UNITED NATIONS STATISTICAL COMMISSION and WP. 35
ECONOMIC COMMISSION FOR EUROPE ENGLISH ONLY

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iv): Data editing by respondents and data suppliers

"Outsourcing" of plausibility improving measures
Supporting Paper

Submitted by the Federal Statistical Office, Germany [1]

I INTRODUCTION

1. The collection of information belongs to the main tasks of statisticians. It can be performed in several ways: the most exact method is the technical measurement which ensures objective information on an reality excerpt. Unfortunately many interesting objects cannot be measured exactly so that statisticians have to use other means like face to face or self-interviews and the use of existing data / information. In opposite to the exact technical measurement all of the last mentioned alternatives lead to more or less influential errors in statistical data in the beginning of the statistical production process and influence last but not least the data quality in terms of accuracy and timeliness on one hand and on the other hand the efficiency of survey processing. Continuous demands for qualitative higher statistical results under the condition of limited resources increase the pressure for the optimisation of the data collection process.

2. Being aware of the tremendous effects of erroneous answers on data quality and efficiency of the statistical production processes statisticians have ever tried to optimise the data collection process[i] with special focus on the data collection instrument. It should be mentioned the use of sophisticated questionnaire testing methods and data collection strategies as well as the use of progressive information technology (IT). Well known examples are the use of electronic questionnaires often for household surveys from 1985 on. In opposite to that the use of technical equipment for the reception of answers for establishment surveys was very seldom in the past. One main reason was the effort for the administration/performance of computer assisted (self) interviews. With the further dispersion of the Internet the situation for computer assisted interviews improves tremendously and offers statisticians better possibilities for the collection of plausible data in establishment surveys. Electronic questionnaires submitted via Internet are nowadays regarded as state of the art. An other important development affecting this area is the wide spread use of IT in enterprises, public services and households, which offers statisticians better possibilities to integrate plausibility checks in commercial software or the direct use of administrative / business data for the production of statistical results.

Due to the increasing importance of the data winning process and the fact, that the SDE work session discusses this topic for the first time the aims of the contribution are to:

[1] Prepared by Elmar Wein, elmar.wein@destatis.de

- describe the data winning process including it's typical errors,
- describe potential outsourced measures with regard to data editing specific aspects – supplemented by examples,
- facilitate the clarification of terms,
- give recommendations concerning future work.

3. The term "data winning process" will be regarded as a generic term for a process where statistical relevant data or information will appear for the first time. With regard to the topic it seems to be useful to distinguish the data winning process by the kind of the origin of the data because of many different aspects. In the case of primary statistics it is called the *data collection process* where the aim is to receive special information for a specific survey. In the case of secondary statistics it is proposed to call this process the *external data winning process*.

II PLAUSIBILITY IMPROVING MEASURES FOR DATA COLLECTION PROCESSES

II.1 The data collection process

4. The data collection process is – depending on the data collection mode and type of instrument – largely under the control of statisticians and enables them to establish a wide range of – let's say – "plausibility improving measures". It consists of several sub processes or work packages and begins with the preparation of the data collection i.e. the choice of the respondents, the announcement of a survey, the data collection phase itself, the checking of answers, the transfer of answers / data to a statistical office, the remainder of missing respondents on the basis of completion checks. The process ends when a sufficient level of data / information has been received and transformed in a computational format. This special end of the process seems to be surprising at a first glance but it represents best practices and facilitates comparisons between electronic and paper-and-pencil-questionnaires.

5. Many types of errors may occur during the data collection process. With regard to this topic the definition of errors is restricted on *obvious* implausible data which means that these errors can be detected by checks. So deteriorations of distributions due to interviewer effects are out of scope.[ii] Typical errors made by respondents are: misunderstanding, incomplete answers, bad memory, coding errors like misclassification. Other sources of errors are incorrect interviewer activities i.e. inadequate choice of respondents, routing errors in questionnaires, deviations from original question and answer texts, and incorrect signing and coding.[iii] Last but not least errors may also be caused by statisticians in a statistical office while entering data and coding.

II.2 Optimal preconditions for the employment of plausibility improving measures

6. The improvement of the plausibility of answers is a domain of the questionnaire development. The upcoming of modern information technology in the eighties enables the integration of plausibility improving measures in data collection instruments and lead to a cooperation of questionnaire developers and data editing methodologists. In principal it seems reasonable to make a difference between questionnaire development and data editing methodology in such a way that questionnaire development encompass all interactions between respondents and data collection instruments or rather interviewers. The aim of questionnaire development shall ensure that respondents will report the truth. The area of data editing begins *after* giving an answer – even in the data collection process and ends when the interaction with a respondent / interviewer starts again which means in the moment when an error message is displayed. While introductory information, question and answer texts belong to the area of

questionnaire development, the integration of "plausibility improving measures" falls in the responsibility of data editing methodologists. As error messages and instructions for corrections represent dialogues with respondents an intensive cooperation between questionnaire developers and data editing methodologists is necessary in such a way that questionnaire developers should also improve error messages and correction instructions. Summing up all aspects it seems useful to give the last responsibility for a data collection instrument in the hand of the questionnaire development. However the proposed distinction will be used for theoretical reasons and perhaps it may help to clarify responsibilities in practice. The mentioned examples represent only a small part of the cooperation between questionnaire development and data editing methodologists.

7. Questionnaire developers possess a wide range of methods for the optimisation of data collection instruments. One of them is the pretest which represents the test of a new/redesigned questionnaire preferably by critical respondents. The focus of pretest techniques is set on a respondent's behaviour on question texts and should be extended in the case of electronic questionnaires on included checks, error descriptions and instruction for corrections.[iv] Modern IT-tools offer the possibility of recording all data entries and may therefore deliver valuable information on errors.[v]
As the size of a pretest sample exceeds in dependence of a chosen pretest design very seldom the amount of 50 respondents a pretest report can only deliver hints on possible errors before a survey starts. Advanced pretest designs consists of two phases where the second phase contains the test of an optimised questionnaire.[vi] These tests may deliver more reliable information on possible errors. Later developments of pretest methods lead to an augmentation of pretest samples. So pretests become more and more an important source of information for the planning of data editing activities.

So it is absolutely necessary that data editing methodologists participate in the analysis and optimisation of electronic questionnaires.

8. While it is good practice to test a new questionnaire pretests – regarded as post-tests – should be initialised by data editing activities. Statistics on errors from the data cleaning process deliver valuable information for the optimisation of existing questionnaires. A very simple but efficient way to retrieve the respective information is the comparison between raw and plausible data. A crucial point of this analysis is the necessity of a flag which indicates whether a record is based on an electronic or paper-and-pencil-questionnaire because this information helps to optimise the choice and integration of checks in electronic questionnaires.

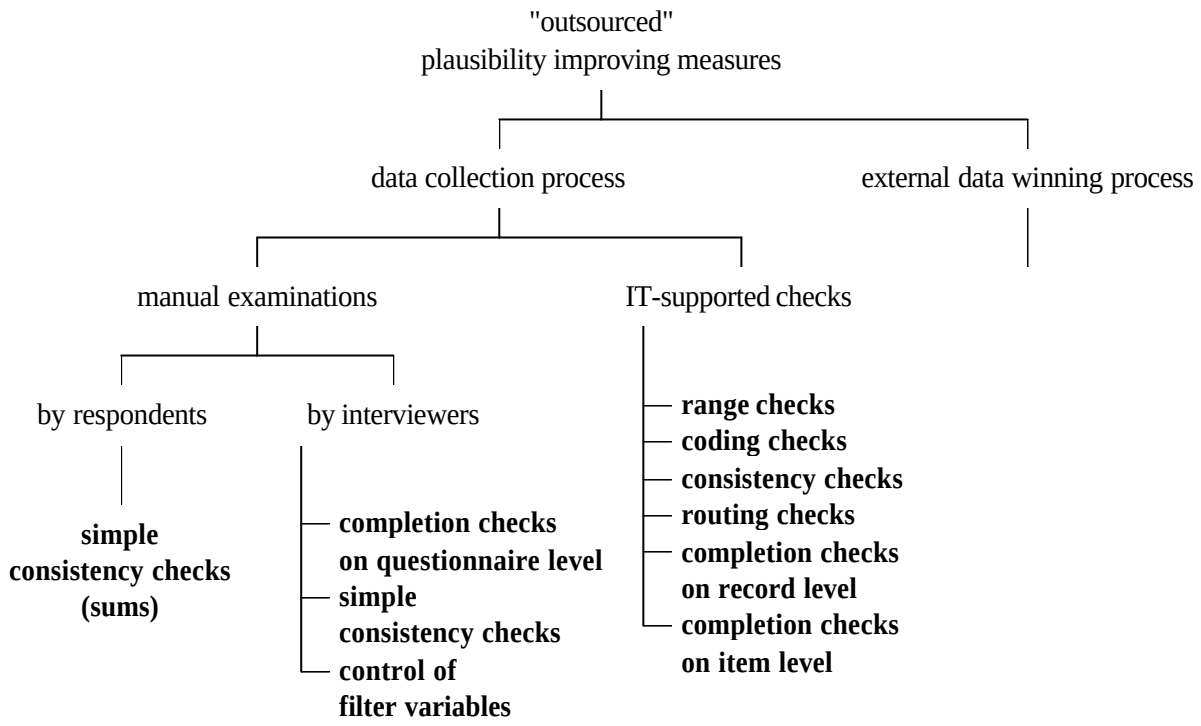
9. Another important measure – not belonging to questionnaire development – is the pilot study which represents the test of a survey with a smaller sample under real conditions. The focus is set on organisational and computational aspects and less on the data collection instrument. Due to the larger sample size a pilot study delivers more detailed information on types and amounts of errors and the effort for their correction. So a report of a pilot study should contain information on:

- the type and number of occurring errors,
- the effort needed for their correction,
- inappropriate checks, descriptions and instructions,
- recommendations for an optimisation of an existing data editing strategy.

II.3 Plausibility improving measures as part of the data collection process

10. The plausibility improving measures which can be integrated in the data collection process can be summarized in the following figure:

Figure 1: Plausibility improving measures of the data collection process



11. Obviously outsourced plausibility improving measures of the data collection process cover a wide range of activities which have the following common attributes:

- they are performed during the data collection by respondents/data suppliers[(2)],
- they require the existence of data or information,
- they originally belong to the methods of data editing.

The previously mentioned examples demonstrate the heterogeneity of the different methods. So it is proposed to designate them as “measures”. The fact that respondents perform the corrections classifies them as “outsourced” measures. They include checks during a computer assisted telephone interview because these corrections are initiated by respondents.

12. The use of plausibility improving measures in the data collection process depends on the survey contents, the type of the data collection instrument, the data collection mode, and the abilities of the respondents. In general a distinction is made between the use of paper-and-pencil-questionnaires for self- or face-to-face-interviews, and the use of electronic questionnaires in combination with self-interviews (CASI) or CAPI and CATI[(3)].

13. Electronic questionnaires can be provided via Internet or CD-ROM for self-interviews as well as on notebooks for face-to-face-interviews. The medium used for the provision of an electronic questionnaire (still) determines it’s size and the number of implemented checks as lot’s of respondents don’t possess a fast access to Internet or use less powerful IT-equipment. So there is often a challenge to find an optimum between the number and design of checks and acceptable download / initialisation times. Figure 2 contains a screenshot of an Internet questionnaire provided by Destatis for EU Intrastrade reports via the Internet:[vii]

[(2)] The term keeps in mind the situation of secondary statistics with different processes.

[(3)] CAPI: Computer assisted personal interviewing, CASI: Computer assisted self-interviewing, CATI: Computer assisted telephone interviewing.

Figure 2: Internet questionnaire for the EU Intra-trade reports via the Internet

II.3.1 Plausibility improving measures in paper-and-pencil-questionnaires

14. Paper-and-pencil-questionnaires in combination with self-interviews enable only few possibilities for the improvement of the plausibility of answers. Examples are instructions to check answers or auxiliary fields for the computation of sums. They represent weak means because their use and power depends to a great extent on the cooperation of the respondents.[viii]
Another measure may be the integration of additional, sound characteristics which facilitate the correction of errors in subsequent data editing processes.

15. Paper-and-pencil-questionnaires in combination with face-to-face-interviews provide better opportunities for the implementation of plausibility improving measures. Concerning the contents of interviewer checks one should pay attention to the fact that an interview creates a kind of tension between respondents and interviewers which improves the willingness to give answers. It should not be disturbed by too long breaks due to complicate checks. Checks on the consistency of different characteristics represent in this context a limit of interviewer checks. So interviewers should concentrate preferably on completeness checks and characteristics which are essential for the data editing process i.e. routing variables or critical topics of a questionnaire.

II.3.2 Implementation of checks in electronic questionnaires

16. An electronic questionnaire generally offers good opportunities for the implementation of checks. Subject matter statisticians often tend to integrate as much checks in electronic questionnaires as possible while they ignore the risk of an increasing refusal rate. Lots of checks regularly increase the respondents' burden which may finally endanger the obligation to respond. In opposite to that a questionnaire without any checks is not conform with the state of art which is often defined by dynamic Internet applications. Furthermore it may diminish consumers' confidence in statistical results.

17. One can generally say that all activities like automatic corrections should not be integrated in electronic questionnaires because they lead to the impression that the respondents' answers are irrelevant.

18. With regard to the decision on the integration of checks in electronic questionnaires they are assigned to the structure plausibility and interplausibility.[ix] Checks which are used to improve the

structure plausibility are coding checks, range checks, completion checks on item and record level, and checks which ensure a correct routing through a questionnaire. These checks are similar to those which are implemented in common homebanking software or Internet shopping applications and their integration seem to be generally smoothly.

19. So the area which remains problematic deals with the integration of consistency checks that permit or forbid certain combinations of characteristic values. Their integration in questionnaires depends on numerous aspects:

- technical equipment of respondents / interviewers
  Checks increase the amount of bits to be downloaded from the Internet or to be initialised on computers. So it is good practise to assume a less powerful IT-equipment of respondents. This assumption ensures that an adequate number of checks is integrated in an electronic questionnaire.
- possibilities of navigation in an electronic questionnaire
  A consistency check which requires an elaborate navigation of twenty positions back to a previous question should not be integrated in a questionnaire if there is no powerful navigation available.
- avoidance of confrontation with information given during a previous round of a survey
  In general it is convenient to find a web form filled out with once given administrative information. If you confront respondents with previous answers (especially in checks) you create an impression of an overall watching.
- complexity of an error
  It is often influenced by the number of involved characteristics and the required subject matter knowledge for a correction. The decision on the integration of a consistency check depends on the involved characteristics, the needed detailed subject matter knowledge for a correction and the respondents who fill out the questionnaires. However a practical recommendation may be to implement only consistency checks, which compare the results of not more than three characteristics on a high abstract level.
- avoidance of follow-up errors
  Do some (consistency) checks to be performed during the data collection process hinder following, more complicate errors which could only be discovered by complex checks in a statistical office? Here we see a need for further research on an optimal choice of checks for the integration in electronic questionnaires.
- output oriented implementation of checks
  Consistency checks shall ensure the provision of (real-time) statistical results at a very early phase of the statistical production. This consideration is a reaction on the increasing demand for up-to-date statistical results.
- facilitation of the statistical production
  Checks are implemented to improve the plausibility of characteristics which function like “anchors” in a following internal data editing strategy.

It is obvious that the aspects may lead to inconsistent choices of checks to be implemented. So subject matter statisticians, questionnaire and data editing methodologists should made decisions which take into account the specific demands of the corresponding statistics.

20. Besides the discussion of the number and complexity of checks a point of discussion is their placement in electronic questionnaires. If an answer will be checked immediately the respondent is still familiar with her answer and can easily reach the position in a questionnaire. One disadvantage of this placement may be an increase of refusals especially when numerous and complicate checks occur in the beginning of an electronic questionnaire.

In opposite to that checks are often placed at the end of Web questionnaires that means before a filled out form will be sent to a statistical office. The assumption is that a respondent's disposition to refuse will decrease when she has completed a questionnaire. The main disadvantages of this location are a higher effort for navigation and higher demands on the descriptions of possible erroneous fields – if no links are provided.

21. Plausibility checks generate error messages which initiate further actions. With regard to their contents the following considerations may be helpful:

- The message should consist of an error description which is expressed in a neutral and objective way. Especially the use of exclamation marks can be considered as problematic. The error description should contain an unique error identifier which facilitates respondents' queries.
- An instruction for correction should complete the error message. An important aspect is that it should not influence the correction. A crucial point may be the sequence of the listed characteristics which are involved in a check . It should vary from respondent to respondent because they tend to pick up the first alternative.
- Understandable descriptions of the involved characteristics and possibilities for an easy navigation, e.g. hyperlinks, should complete the messages.

22. Checks in electronic questionnaires may lead to a higher burden for respondents which can sometimes hardly be compensated by a dynamic routing, the use of existing information, and a context oriented help. So there is a need for incentives to enhance respondents' readiness for cooperation. A successful approach may be to give respondents some incentives like a temporary access to statistical data which are relevant for their management activities.

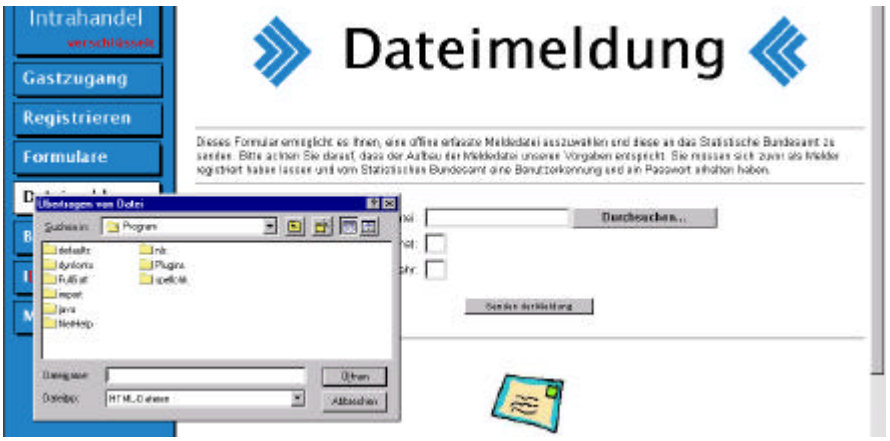
23. In practice statisticians want to summarize all checks which are integrated in electronic questionnaires. So we would like to call them "questionnaire checks".

24. Statisticians expect a mix of data collection modes for the next 10 to 15 years.[x] Some respondents use electronic other paper-and-pencil-questionnaires. So there is a need to check given answers from paper-and-pencil-questionnaires. In this context the question often arises: How should checked data, collected by electronic questionnaires, be handled in the following production process of a survey? Some statisticians prefer a general new checking because this step leads to a unique treatment of all statistical data. Others refer to the fact that the data have been cleaned. An additional effort and time losses have to be quoted against a better control over errors.

III PLAUSIBILITY IMPROVING MEASURES FOR DATA DELIVERIES

25. Besides the collection of data for statistical purposes there is a growing tendency to compute statistics on the basis of existing registers. The wide existence of IT-equipment in establishments leads to a growing electronic supported data winning: the increased use of accounting information of enterprises for statistical purposes – better known as electronic data interchange (EDI). Electronic data interchange seems to be a mixture between data collection and data delivery as this "data collection mode" is mostly offered in the context of surveys. This way of data transfer is possible in the case of a high compatibility between survey contents and available accounting information. As there are lots of similarities to secondary statistics it will be summarized under this chapter. These different ways are summarized under the term "data deliveries". An example for electronic data interchange is the data transfer possibility within the w3stat system. Figure 3 contains a dialogue box for the transfer of data from an enterprise to Destatis:

Figure 3: EDI dialogue box as a part of the w3stat system



26. The respective process where the data is produced is called the "external data winning process". One specific problem to be handled during the planning of data editing is the judgement of the external data quality. A second specific aspect of secondary statistics is to get more control over the external data winning process. At first the external data winning process in the case of a secondary statistic will be described.

III.1 The external data winning process

27. It begins with the completion of forms or making of data entries in a database and ends sometimes with the checking of delivered data. It differs from the data collection process because relevant data or information is completely won without control of statisticians. The only part which often remains under the control of the statisticians is the data delivery including, completeness and consistency checks of delivered data / information. The missing control over this process may induce problems for the planning and performance of data editing.
Besides this way of external data winning the collection of information via forms represents more and more an exception.

28. Typical errors which may occur during the external data winning processes are often data delivery errors especially in the case of first deliveries due to discrepancies between the real and the data structure agreed upon. Additional data editing problems are caused by the fact that (German) public services are only allowed to collect a minimum of information which is absolutely necessary for administrative activities. Additional information, which may facilitate data editing, is simply missing.

III.2 Judging the quality of external data / information

29. Statisticians often can not influence the data quality but in many cases they can receive information about it which may facilitate the planning of data editing. Useful indicators are in general:

- incomplete and different meaning of the data / information
  The meaning of the data / information may be streamlined to meet the specific needs of an administration / enterprise, or it is simply missing. Other problems are the use of different classifications or the aggregation of information. Existing (legislative) descriptions may give useful information on the meaning of the data.
- use of data / information
  The use of the data / information heavily influences the data quality. One can expect for in-

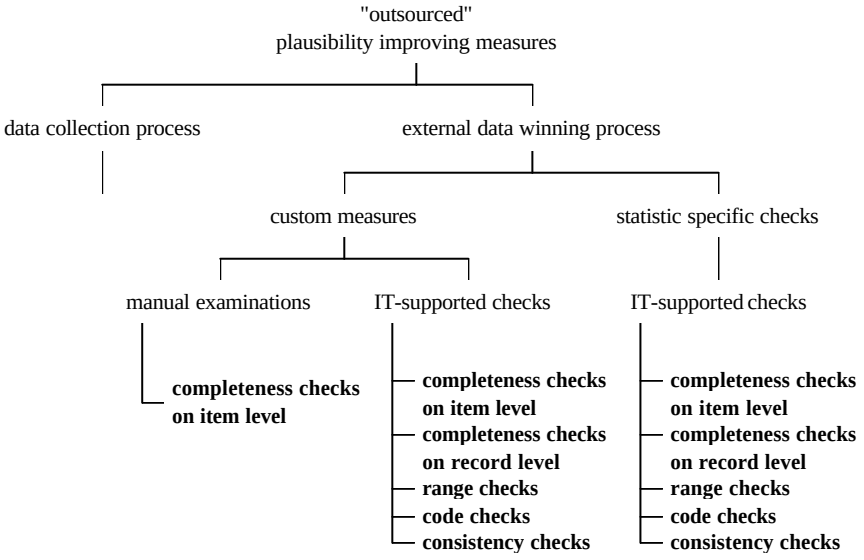
stance a high level of data quality if the data / information is used for financial decisions or expenditures. On the other hand one can assume less incentives to achieve and maintain a high data quality if the number of citizens in a register determines the income of a mayor.

- organisational aspects
  This topic covers aspects like the representation and administration of a register (centralised versus decentralised) in combination with coordination mechanisms in the case of decentralised registers, plans for a periodic register maintenance and the qualification of the personnel who handles the register.
  An important aspect in the case of a decentralised register is the existence of an uniform and unique index used for the summing up of records.
  Other organisational aspects cover the data winning mode: direct data entry in a database in combination with checks versus the use of forms at first and data entry in a subsequent phase.
- information technology
  The crucial point of this topic is the question whether data / information is checked or not. Information on checks which are implemented in a used software, can be retrieved from the respective manuals or developer / user of the software.

III.3 Outsourced plausibility improving measures

30. The outsourced plausibility improving measures of the external data winning process can be summarised in the following figure:

Figure 4: Outsourced plausibility improving measures of the external data winning process



31. It becomes clear that the range of the (IT-supported) checks of the external data winning process resembles the ones of the data collection process. Furthermore IT-supported checks as custom measures are highly identical with statistic specific checks. One problem in this context is: “How reliable are they?”.

32. The possibilities of statisticians to integrate statistic specific checks in commercial software are limited because their integration causes additional effort. It is in general very important to improve the benefit for the data suppliers to find out common areas of interest, or to overtake the additional effort.

33. Means to improve the quality of information / data are the provision of statistical software or integration of statistical IT-modules in commercial / administrative software. One example is the program "FLIRT*FRA" which was developed on behalf of Fraport public company in cooperation with the

consortium of German airports and Destatis. The system generates automatic flight reports which fulfil the demands of airlines and airports as well as the ones of German official statistics.[xi]

34. The most effective arguments for the conviction of data suppliers to implement statistical software (with checks) or to integrate further checks are: the reduction of the burden for enterprises as an argument for the purchase of a software, a higher data quality with positive effects for the specific work of establishments, and incentives like the temporary right to use statistical databases.

35. Finally it may be useful to define terms for different checks: "statistic specific checks" are specified by statisticians and in opposite to that "data supplier specific checks" are specified by externals.

IV SUMMARY AND CONCLUSIONS

36. The wide spread use of IT-technology in establishments and households and the further dispersion of the Internet offer statisticians better possibilities to receive plausible data. The integration of checks in electronic questionnaires is determined by interactions with respondents but their integration related to aspects of the statistical production is nevertheless very important too.

37. The UNECE conference discusses this topic for the first time. So it seems to be practical to start with the description of it's borders and necessary enhancement of the data editing specific terminology and continue the discussion in the most promising areas.

38. The winning of data / information for statistical purposes will be determined by the following developments in the next decade of years:

- The wide spread use of information technology will increase the electronic data interchange.
- There will be a mix of instruments used for data collection with an increasing part of electronic questionnaires and a decreasing one of paper-and-pencil-questionnaires.

The consequence of these developments is a need for further research …

- on the integration of improving plausibility measures in the data collection process with regard to a better coordination between plausibility improving measures as well as checks in electronic questionnaires. Another direction may be the optimisation of the integration of checks in questionnaires to improve the preconditions of the production of statistics.
- on the handling of non response in the case of a mixed data collection mode.

---

[i] Don A. Dillman (1978). "Mail and Telephone Surveys: The Total Design Method". New York

[ii] More information on effects of interviewer behavior and data collection mode on data quality can be found in: Karl-Heinz Reuband (1998). "Der Interviewer in der Interaktion mit dem Befragten – Reaktionen der Befragten und Anforderungen an den Interviewer", in: "Interviewereinsatz und -qualifikation", Destatis (Eds.), Wiesbaden.

[iii] Lars Lyberg, Daniel Kasprzyk (1991). "Data collection methods and measurement error: an overview", in: "Measurement errors in surveys", Paul P. Biemer et al (Eds). New York

[iv] Erwin K. Scheuch (1996). "Die Notwendigkeit von Pretests zur Vorbereitung statistischer Erhebungen", in: "Pretest und Weiterentwicklung von Fragebogen", Destatis (Eds.), Wiesbaden.

[v] Blaise Manual.

[vi] Manfred Ehling, Rolf Porst (1996). "Pretest zur Erhebung Neukonzeption der laufenden Wirtschaftsrechnungen", in: "Pretest und Weiterentwicklung von Fragebogen", Destatis (Eds.), Wiesbaden.

[vii] Destatis (2001). "Meldung zur Intrahandelsstatistik mit w3stat via Internet", in: Methoden-Verfahren-Entwicklungen (1/2001), S. 7ff.

[viii] Destatis (2003). Internal manual for the development of questionnaires. Wiesbaden.

[ix] Destatis (2003). Internal manual for the planning, performance and optimisation of data editing (draft). Wiesbaden.

[x] Erich Wiegand (2000). "Chancen und Risiken neuer Erhebungstechniken in der Umfrageforschung", in: "Neue Erhebungsinstrumente und Methodeneffekte", Destatis (Eds.), Wiesbaden.

[xi] Destatis (1996). "Luftfahrtstatistik – Ein Beispiel von Rationalisierung durch PC-gestützte weitgehend automatisierte Flugberichtsgenerierung auf den Flughäfen.", Methoden-Verfahren-Entwicklungen (2/1996), S. 3ff.