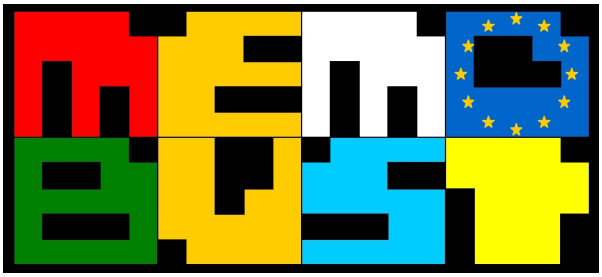


This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Editing During Data Collection

Contents

General section

1. Summary

Data editing is the process of "improving" collected survey data. The improvement involves finding erroneous data and then correcting them. Errors may have happened along the way from the respondent to the survey organisation's data files for various reasons, intended or unintended. Examples include typing errors, wrongly estimated values, misclassifications. Omission or answer denial can also be a source of measurement error. Up to about 40% of statistical agency's resources is spent on editing and imputing missing data (De Waal et al., 2011). In mail business surveys the editing process is performed at the post-collection phase of the survey. The advent of computer technology has enabled statisticians to shift data editing to the data collection stage. Some types of data editing tasks can be performed at the data collection phase. Editing was first incorporated into data collection in the CATI mode. The interviewer is assisted by an electronic questionnaire, which is a program running on his computer. The program contains a built-in set of editing rules, called *edit checks or edits*. These rules assess whether the response is allowed by survey criteria or should be discarded, that is whether an edit is satisfied or violated. Mobile computers extend the field of editing to CAPI. The interviewer conducts a face-to-face interview using an interactive computer program with embedded edit checks. Computer self-administered questionnaires also adopt editing rules, in which the editing process is performed by the respondent. The increasing use of the Internet entails a shift to another mode of survey data collection: online data collection. The prevalent self-administered data collection mode in business surveys and the use of computer questionnaires with incorporated edits enable the editing process at the respondent level. This solution results in many benefits: it decreases costs, improves data quality and response rates and lowers the perceived response burden. For the general issues of data editing in business surveys the user is referred to the topic "Statistical Data Editing".

2. General description

2.1 Communication of the need for correction

The goal of editing at the time of data collection is to take advantage of the measurement instrument to improve quality of the data and reduce the costs of the post-collection process. Data typed into the questionnaire are checked for their correctness. This requires to define the conditions that must be met to assume the response is accurate. The response item has a built-in edit rule to inform the user about an error in case the rule is not satisfied. This leads to the definition what is meant by assuming data are erroneous or data are supposed to be suspicious. Typically, validation in software technology, when a reaction is expected from a user or a user should be informed about something, is notified in a dual way: like an error marked in red colour which means the situation is unacceptable and must be changed in order to continue and a warning which notifies the possibility of incorrectness or to draw attention to a certain aspect of working being the consequence of earlier choices. Moving to the editing field rules that must be satisfied unconditionally – called hard edits – prevent the user from going further or from submitting data to the statistical agency. A second kind of edit rules can be called warnings or soft edits. These kind of edits only notify users that an item should be assessed for its adequacy. In this case three types of resolution of that kind of failures can pointed out: correction, comment or no action. However, no action should be confirmed by respondents as their selection.

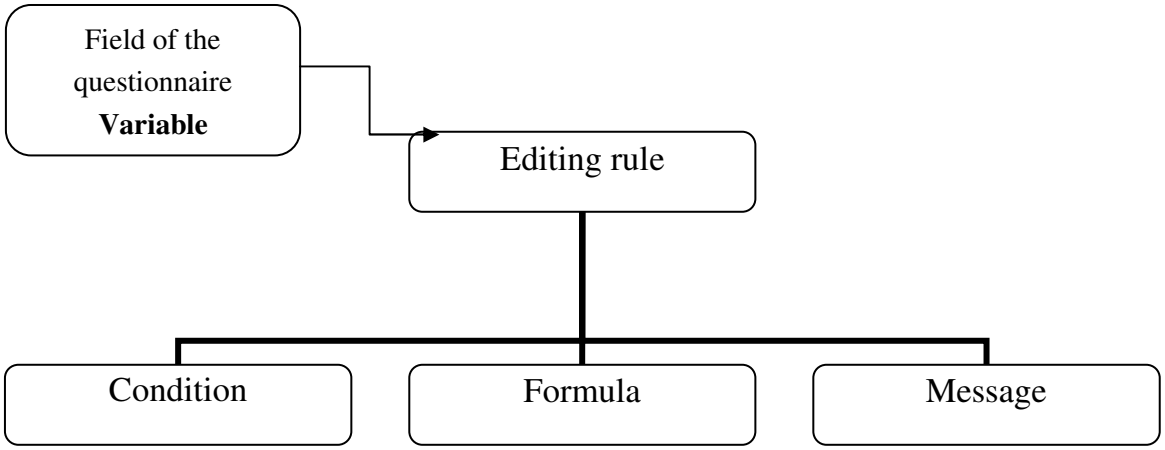
Another treatment of error messages can be pointed out, namely “unsolicited clarification” (Haraldsen, 2013).

2.2 Types of edit rules

An edit rule can be understood as a logical formula triggered when a condition is met under which the variable is tested for its correctness.

IF (variable ∈ editing set) THEN (testing formula).

Below the schematic diagram of an edit check in the questionnaire is presented:



Correctness may depend on the type of the variable:

- a formula that allows keying only some type of characters, e.g., numeric characters,
- a required item check – the edit rule stipulates that the field cannot be empty: null values are not allowed,
- formulas for numeric fields – Let X be a variable denoting, for example, turnover, then edit rules for instance can check:
  - Minimum value ≤ X ≤ maximum value (range constraint),
  - X ≥ 0 (non-negative number requirement),
- a formula that sets a length limit for text variables, the number of characters to be entered is limited,
- an edit rule allows only a specific pattern in the field, for example an e-mail address must contain the @ character.
- edits that check relationships between two or more values:
  - balance edit – checking if the sum of selected items equals a total value,
  - logical formula – various types of relationships between variables (also called inter-item rules), e.g., equality, inequality, greater than, less than, ratio edit, other types of logical relations between two or more variables.

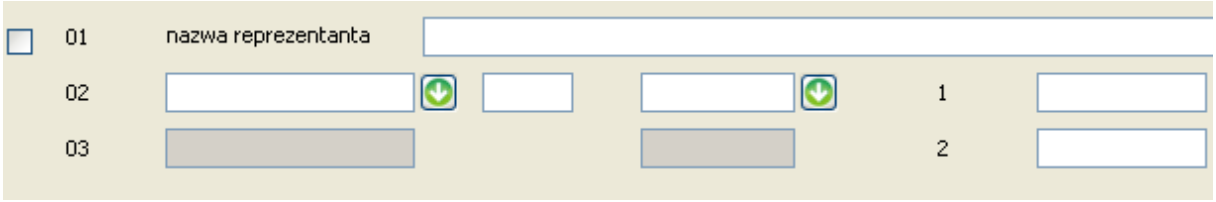
2.3 *The measures to avoid errors*

The items of electronic instruments and their features can be use used as cues to help respondents to complete the responses. There is a possibility of an interactive way of communication between a respondent and a questionnaire. Moreover, the greater usage of a web data collection is an opportunity to tailor the measurement instrument to an individual respondent context. Haraldsen (2013) talks about "questionnaire communication" instead of questionnaire design, stressing the role of the questionnaire as a way to communicate the request for business data. The context is set to technological environment as a shift from paper one-way communication to a dynamic two-way self-administered exchange of information.

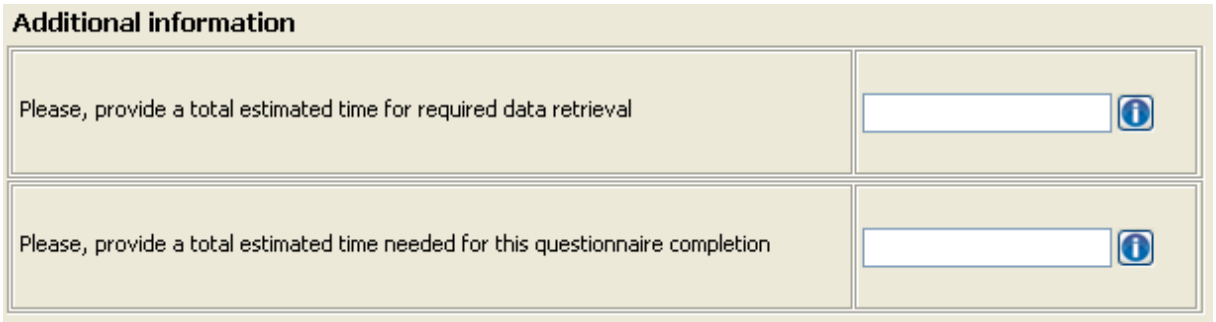
- Information from the business register determines the obligation to convey data to various types of surveys. According to size and kind of activity a list of such surveys, devoted only to the distinguished respondent, can be presented after the user logs in to the web portal which is a communication point for data collection by using electronic questionnaires.
- Access to certain modules of the questionnaire can be determined from answers to previous questions. This means that certain skips and filters can activate when the questionnaire is loading. Also, only selected variables can be enabled for editing. Not only can this improve data consistency but also diminish the response burden. Whether a variable is enabled for editing may depend on previous answer(s). Automatic routing can sometimes lead to a gap in the numbering. Below is an example of such a result. A solution can be to use a two-level numbering.

- Some values can be chosen only from a predefined set of values. The idea is to take advantage of the meta-data environment. The questionnaire items are sometimes based on classification tables. Examples of such classifications are the Statistical Classification of Economic Activities (NACE), the Classification of Products by Activity (CPA), and a table of units for the questionnaire element. The figure below presents a possible solution of limiting the choice to the table containing the CPA nomenclature and table of units.

This is marked by an icon with a green downturned arrow. There is also a possibility to enter values by keying them, but in this case, if the values are not in the table an error message is triggered.



- In longitudinal surveys, editing during data collection should account for relationships between current data and data from previous periods. The motive for doing so is to improve consistency and enable the respondent to pay attention to which data have been submitted previously. This can lead to lowering variability and avoidance of outliers. Another benefit of this approach is that the respondent is presented with values from earlier periods, which reduces the response burden. Whether historical data should be presented or not is a question that has not been clearly settled. A study by Holmberg (2002) indicates that presenting data from earlier periods has a positive effect. The study has not revealed undesired effects of repeating data from earlier rounds or underestimation. Holmberg advocates this approach in surveys with a high degree of data variability. The fear of conformity to and replication of previously reported data in current rounds was not confirmed. On the other hand, a study by Phillips et al. (1995) recommends a more conservative use of historical data. It stresses respondents' inclination to conform to information from previous survey rounds even if the presented data were spurious.
- In Haraldsen (2013), which is chapter 8 of a book devoted to designing questionnaires for business surveys, one can find useful information on how the technological aspects of business web questionnaires can assist shifting from presenting one and the same general approach to all respondents to a more personalised one. Information about a possible error or a request for a confirmation of data entered can result from the analysis of provided responses. A more active dialog can be based on data generated through processes of the questionnaire completion (paradata), registered behind the scenes. The figure beneath provides an example of attaching an icon with sign i (information) close to the field. By clicking on that icon the respondent is assisted with additional clarification about the item.

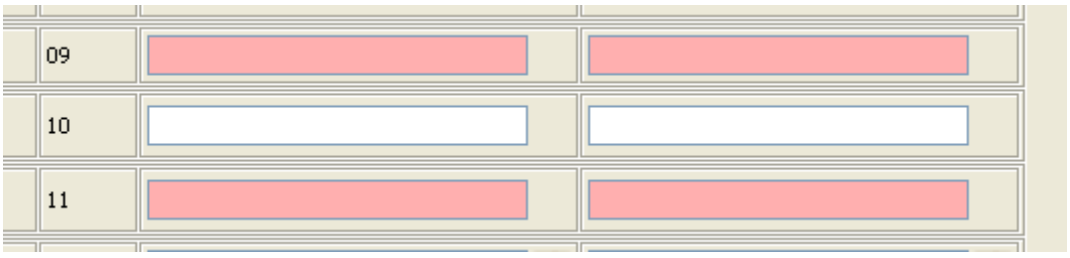


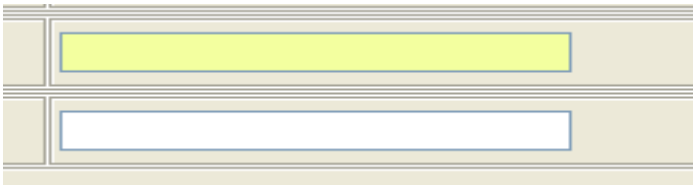
2.4 Presentation of the messages to the respondent

Implementation of edit rules into an electronic questionnaire poses a problem how to efficiently signify the error messages to users. The messages are to be recognised and comprehended. The visual side of a user interface includes a suite of elements such as graphics, colours, and fonts. Another aspect involves the phrasing of the error message. Usability test results suggest that words like “error” or “mistake” should be avoided because of their strong judgemental sense. Politeness and sensitivity towards the user is advisable as well as avoidance of jargon, e.g., computer terminology. The wording of error messages should be similar to that used in questions and associated with the subject of the survey. Error messages should be accompanied by the following attributes: item number, item topic and actual response (Murphy et al., 2001). In the interaction between a user and a computer program a message with a red icon signifies an error as a result of the execution or submission of an incorrect value. By analogy, a similar solution can be used in surveys. Another icon with an exclamation mark is used to signify soft errors. Beneath are examples of these icons:

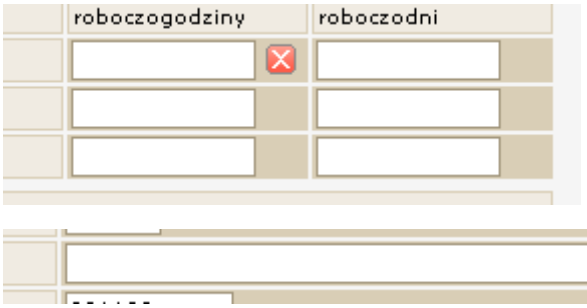
Schonlau et al. (2002) advise placing the message as close as possible to the questionnaire item which it concerns. The message should be displayed either directly above or below the incorrect item. On the other hand, the study conducted by Mockovak (2005) showed no clear significance between different approaches to the placement of messages. Three kinds of solutions were tested. The first two involved placing the message above the item that triggered the edit and directly under that item, respectively. In these cases the error message was displayed after all the items on the page had been completed. In the final solution, the message was placed directly under the item and displayed as soon as the user left the field. The variation in placement and timing of the messages did not have a clear impact on noticing them. It also did not have a significant effect on the resolution of the problem indicated by the message or on following instructions contained in the message text, after the message had been noticed by the respondent. However, participants expressed clear preference for the message under the item.

The following examples present ways of marking erroneous fields using colours.

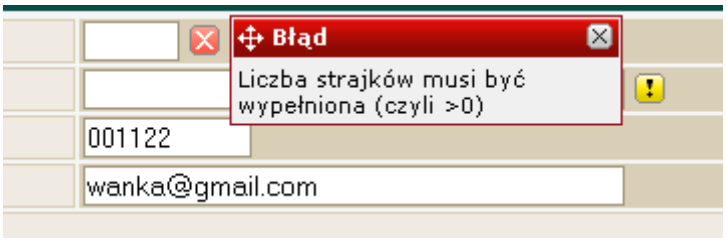


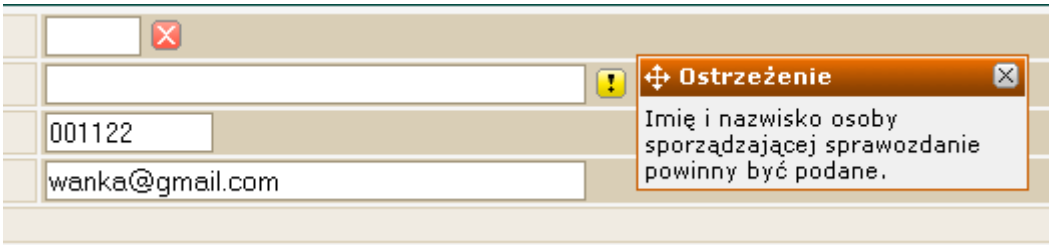


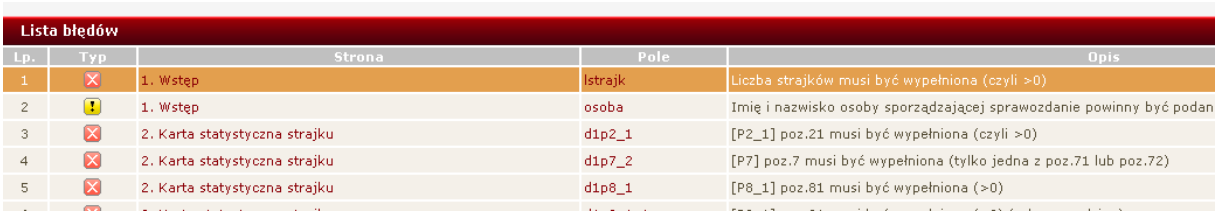
The two examples below are taken from a Polish reporting web portal and present marking the erroneous field by using icons.



The following examples present solutions for displaying error messages (examples taken from a Polish reporting web portal). The first two figures provide messages close to the fields which have been marked as erroneous. Clicking on the item displays the message as a pop-up box. The last example presents a solution where a list of errors is gathered in a table exposed at the bottom of the web page.







Lista błędów

Lp.	Typ	Strona	Pole	Opis
1		1. Wstęp	strajk	Liczba strajków musi być wypełniona (czyli >0)
2		1. Wstęp	osoba	Imię i nazwisko osoby sporządzającej sprawozdanie powinny być podan
3		2. Karta statystyczna strajku	d1p2_1	[P2_1] poz.21 musi być wypełniona (czyli >0)
4		2. Karta statystyczna strajku	d1p7_2	[P7] poz.7 musi być wypełniona (tylko jedna z poz.71 lub poz.72)
5		2. Karta statystyczna strajku	d1p8_1	[P8_1] poz.81 musi być wypełniona (>0)

Timing of edit rules – The question is how the edit messages should be presented for their maximum effect. The possible solutions can be: present the message immediately after the field has been left, after the page was filled or at the end of the questionnaire entry. Immediate edits allow the respondent to correct the error straight away and can prevent similar mistakes later on (Skentelbery and Davies, 2012). From the other side, edits involving more than one variable raise the issue of waiting with edit execution for last variable completion. Whatever the case, usability studies point to the expectation of users that the form checking can be run iteratively. Another case is to prevent the user from entering some sort of keys, for example permitting only numeric keys. Programming formatting edits are examples of editing to prevent errors. This kind of edit checks should be triggered immediately. Edit rules should also be executed to reflect relationships between two or more variables. Such actions should be deferred, which requires additional functionality, where the user should be able to manually start an editing action as a batch operation. This, in turn, raises the question of whether the editing action should be triggered at the moment of completion or when the questionnaire is submitted over

the internet. Usability tests discourage this last solution, since performing actions that combine multiple functions is perceived as confusing (Anderson et al., 2005).

2.5 Testing and evaluation of editing strategy

Usability testing – Generally, usability testing results suggest the need for a good visual questionnaire design that uses fonts of different size and colour for questions and answers, can facilitate the answering process and reduces the completion time (Hansen and Couper, 2004). Though usability tests have their limitations, as they are conducted on a small number of users and try to test an entire questionnaire, not only the editing aspect, they can be a source for best practices for designing edit rules.

Analyses of collected data – Business surveys have a longitudinal nature. This grounds the possibility to evaluate the data collection instrument. A way to evaluate the set of built-in editing rules can be the number of non-response items. An issue when respondents tried to fit values to the upper bound of a range edit when it exceeded the range (Anderson et al., 2005) can be an example for tracking too rigorous edit rules in questionnaires.

User's centred design -– Usability principles advocate the basic rule: user needs should be at the centre of the design. All tasks to be performed should be under the user's control. Throughout the response process, during the data entry stage, edit messages can appear several times and in various forms. The user needs to be able to choose the right moment to deal with them and to ensure the action taken is effective, which requires inter-connectivity between edit messages and the item back and forth as desired. The policy on how data with unresolved items submitted will be treated should be included in the instruction manual. It should be clear whether data marked as erroneous can be submitted. In other words, the question is whether strict conformity to edit rules should be required or rather whether users should be allowed more freedom in this matter, which will make them more likely to provide data, thus reducing the rate of non-response. The principle of emotional design (Norman, 1990) states that errors can result from various sources. This calls for a consistent design that accounts for the possibility of various errors. Another purpose of design is to counteract errors.

The burden – Incorporating edit rules into the questionnaire does not necessarily increase response burden (Anderson et al., 2005). Usability studies showed that some automatic checking of data entries are awaited by users to be performed by a computer. If the goal of edit checks is clearly understood by respondents the tolerance and acceptance for them can be easier gained. The limit for the scope of edits can be drawn from usability testing. The aim of edit rules is to improve data quality and not to encourage non-response.

Testing proposals – Skentelbery and Davies (2012) give good examples of testing online edits set-ups in their paper "Editing Challenges for New Data Collection Methods". They bring up the research stating that for obtaining quality data the paging questionnaire design is the best option bearing in mind that two approaches are possible: paging and scrolling survey design.

2.6 Electronic documents

Typically, electronic processing involves implementing algorithms performed by a computer. An electronic questionnaire is simply a computer program. It seems useful to create a universal system, understood as a prototype program that could envelope a set of statistical variables and their validity

rules. In this way, variables and edit rules are combined, which gives shape to the definition layer of the output questionnaire. The questionnaire itself is designed to be a complete electronic document. This is why, a unified system combining the outer and inner part of the questionnaire should be created. The outer part refers to a computer program executed by the respondent, regardless of whether it is executed locally or remotely. The inner part, the core of the system, comprises the questionnaire definitions. The new technology supplies powerful tools that could be used to create such a unified solution. The extensible mark-up language seems to be well suited to the purpose of defining structural documents.

2.7 Conclusions

The goal of incorporating edits into the electronic questionnaire is to decrease measurement error in surveys. In the context of business surveys their unique features should be remembered when adopting a strategy for data collection editing. First, the response process is more burdensome than in social surveys. The data most commonly reside in business records and their retrieval requires time and effort. The response of a single unit may have an influential character. This is related to outliers. Some types of editing may require an aggregate level outlook. These features determine the types of edits used in questionnaires and also the scope of them. Data may be received with unresolved edit checks in order to avoid non-response. The compulsory requirement of edits resolving may be reserved for a "critical" set of items (Anderson et al., 2005). On the other hand the need for continuous evaluation of data collection instruments can be an opportunity for improvements. The crucial principle can be drawn from the usability principles that put the user control of the response process at the core of the design.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Anderson, A., Murphy, E., Nichols, E., Sigman, R., and Willimack, D. (2005), Designing interactive edits for U.S. electronic economic surveys and censuses: Issues and guidelines. Proceedings of UNECE Conference of European Statisticians, Ottawa, Canada, May 2005.

Hansen, S., Couper, M. (2004), Usability Testing to Evaluate Computer-Assisted Instruments. In *Methods for Testing and Evaluating Survey Questionnaires*, Chapter 17, Wiley, New York.

Haraldsen, G. (2013), Questionnaire Communication in Business Surveys. In Snijkers, G., Haraldsen, G., Jones, J., and Willimack, D., *Designing and Conducting Business Surveys*, Chapter 8, John Wiley & Sons.

Holmberg, A. (2002), Pre-printing Effects in Official Statistics, an Experimental Study. *International Conference on Questionnaire Development, Evaluation, and Testing Methods*, Charleston, SC.

Mockovak, B. (2005), An evaluation of different design options for presenting edit messages in web forms. Bureau of Labour Statistics.

Murphy, E., Nichols, E., Anderson, A., Harley, M., and Pressley, K. (2001), Building usability into electronic data-collection forms for economic censuses and surveys. *The Federal Economic Statistics Advisory Committee 2001 Conference*.

Norman, D. (1990), *The Design of Everyday Things*. DoubleDay.

Phillips, J. M., Mitra, A., Knapp, G., Simon, A., Temperly, S., and Lakner, E. (1995), The Determinants of Acquiescence to Preprinted Information on Survey Instruments. *Proceedings of the Survey Methods Research Section*, American Statistical Association, 1169–1171.

Schonlau, M., Fricker, R. D., and Elliott, M. N. (2002), Guidelines for designing and implementing internet surveys (chapter five).

Skentelbery, R. and Davies, C. (2012), Editing Challenges for New Data Collection Methods. Working Paper No. 18, Work Session on Statistical Data Editing, Oslo, Norway, 24-26 September 2012.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, New Jersey.

Interconnections with other modules

8. Related themes described in other modules

1. Response – Response Process
2. Statistical Data Editing – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Questionnaire Design-T-Editing During Data Collection

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	13-03-2012	first version	Paweł Lańduch	GUS (Poland)
0.2	31-03-2013	second version	Paweł Lańduch	GUS (Poland)
0.3	10-12-2013	third version	Paweł Lańduch	GUS (Poland)
0.3.1	20-12-2013	preliminary release		
0.4	18-02-2014	version revised after EB review	Paweł Lańduch	GUS (Poland)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:27