as words, letters, and drawings.

For the post processing of data, we may cut the code space in two parts: the part that contains the correct records (records that pass all edits), and the part that contains records with at least one error. The part which contains the corrupted records is called the *edit space* $\mathcal{E}$, and we say that a record $\mathbf{a}$ *fails* if $\mathbf{a} \in \mathcal{E}$. The part of $\mathcal{A}$ which contains the correct records is the complement of $\mathcal{E}$ in $\mathcal{A}$, and will be called $\mathcal{E}^*$. Editing after data entry consists of three things: first to determine if a record is in $\mathcal{E}$ or not, and if it is, find out how it is corrupted exactly. The second step is to decide which fields to change. A common choice in automated processing is to change as few variables as possible. This is called the principle of Felligi and Holt (Fellegi and Holt 1976). The third step consists of computing the new values and replacing the old ones.

Thus, the process of micro editing is completely determined by the code space and its edit space. Both are in turn determined by the metadata corresponding to the questionnaire.

In practice, the corrupt part of space is determined by a set of edit rules, or *edits*, each of which determine a piece of $\mathcal{E}$. In the case of the CWD example above, these rules can be stated as follows:

$$\begin{aligned} \text{budget is empty} \quad & e_1: \quad b = \mathsf{NA} && (3)\\ \text{budget smaller than 200 kEUR} \quad & e_2: \quad b < 200 && (4)\\ \text{surface is empty} \quad & e_3: \quad s = \mathsf{NA} && (5)\\ \text{surface is negative} \quad & e_4: \quad s < 0 && (6)\\ \text{outliers are suspect} \quad & e_5: \quad f(b,s) > g, && (7) \end{aligned}$$

where we labeled the edits $e_1 \ldots e_5$ for future reference. In the last edit (e_5), $f(b,s)$ is an outlier detection function which obtains larger values when the absolute ratio $|b/s|$ is unusually small or large. Records fail the edit if $f(b,s)$ exceeds some predetermined value g. (One must also define the outlier function when one or both of the data b or s is missing or 0. We will not discuss this here). Records which satisfy one or more of the rules above, are of course corrupted.

Now that we have a mathematical description of the edits, we would like to explicitly determine $\mathcal{E}$ and $\mathcal{E}^*$. To do this, we proceed as follows: first we determine from the edits [Eq. (3)-(7)] the relevant separate (disjunct) parts of $\mathcal{B}$ and $\mathcal{S}$.

$$\begin{aligned}\mathcal{B} &= \{\mathsf{NA}\} \cup \{b < 200 \wedge f(b,s) \leq g\} \cup \{b < 200 \wedge f(b,s) > g\} \\ &\quad \cup \{b \geq 200 \wedge f(b,s) \leq g\} \cup \{b \geq 200 \wedge f(b,s) > g\} \qquad (8) \\ \mathcal{S} &= \{\mathsf{NA}\} \cup \{s < 0\} \cup \{s \geq 0\}. \qquad (9)\end{aligned}$$

The edit which depends on a combination of b and s [Eq. (7)] is used only to split up the set of possible values for b. Doing the same for s would lead to double counting of edits. The choice to apply edit (7) to split up the domain of b and not of s is arbitrary. The complete code space $\mathcal{A}$ can be retrieved by making cartesian products of the subsets in Eq. (8) and Eq. (9). Since there are 5 subsets in $\mathcal{B}$ and 3 subsets in $\mathcal{S}$, there are $5 \cdot 3 = 15$ possible *record types*, 1 of which is in $\mathcal{E}^*$, and 14 in $\mathcal{E}$.

The space of correct records $\mathcal{E}^*$ is given by

$$\mathcal{E}^* = \{b > 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\}. \qquad (10)$$

It is clear that any record $\mathbf{a} \in \mathcal{E}^*$ passes all edits given in Eqs. (3)-(7). The edit space $\mathcal{E}$ is built up from the cartesian products of the remaining subsets of $\mathcal{B}$ and $\mathcal{S}$. In total there are $3 \cdot 5 - 1 = 14$ disjunct subsets in $\mathcal{E}$:

$$\begin{aligned}\mathcal{E} &= \{\mathsf{NA}\} \times \{\mathsf{NA}\} \\ &\quad \cup \{b < 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\} \\ &\quad \cup \{b < 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\} \\ &\quad \cup \{b \geq 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\} \\ &\quad \cup \{b \geq 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\} \\ &\quad \cup \{\mathsf{NA}\} \times \{s < 0\} \\ &\quad \cup \{b < 200 \wedge f(b,s) \leq g\} \times \{s < 0\} \\ &\quad \cup \{b < 200 \wedge f(b,s) > g\} \times \{s < 0\} \\ &\quad \cup \{b \geq 200 \wedge f(b,s) \leq g\} \times \{s < 0\} \\ &\quad \cup \{b \geq 200 \wedge f(b,s) > g\} \times \{s < 0\} \\ &\quad \cup \{\mathsf{NA}\} \times \{s \geq 0\}\end{aligned}$$

$$
\begin{aligned}
&\cup\{b < 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\} \\
&\cup\{b < 200 \wedge f(b,s) > g\} \times \{s \geq 0\} \\
&\cup\{b \geq 200 \wedge f(b,s) > g\} \times \{s \geq 0\}. \qquad (11)
\end{aligned}
$$

Micro-editing software based on the Felligi-Holt paradigm always implement some algorithm to (implicitly) determine the record type. For every record type, the minimum number of fields that have to be changed is determined, and some imputation method is chosen.

So far, we assumed a PAP questionnaire, and the respondent can fill in anything. Suppose that an electronic questionnaire is used with built-in edits such that no values out of the predefined ranges can be filled in. In other words edits (4) and (6) are built in. This essentially means that for the post processing step the code space associated with the questionnaire has changed. The new code space will be called $\mathcal{A}' = \mathcal{B}' \times \mathcal{S}'$ where $\mathcal{B}'$ is given by

$$
\begin{aligned}
\mathcal{B}' &= \{\mathsf{NA}\} \times \{b \in \mathbb{R} \mid b \geq 200\} \\
&= \{\mathsf{NA}\} \cup \{b \geq 200 \wedge f(b,s) \leq g\} \cup \{b \geq 200 \wedge f(b,s) > g\}, \qquad (12)
\end{aligned}
$$

and $\mathcal{S}'$ is given by

$$
\mathcal{S}' = \{\mathsf{NA}\} \cup \{s \in \mathbb{R} \mid s \geq 0\}. \qquad (13)
$$

The first line in Eqs. (12) and (13) corresponds to the situation in Eq. (2). The second line in Eq. (12) is again a disjunct separation of the domains as in Eq. (8). This is a result of analysing the remaining edits given by Eqs. (3), (5) and (7). The domain in Eq. (13) does not have to be split up further.

There are 6 record types coming from the electronic questionnaire. The space of correct records $\mathcal{E}'^*$ is exactly the same as given in (Eq. 10), so $\mathcal{E}'^* = \mathcal{E}^*$. The edit space $\mathcal{E}'$ is given by

$$
\begin{aligned}
\mathcal{E}' = &\{\mathsf{NA}\} \times \{\mathsf{NA}\} \\
&\cup\{b > 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\} \\
&\cup\{b > 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\} \\
&\cup\{\mathsf{NA}\} \times \{s \geq 0\} \\
&\cup\{b > 200 \wedge f(b,s) > g\} \times \{s \geq 0\}. \qquad (14)
\end{aligned}
$$

If a record is to be corrected using as much of the (assumed) correct information in the record to find new values for erroneous fields, then for every record type in $\mathcal{E}'$ (or $\mathcal{E}$) a separate imputation method must be chosen. Moving edits into the electronic questionnaire severely limits the combinatorial explosion of corrupt record types. Thus, the number of solutions that have to be implemented to (automatically) correct the data possibly reduces significantly.

The above analysis makes clear that the complexity of the editing process after data retrieval can be severely reduced just by including some simple edits in the electronic questionnaire. The argument presented here applies also for questionnaires containing categorical or mixed categorical and numerical data. It must be noted however, that the argument only holds if the respondent is somehow forced to give an answer which obeys the edit rules. For simple range edits and format edits this is already commonly implemented at Statistics Netherlands. When the respondent has the option to either correct or neglect the edit, the code space does not reduce from $\mathcal{A}$ to $\mathcal{A}'$ and all the edits stay in place.

Finally, in a mixed mode setting, the editing process after data retrieval is confronted with two types of records: $\mathcal{A}$-records coming from PAP, and $\mathcal{A}'$-records coming from CASI. One can then choose to create the full editing process for $\mathcal{A}$, and profit from the fact that the CASI-records have less errors. Alternatively, a simplified process can be set up, where the only $\mathcal{A}$-records that are led into the editing process are the records that are also $\mathcal{A}'$-records or can be turned into $\mathcal{A}'$-records by deductive imputation. The choice will depend on the type of built-in edits and the ratio of PAP *versus* CASI response. The latter option amounts in neglecting records in the data set, which is feasible only for (very) small numbers of records.

3.3 Plausibility index

The plausibility index P is a quantity, commonly used in data editing practices, which indicates the "corruptness" of a record by computing a number, say between 0 and 1, for every record. A higher P means a more erroneous record. The main question here is: how can the plausibility indices for $\mathcal{A}$ and $\mathcal{A}'$ be compared?

Suppose we have a PAP questionnaire, with a code space $\mathcal{A}$ and a plausibility index taking values between 0 and 1. Technically, P is a map

$$P : \mathcal{A} \to (0, 1). \tag{15}$$

When a number of edits are built in some CASI questionnaire, we are left with a remaining code space $\mathcal{A}' \subset \mathcal{A}$. We would like to determine the plausibility index P' for $\mathcal{A}'$ from P.

The easiest option is to take for P' the restriction to $\mathcal{A}'$. That is, the same formula is used for P' as for P, and it is taken for granted that the values of P' are probably not in $(0, 1)$, but in some subset of that range. This basically means that $\mathcal{A}'$-records are interpreted as "cleaner" $\mathcal{A}$-records.

On the contrary, one can argue that the plausibility of a record should be evaluated by judging the occurring errors with respect to the possible errors. One way to obtain such a “relative” plausibility index is to define P' as follows:

$$P'(\mathbf{a}') = [\max_{\mathbf{a}'\in\mathcal{A}'} P(\mathbf{a}') - \min_{\mathbf{a}'\in\mathcal{A}'} P(\mathbf{a}')]^{-1}[P(\mathbf{a}') - \min_{\mathbf{a}'\in\mathcal{A}'} P(\mathbf{a}')], \tag{16}$$

for all $\mathbf{a}'$ in $\mathcal{A}'$. It is easily seen that this index has values in $(0, 1)$. Other scaling methods, based for instance on the relative volumes of $\mathcal{E}$ and $\mathcal{E}'$ are conceivable as well. Thus far, there is little or no research in this direction.

3.4 Discussion

The examples given in section 3.1 show that CASI and PAP data are treated mostly similar in the editing process, since at arrival, data is converted to a single format and gathered in a single file. There are differences however: paper forms are digitised by typists, and can contain format problems such as characters in numeric fields. Uploaded data may contain file format problems, such as swapped columns. The process of data conversion before it is delivered to editing software must be seen as part of the editing process. It is in this stage of data processing where the first differences between editing of PAP and CASI data occur.

It follows from the discussion in the previous subsections, that data coming from CASI and PAP questionnaires should be viewed as if coming from different questionnaires, since the set of possible records from CASI and PAP are not the same. Comparing the record quality (which is a *value judgement*) is therefore not straightforward.

Using CASI questionnaires with built-in edits instead of PAP questionnaires does not only reduce the number of checks that need to be performed on the data. Some CASI-specific edits, like checking record length, are introduced. It is yet unclear if data from audit trails could also be used in judging data quality.

At International Trade, the CASI-specific checks are done before merging the PAP and CASI data into a single file. However, since the CASI has built-in edits, the record set after merging is still not uniform with respect to editing. To obtain a truly uniform set of data, the PAP data should be processed to pass all the edits which are built into the CASI questionnaire before the two streams are merged.

4 Quality of data

There are numerous sources in literature discussing quality issues in National Statistical Institutes (NSI's). Recent examples include Brackstone (2003); Lyberg (2000) and Karr *et al.* (2005). The consensus nowadays is that a statistical process is viewed as a customer-supplier chain, with the users of statistical data at the end of the chain, and respondents at the beginning. Surveyors can be seen as customers of respondents and editors can be seen as customers of surveyors. Overall quality is defined in terms of a set of demands, posed by the end user. These demands then propagate backwards over the supply chain, yielding specific demands for each step in the statistical process. Karr *et al.* (2005) define end-user quality of statistical data in decision theoretic manner:

> *Data quality is the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions.*

This definition leads then to quality aspects such as accessibility, timeliness, coherence, interpretability, and relevance for the user. From there, one can work back over the supply chain to define quality aspects of the raw data delivered by respondents.

4.1 Definition of data quality for editing

Although the data quality aspects validity, completeness, consistency and accuracy are often mentioned in various papers mentioned above, one rarely finds quantitative measures of (raw) data quality.

Raw data quality must not be confused with the usual statistical quality measures such as variance estimates and confidence intervals. These numbers say something about aggregated data on the (near) end-user level. The fact that editors can be seen as a customer of the surveyors leads to specific quality demands, which seems to have been overlooked in the literature. This is a pity, since it is one of the few areas where the sense of quality can be made quantitative. In the spirit of the definition of Karr *et al.* (2005) we define raw data quality for data editors as follows:

> *The ability of data to be checked and corrected efficiently and economically so that a database of valid, complete, consistent and accurate records can be delivered for aggregation.*

This definition directly leads to the following demands on the data delivered by data-gathering parties.

4.1.1 Documentation

The nature of records which are offered for editing must be clear. This includes both a description of the file format as well as a description of the relevant code space. For example by documenting the edits, built into CASI questionnaires.

4.1.2 Simple structure of code space and edit space

Editing is easier when the number of choices that have to be made during the error correction process is smaller. Reducing the complexity of code space and edit space is one way to achieve this.

Consider again the example given in Section 3 of the two variables from the CWD questionnaire. In Tables 3 and 4, the record types mentioned in the example are tabulated. Table 3 shows the record types for the PAP questionnaire, and Table 4 shows the record types for the (fictional) CASI questionnaires with the range edits built in.

In the 3rd and 4th column, the failed edits and the fields that must be edited are listed for each record type. Removing edits by building them into a questionnaire has two concequences, namely

1. it reduces the number of options (fields) for error correction, and
2. it reduces the number of fields to be edited.

For example, record type 14 in Table 3 concerns records for which the budget-to-surface ratio is suspiciously high or low. It is not *a priori* clear if either the budget variable or the surface variable should be edited. Another example is record type 2 in Table 3. Since both the budget is out of range, and the surface is left empty, both the variables have to be edited in order to get a fully consistent record. In the CASI case, edit nr. e_2 [Eq. (4)] cannot be failed anymore, thus avoiding this problem.

4.1.3 High plausibility (low plausibility index)

Analysing the structure of code- and edit space allows one to compare the quality of different modes and questionnaires. Within one questionnaire, a file of records is easier to edit when there is less to edit. Ideally, there are no records containing errors and all records have plausibility index zero. In practice, this is of course as much dependent on the respondent as on the design of the questionnaire. The high plausibility requirement thus partially translates to a requirement the surveyor poses to the respondent.

4.2 Measuring quality

Authors discussing the effect on quality of implementing edits into CASI's use various quality indicators which are hard to compare. For example, Sweet and Ramos (1995) count the number of failed edits, and conclude after a statistical analysis that significantly less edits are failed in a CASI questionnaire compared with the PAP version. Cohen (2003) reports the fraction of records failing at least one edit. De Leeuw (2005) notes that self-administered modes yield slightly higher quality, without quantifying this precisely.

Most authors who quantify record quality somehow count the number of failing records. In their paper on data quality, Karr *et al.* (2005) suggest the fraction of fully intact records as a measure. However, as was made clear in sections 3.2 and 3.3, comparing these fractions between CASI and PAP questionnaires, or between any two questionnaires is not without problems since the records are associated with different code spaces. One should somehow take into account which edits are failed and how important those failures are. Another question is how to actually compute the fraction. How does one count nonresponse, or empty records? In some cases, such as in the International Trade statistic, an empty record can be a valid answer. Defining the fraction of error free records is not always straightforward.

4.2.1 An example

In order to gain some insight in the quality effect of building edits into CASI questionnaires, consider again Tables 3 and 4. In the fifth column of Table 3, the percentage of observed records is listed for every record type. For example, 5.56% of the respondents fail edit e_2 [Eq. (5)]. That is, a too low building budget is filled in. The percentage of records without any budget or surface errors is 76.84.

In the last column of Table 3, an estimate of the contribution of the records to the total budget is given. Records where the budget b_i is missing (failing edit e_1) have been inputed with $b_i = \bar{x}s_i$, where $\bar{x}$ is the average price per square meter. When the surface s_i was also missing (failing e_3), the overal median budget was imputed. It can be seen that the outliers (record type nr. 14, failing e_5) consist of only 1.25% of the records, but contribute significantly (48.76%) to the total budget.

In Table 4, the percentages are computed for the CASI case, where it is impossible to fail edits e_2 (too low budget) and e_4 (surface below 0). The numbers in Table 4 are computed using the same data set as for Table 3, but simply ignoring edits e_2 and e_4. By comparing Table 3 with Table 4, it can be seen that building in the edits, yields 5.57% less failing records, simply because some

Table 3. Record types, failed edits, fields to edit and observed percentage of records per type for the PAP CWD questionnaire. The percentages are based on 881 records (building projects).

	Record type	failed edits	fields	rec. (%)	budg. (%)
1	$\{\mathsf{NA}\} \times \{\mathsf{NA}\}$	e_1, e_3	b, s	2.50	0.28
2	$\{b < 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\}$	e_2, e_3	b, s	0.91	0.01
3	$\{b < 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\}$	e_2, e_3, e_5	b, s	0.00	0.00
4	$\{b \geq 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\}$	e_3	s	9.65	16.24
5	$\{b \geq 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\}$	e_3, e_5	s	0.00	0.00
6	$\{\mathsf{NA}\} \times \{s < 0\}$	e_1, e_4	b, s	0.00	0.00
7	$\{b < 200 \wedge f(b,s) \leq g\} \times \{s < 0\}$	e_2, e_4	b, s	0.00	0.00
8	$\{b < 200 \wedge f(b,s) > g\} \times \{s < 0\}$	e_2, e_4, e_5	b, s	0.00	0.00
9	$\{b \geq 200 \wedge f(b,s) \leq g\} \times \{s < 0\}$	e_4	s	0.00	0.00
10	$\{b \geq 200 \wedge f(b,s) > g\} \times \{s < 0\}$	e_4, e_5	s	0.00	0.00
11	$\{\mathsf{NA}\} \times \{s \geq 0\}$	e_1	b	1.59	1.25
12	$\{b < 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\}$	e_2	b	5.56	0.08
13	$\{b < 200 \wedge f(b,s) > g\} \times \{s \geq 0\}$	e_2, e_5	b	1.70	0.01
14	$\{b \geq 200 \wedge f(b,s) > g\} \times \{s \geq 0\}$	e_5	b or s	1.25	48.76
15	$\{b \geq 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\}$	–	–	76.84	33.37

Table 4. Record types, failed edits, fields to edit, and percentage of records per type for the (fictional) CASI CWD questionnaire with range edits built in. The percentages are determined by assuming that built-in edits are satisfied by all records.

	Record type	failed edits	fields	rec. (%)	budg. (%)
1	$\{\mathsf{NA}\} \times \{\mathsf{NA}\}$	e_1, e_3	b, s	2.50	0.28
2	$\{b > 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\}$	e_3	s	10.56	16.25
3	$\{b > 200 \wedge f(b,s) > g\} \times \{\mathsf{NA}\}$	e_3, e_5	s	0.00	0.00
4	$\{\mathsf{NA}\} \times \{s \geq 0\}$	e_1	s	1.59	1.25
5	$\{b > 200 \wedge f(b,s) > g\} \times \{s \geq 0\}$	e_5	b or s	2.95	48.77
6	$\{b \geq 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\}$	–	–	82.41	33.45

records are taken out of the code space. The effect is completely due to building in e_2, since this is a frequently occurring error: $0.91 + 5.56 = 6.47\%$ of the records in the data set fail this edit. Edit e_4 is never failed in the current data set[4]. Thus, building edit e_4 into the CASI does not improve plausibility, but it does improve the quality of edit structure since less edits need to be checked, which simplifies the editing process. Building in edit e_2 improves both aspects of raw data quality.

Table 5. Record types, failed edits, fields to edit, and percentage of records per type for the (fictional) CASI CWD questionnaire with range edits and outlier detection built in. The percentages are determined by assuming that built-in edits are satisfied by all records.

	Record type	failed edits	fields	rec. (%)	budg. (%)
1	$\{\mathsf{NA}\} \times \{\mathsf{NA}\}$	e_1, e_3	b, s	2.62	2.50
2	$\{b > 200 \wedge f(b,s) \leq g\} \times \{\mathsf{NA}\}$	e_3	s	11.07	10.56
3	$\{\mathsf{NA}\} \times \{s \geq 0\}$	e_1	s	0.00	1.59
4	$\{b \geq 200 \wedge f(b,s) \leq g\} \times \{s \geq 0\}$	–	–	86.33	82.41

In Table 5 a similar analysis is shown, for the case where e_2 (no budget), e_4 (wrong surface), and e_5 (outlier detection) are built into the CASI questionnaire. It is assumed that no outliers occur in the data from CASI with built-in outlier detection. To compute the percentages, records containing outliers were ignored here. It is clear that the estimate for the total budget becomes much more reliable. About 82.41% the budget comes from error-free records. Another 10.56% comes from records where the budget has no errors that can be detected.

Finally, note that since all the record types are disjunct and make up the code space completely when combined, the last two columns in Tables 3, 4, and 5 are distributions over the code space. They can thus be seen as probabilistic measures over the code space. The combined analysis of code space structure and the associated measures over code space can provide useful handles to judge for instance which edits to build into the questionnaire.

4.3 An open question

It is shown in the previous sections that building edits into CASI questionnaires leads to a reduction of erroneous record types, and thus to a reduction of the complexity of the editing process. A second way to achieve a smaller code space is to reduce the number of redundant variables in the questionnaire. After all, reducing the number of variables reduces the number of possible mistakes. For example, the variable s (surface) in the CWD questionnaire is redundant in the sense that it is not aggregated for publication. The surface is only requested from the respondent to be able to check for outliers [edit e_5, Eq. (7)]. However, adding this variable adds two extra edits: e_3 and e_4, [Eqs. (5) and (6)], thus adding to the complexity (number of choices to be made) of the editing process.

This first leads to the question of how requesting redundant information affects data quality. Secondly, if data quality (after editing) is in general better when redundant information is present, how many redundant variables are necessary or optimal? These questions are quite fundamental for questionnaire development and are unanswered at the moment. Analysing code spaces and their associated measures can provide useful insight into these questions since it offers a tool to compare different (versions of) questionnaires.

5 Conclusions and recommendations

5.1 Conclusions

Based on the foregoing discussions, the following conclusions can be drawn.

1. Statistics Netherlands (SN) is aiming to retrieve as much data via CASI (or registration) as possible. Respondents have a right to fill in and send in questionnaires on paper, so 100% electronic retrieval is not (yet) possible.

2. In general, high percentages of electronic response (as opposed to paper response) are obtained. However, the amount of paper response is high enough to regard all electronically available business surveys as mixed mode surveys.

3. Statistics Netherlands builds edits into electronic establishment questionnaires quite commonly. The ratio of edits/field is about 2 and concerns mostly range edits, type (format) edits, automated summation and sometimes routing. Edits are mostly hard edits (CHECKS in Blaise). No edits regarding accuracy, such as ratio edits or longitudinal edits are built in.

4. Built-in edits are not always perceived as part of the editing process by SN's employees, probably due to the simplicity of the edits.

5. Edits which are built into CASI questionnaires at SN are not documented yielding record formats which are not fully specified for post processing. This is a lack in the process documentation.

6. The decision about what edits to build in is made by questionnaire designers, statistics departments and their editors. There is no systematic way to determine what edits to build in. This could compromise comparability of variables across questionnaires and yields a danger of over editing.

7. Audit trails can be recorded for CASI's and are used in usability studies. The possible use for determination of data quality has not been fully explored yet.

8. In general, respondents are found to react positively to edits, provided they are built in in a user friendly way. A high percentage (over 50%, but usually higher) will react to an edit by correcting. Even advanced accuracy-measuring edits, such as longitudinal edits have been implemented at statistical bureaus.

9. CASI records reduce the number of possible edits with respect to PAP edits, but can also yield specific extra edits, such as data format errors when respondents can produce their own file.

10. Incoming CASI and PAP records mostly go through exactly the same editing process at Statistics Netherlands. PAP and CASI records are prepared to be stored in a single file after which they enter the same editing chain.

11. CASI and PAP records are elements of different code spaces, making the comparisons of (raw) data quality, using for instance plausibility indices, problematic.

12. Raw data delivered to editors at Statistics Netherlands can be seen as a semimanufacture. The quality aspects of raw data include documentation, code space structure and plausibility. At the moment, only the plausibility index is frequently monitored, although current methods do have comparability issues, as mentioned in conclusion Nr. 11.

13. The effect on adding redundant variables to a questionnaire on the editing process and on data quality is not known. This kind of information would be valuable for the questionnaire development process.

5.2 Recommendations

Based on the findings, we have the following recommendations.

1. Move towards building edits into CASI questionnaires which aim to improve accuracy, such as longitudinal edits.

2. Document the edits built into CASI questionnaires in standard process documentation.

3. Develop standard practices and edits for commonly occurring variables and combinations of variables.

4. Develop quality measures for (raw) data which are comparable for different modes and questionnaires. The analysis put forward here in Sections 3.2 and 4.2 can serve as a research direction.

5. Consider and develop the CASI edits together with the post data gathering editing process at SN. The editing procedures which are done after data gathering are now developed regardless of edits which have been built into CASI questionnaires. In a more integrated approach, one might significantly reduce the data editing workload by smart decisions on built-in edits. An analysis as shown in Chapters 3 and 4 can help with these decisions.

6. Research the possibility of using audit trails in data quality measurement.

7. Ideally, in a pilot study for new surveys, some CASI's will have built-in edits and some not. That way, commonly made errors can be analysed as described for instance in Section 4.2. It also provides insight into the effect of over editing. A pilot study like this, combined with a usability study, can serve as a rational basis to choose which edits should be implemented in the final version of the CASI.

8. Research the effect of requesting redundant information in questionnaires on editing processes and data quality.

Notes

[1] *Productiestatistiek (PS)*

[2] *Bedrijfstelling Starters (BS)*

[3] *Kortetermijn Statistiek (KS)*

[4] More up to date datasets show that it is failed occasionally

References

A. E. Anderson, S. Cohen, E. Murphey, E. Nichols, R. Sigman, and D. K. Willimack. Changes to editing strategies when establishment survey data collection moves to the web. Technical report, US Census Buereau, 2003. available *via* http://www.websm.org.

I. Arbuez, M. Gonzalez, M. Gonzalez, J. Quesada, and P. Revilla. EDR impacts on editing. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.27, available *via* http://www.unece.org/stats.

D. G. Beckler. Electronic data reporting and data collection edits at the national agricultural statistics service. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.29, available *via* http://www.unece.org/stats.

M. Booleman, R. van Brandeburg, H. Brouwer, B. Diederen, J. Florie, A. Gras, G. Hanssen, C. A. Hertog, G. Krewinkel, P. Michels, F. van de Pol, B. Resing, L. Scholten, M. Schütt, G. Slootbeek, M. Tó th-Pál, and K. Vennix. Methodologie van de meetlat. Technical report, RSM/HIP/HIH, 1997.

G. Brackstone. Managing data quality in a statistical agency. *Second meeting of the statistical conference of Americas of th ECLAC*, 2, 2003.

S. H. Cohen. Editing strategies used by the U.S. bureau of labour statistics in data collection over the internet. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2003. Work session on Statistical Data Editing WP.36, available *via* http://www.unece.org/stats.

E. D. De Leeuw. To mix or not to mix data collection modes in surveys. *J. Off. Stat.*, 21, 2005.

C. Ensinck. Beschrijving sagitta in- en uitvoer. Technical report, CBS-MSO.

P. Fellegi and D. Holt. A systematic approach to automatic imputation. *JASA*, 71:353, 1976.

D. Giesen and T. Hak. Revising the structural business survey: From a multimode evaluation to design. Technical report, CBS-TMO, 2006.

D. Giesen and R. Vis. Tussentijdse evaluatie pilot e-ps. Technical report, CBS-TMO, 2006.

R. Göttgens and P. Baart. Benaderingsstrategie e-waarnemen. Technical report, CBS-BWH, 2006.

M. Hoekstra. Analyse van mode-effecten op bedrijfsênquetes. Technical report, CBS-SOO, 2007. Afstudeer opdracht.

P. J. H. Jongen. Verwerkingsprocessen internationale handel. Technical report, CBS-HIH, 1997.

A. F. Karr, A. P. Sanil, and D. L. Banks. Data quality: A statistical perspective. *Statistical Methodology*, 3, 2005.

W. Koller. The impact of edr on long-established surveys: Statistics austria's experience in the short-term production survey. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.21, available *via* http://www.unece.org/stats.

K. Kurkowski. Modernization of the data collection systems at the CSO of Poland. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.xx, available *via* http://www.unece.org/stats.

L. Lammers. Analyse audit trails e-ps, stageverslag cbs-dmh (2006). Technical report, CBS-DMH, 2006.

D. Laroche. Evaluation report on internet option of 2004 census test. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.23, available *via* http://www.unece.org/stats.

H. J. Lurken. Is het interactief gaafmaken van de productie statistieken industrie met behulp van impect 1 applicatie vir proof? Technical report, CBS-SPI, 2005.

L. Lyberg. Recent advances in the management of quality. In *Statistical Quality Seminar, Cheju Island, Republic of Korea*, 2000.

J. Mol and R. Groen. Aansluitplan dcc 2006. Technical report, CBS Programma e-Waarnemen, 2006.

E. M. Nichols, E. D. Murphy, A. E. Anderson, D. K. Willimack, and R. S. Sigman. Designing interactive edits for u.s. electronic economic surveys and censuses: issues and guidelines. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.22, available via http://www.unece.org/stats.

PRODONNA, 2005. PRoductiestraat DOtNet Nieuwe Architectuur documentation can be found on the CBS-intranet site: http://intranet/BESprojecten/prodonna/index.htm.

P. Spaans. Evaluatie onlinewaarneming "bedrijfstelling starters". Technical report, CBS-WVV, 2006.

P. Spaans, 2007. personal communication.

E. Sweet and M. Ramos. Evaluation results from a pilot test of a computerized self-administered questionnaire (csaq) from the 1994 industrial research and developement survey. Technical report, U.S. Census Bureau, Economical and statistical methods and programing division, ESM-9503, 1995.

H. P. M. Thijssen. Beschrijving hih inputsystemen. Technical report, CBS-MSO, 2002.

M. van der Loo and J. Pannekoek. Advies gaafmaken van de statistiek bouwobjecten in voorbereiding. Technical report, CBS-DMK, 2007.

R. Vis. Vragenlabverslag tweede veldtest elektronische ps. Technical report, CBS-TMO, 2005.

R. G. Vonck and W. B. van der Vegt. Haalbaarheidsonderzoek vragenlijstengeneratieproces. Technical report, Deloitte, 2007.

P. Weir. Electronic data reporting-moving closer to respondents. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2003.

P. Weir. EDR and the impact on editing-a summary and a case study. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.28, available *via* http://www.unece.org/stats.

K. Zeila. Electronic data collection system developed and implemented in central statistical bureau of latvia. In *United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians*, 2005. Work session on Statistical Data Editing, WP.25, available *via* http://www.unece.org/stats.