

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Vienna, Austria, 21-23 April 2008)

Topic (i): Editing of data acquired through electronic data collection.

**USING “TRADITIONAL” CONTROL (EDITING) SYSTEMS TO REVEAL CHANGES WHEN
INTRODUCING NEW DATA COLLECTION INSTRUMENTS**

Supporting Paper

Prepared by Sindre Børke, Statistics Norway¹

I. INTRODUCTION

1. Statistics Norway (SN) collects business data through self administered questionnaires in more than 50 surveys, close to one third in each group; Month, quarter and year. Almost all questionnaires have a paper version. Some electronic reporting was established earlier, but since 1.july 2004 there are possibilities to respond electronically to “all” surveys.
2. Electronic reporting is (for the purpose of this paper) divided into two main categories;
 - Reporting of electronic data (electronic transfers) from enterprise systems, and
 - filling in an electronic, usually web-based questionnaire.
3. There are possibilities to do control/editing when respondents answer and send questionnaires, and when they send data from their enterprise system. The intuitive expectation of this outsourced editing is that the quality of data collected will be better, but the burden on the respondent will increase. In the big picture, we want to measure quality and burden, and pinpoint the optimal outsourcing level. This optimal level may be different in the two categories of electronic reporting.
4. The setup of electronic transfer is done to reduce/avoid manual work when sending. Data, and even competence, needed for correction of incorrect answers is perhaps not present with the actual respondent. If we implement controls at the time of sending data, the respondent will probably feel high burden trying to solve problems identified.
5. Transfer of electronic data will call for changes (supplements) in editing systems in SN. Electronic transfers of electronic data will imply new types of errors, and new aspects of control. Data quality will be very dependent on correct software in the enterprise systems, and stringent (common) use, alternatively local adjustments of the software. This introduces a need to keep under surveillance also updating of software used.

¹ The paper is heavily based on earlier, internal presentations in Statistics Norway, the most important of them being Gustad (2007)

II. THE SCOPE OF THIS PAPER

6. Electronic questionnaires are as a basis handled in the same procedures as paper questionnaires when it comes to controls. But the electronic versions open for blocking of obvious wrong answers and asking the respondent to rethink if answers seem inconsequent or outside a defined range. Using advanced solutions, controls can be done also against known data from sources outside the questionnaire. The main task for this paper is to study whether there are more or less errors in web than in paper questionnaires, no controls implemented, and effects of controls implemented in electronic versions.

7. As far as possible, electronic data to be used in reporting for statistical use should be data also used in the enterprise's own administration. If so, we will expect the data to be of high quality. Even if the definitions in the enterprise system are as asked for, the data in the system still might be wrong. This problem occurs also when the data is found in the system and entered in a paper or web questionnaire, but doing a manual job, at least fundamental errors may be identified by the respondent. We have not had the resources to do a thorough study of data transferred from enterprise systems, but present editing data from one survey, giving a glance into this area.

III. STARTING WITH A HOUSEHOLD CENSUS

8. The first, and so far best, study on differences between paper and web questionnaires in SN was done on the 2001 housing census in Norway (Haraldsen et al, 2002). For technical reasons on the respondents' side, they developed one simple version (scrollable "copy" of paper version, without controls) and one advanced version (one question in one page, and including controls).

Proportion of questionnaires with errors and average number of errors in paper and Web questionnaires.

	Proportion of questionnaires	Average no. of errors	N
Total number of errors	98,6	6,0	
Paper	98,6	6,2	(1858493)
Internet	91,7	3,3	(213334)
Simple Web version	97,4	4,8	(22345)
Advanced Web version	92,1	3,4	(64404)

9. These figures have to be considered against differences in respondents characteristics and perhaps also in differences in the dwellings they live in. As a starting point we have the perhaps expected picture; There are fewer errors in the web responses, and best results when including controls in the web-questionnaire.

IV. THREE DIFFERENT ANALYSES OF THE QUESTIONNAIRE ON ACCOMMODATION DATA (HOTEL STATISTICS)

10. From august (data for July) 2003, the questionnaire for monthly reporting of data to hotel statistics is offered on web. Increasing from a small number of respondents, this way of reporting exceeds 50% from the beginning of 2007. Some controls are implemented.

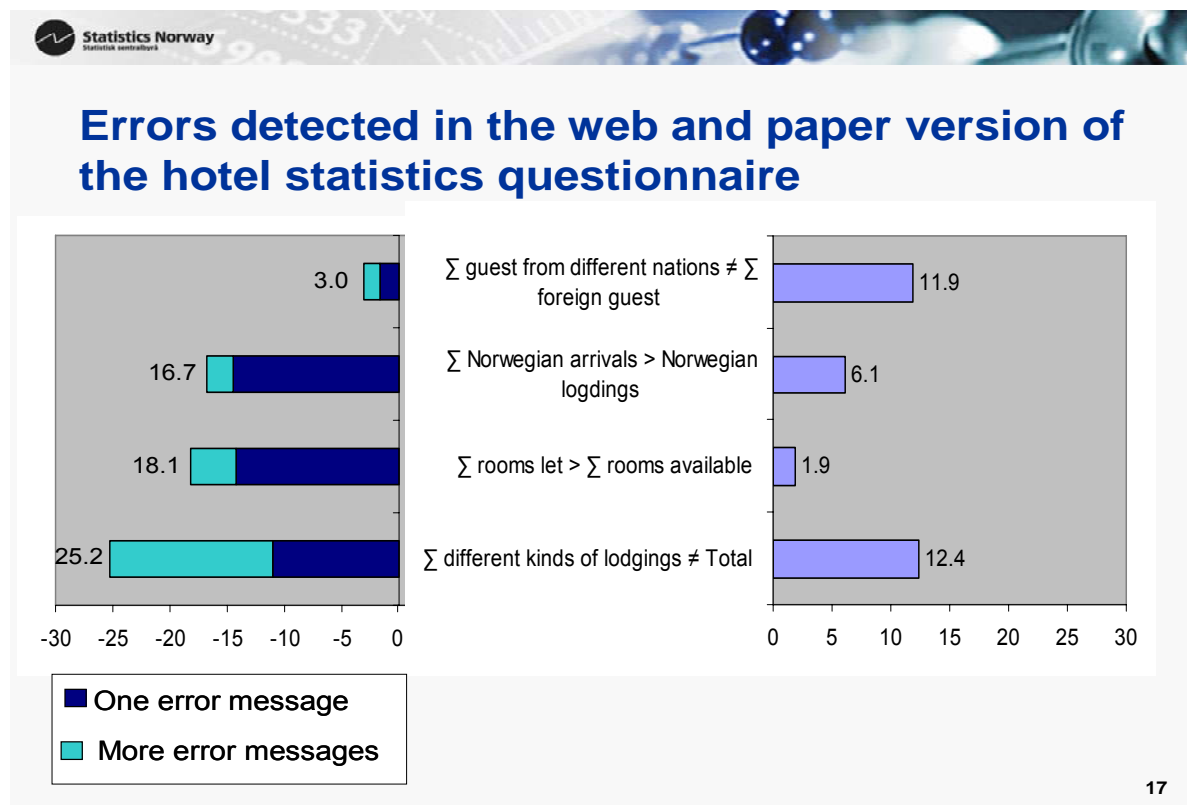
11. A brief analysis (Auno, 2004) of responses based on the total of all questionnaires from the first eight months with electronic questionnaire as an opportunity, found an average of 1.7 warnings per paper questionnaire. Corresponding figure for electronic questionnaires was 0.7. This was measured through the internal editing system run after receiving the data in SN. While 18% of the questionnaires were electronic, only 9% of what was defined as "possible errors", and 6% of the "definitive errors" detected were in these.

12. Thus, this analysis gives the intuitive expected result, web and outsourced editing gives better data.

13. SN has developed software that (among other) may report how often different controls in a web questionnaire are obstructed. Through that, we may measure not only what respondents send, but also if they were stopped or warned when filling in the questionnaire, the outsourced editing. (This paradata concept represents an important possibility in future work.)

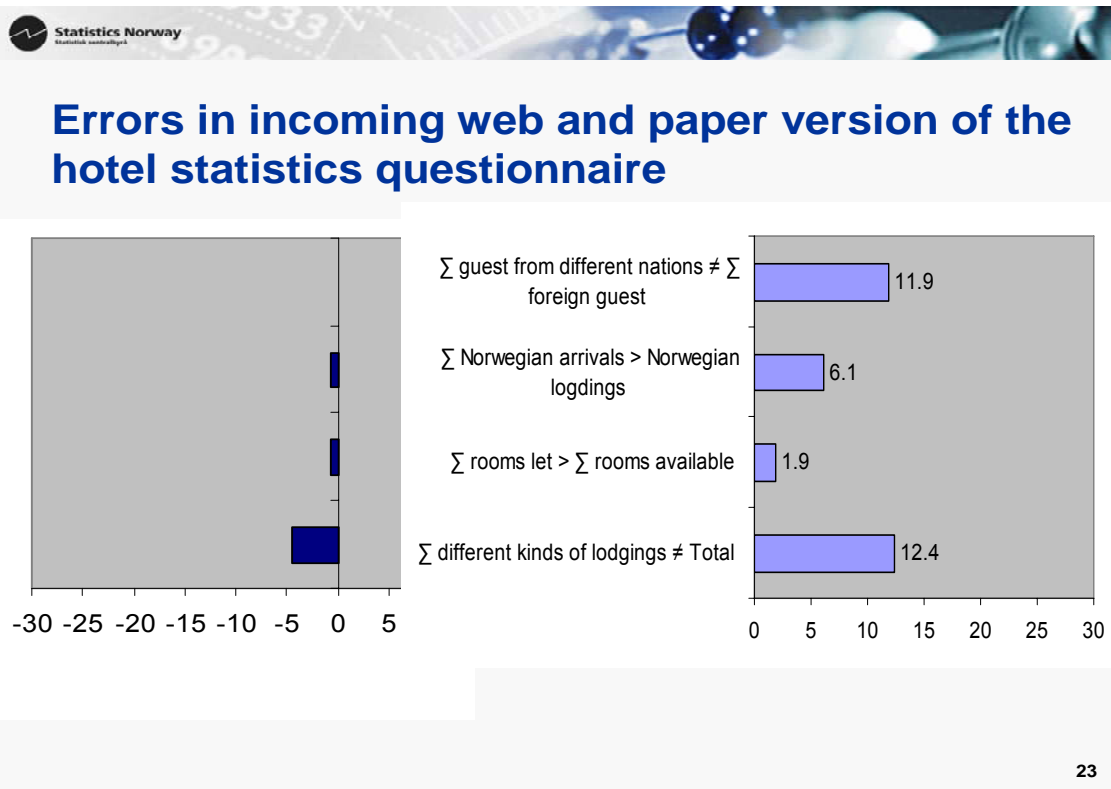
14. Based on data for hotel statistics, November and December 2005, we (Haraldsen et al, 2006) found that the number of erroneous answers in web questionnaire is considerably higher than in the paper version. This is the result when we look at the initial input in the questionnaire, but since errors are corrected before sending, we usually do not identify this problem.

15. As we can see, the number of rooms let exceeds the number of rooms available in 2% of the cases (and approximately 5% of the total needs more than one attempt to find acceptable answers).



Haraldsen et al, 2006

16. As shown below, almost all these errors are corrected before sending, the received data showing a pattern more like that from 2004. Our intension is to study not only the number of errors, but also the changes done in respondent's corrections. This information could be used in the editing process of paper questionnaires. This first study, however, indicates uncontrolled differences between use of electronic and paper questionnaires.



Haraldsen et al, 2006

17. As part of an analysis of the editing system for monthly data on accommodation (Gustad, 2007), we have data comparing errors found in some controls in the data from web and paper. (Data from September 2006.)

18. One control in the editing system picks out questionnaires where answers indicate that less than 10% of the rooms are let out. This control is not implemented in the web-questionnaire. 12% of the paper and 18% of the web questionnaires are marked.

19. Another control specifies the average price of a room to be between 300NOK and 1000NOK per night. The maximum limit is controlled in the web questionnaire, asking the respondent to check again to be sure the answer is right. After receiving all data at SN, the control picks out 41% of the paper questionnaires, and 36% of the web-questionnaires.

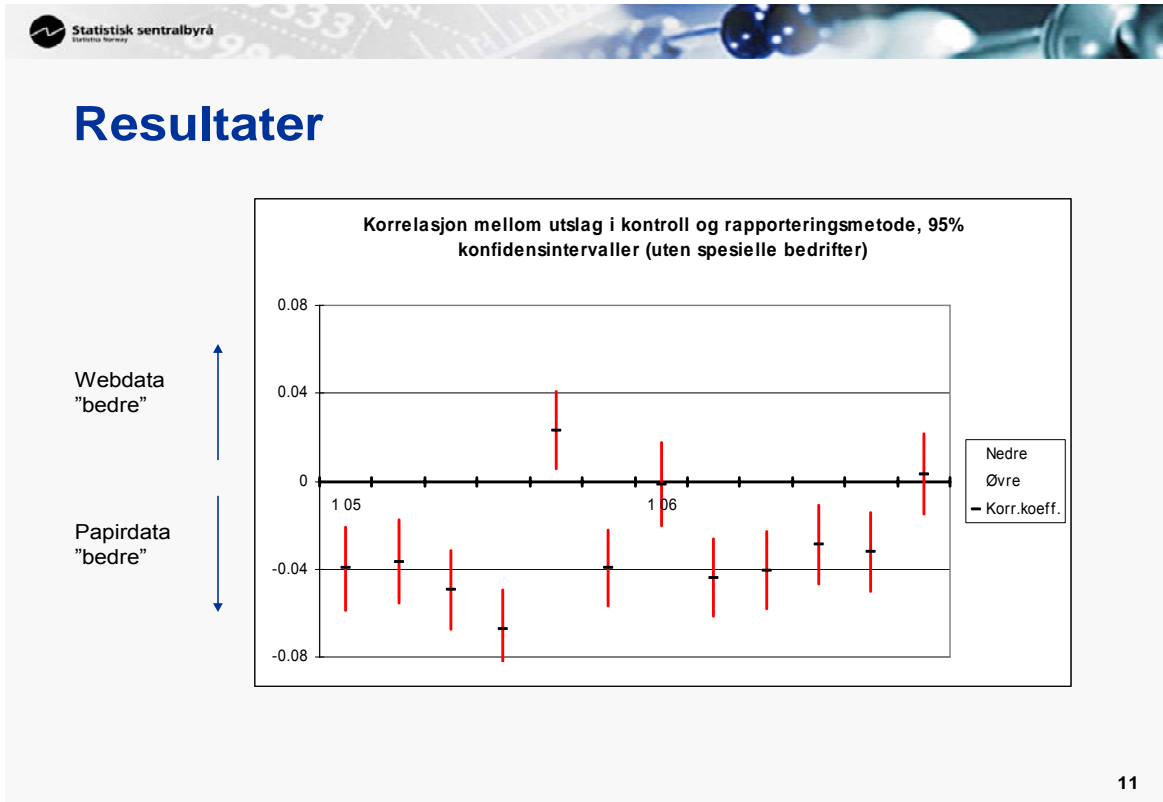
V. PRODUCER PRICE INDEX – MORE WARNINGS, BUT NOT MORE CORRECTIONS IN DATA FROM WEB QUESTIONNAIRES

20. Data collection for producer price index involves 850 companies every month. Following this data collection, we (Kristiansen, 2007) find approximately 45% using web in the beginning of 2005, and 55% at the end of 2006. There are no controls implemented in the web questionnaire. Controls are done in the received data on extreme changes in prices and great influence on the index. The number of questionnaires pinpointed for further investigation is small, usually below 1%.

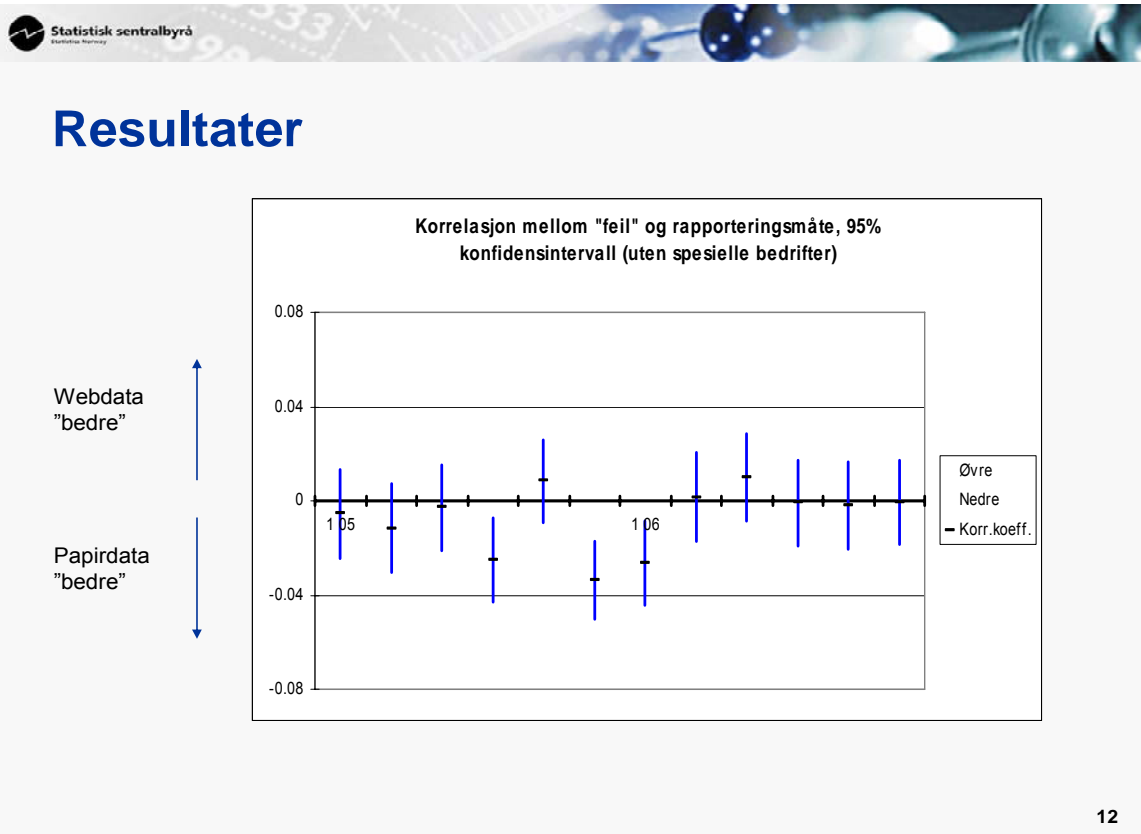
21. Exceeding this level two months in a row, the percents in January/February 2005 are:

Jan/Feb 2005	Within defined range	Outside defined range
Web questionnaire	98.60	1.40
Paper questionnaire	98.73	1.27

22. Looking at the data for two months at a time, and through 2005 and 2006, we find that the data from paper is “better” (“bedre”) in most cases, described through correlation between warnings and instrument.



23. As indicated, this does not lead to more corrections in the web based data. The correlation between errors (“feil”) that lead to corrections, and the data collection instrument shows a more neutral pattern.



24. At first glance, this study confirms that web questionnaires tend to have more wrong or suspicious answers than paper questionnaires. But it also indicates that the explanations can be various. One explanation can be that respondents with correct answers that exceed the limits used in controls seem to use web questionnaire more often than others. So far we have not had the resources to explore this further.

VI. ROAD GOODS TRANSPORT – A TROUBLESOME WEB IMPLEMENTATION

25. To disclose a conclusion, one explanation of increased number of errors in web questionnaires contra paper questionnaires is the early stage of web questionnaires in general, and the SN prioritizing of web questionnaires in all business surveys at the sacrifice of thorough work with each questionnaire.

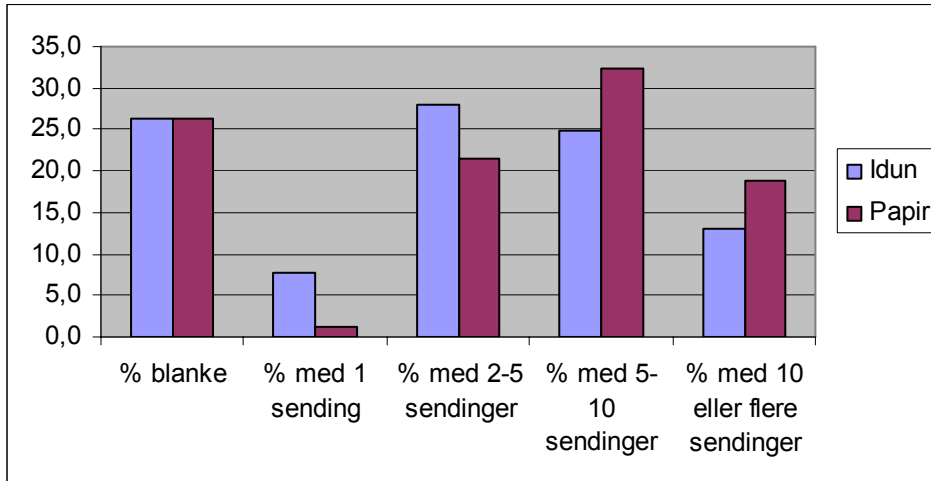
26. Some questionnaires are more difficult to translate into the frames of a web solution. Especially large and varying matrixes give challenges that are not easy to solve. One of these is the questionnaire on road goods transport. In the paper version there is one row for each "sending"², with a number of columns to specify the characteristics for the "sending". In the web version, this is presented as one page for each "sending", with a number of fields to fill in, urging the respondent to ask for another page to register another "sending".

27. In this example we have not used the traditional controls to identify differences, but looking at the data from web and from paper separately gives an interesting pattern³.

² Contract of affreightment

³ We will emphasize that this is under investigation, and must be considered as a most preliminary presentation.

Web (Idun) and paper questionnaires, by number of contracts (“sending”) reported in the questionnaire. Per cent.



28. An explanation can be that the companies having more “sendings” to report more often uses paper, but probably this is not the final answer. (And even if so, what do we do to facilitate a web solution for them?)

VII. SUMMARIZING ON ELECTRONIC QUESTIONNAIRES CONTRA PAPER QUESTIONNAIRES

29. Challenges for the respondent will occur when logging on and finding the right questionnaire, and this may impact the answers in some way. Technical and practical considerations will pose limitations to the developer, making the good questionnaire impossible or very difficult to implement.

30. We have data from a limited number of questionnaires. Superficial glance at a couple of questionnaires not documented here seem to confirm that the pattern is more errors in the initial answers in web questionnaires. But this is changed through controls and warnings presented to the respondent, so that *the answers coming to SN are better from web than from paper in the questions where controls and warnings are implemented*. The answer to this should not be more outsourced controls. We obviously have to work with the electronic questionnaires to improve them to at least the level achieved in paper questionnaires.

31. These brief investigations using the established editing systems as instrument can not reveal quality differences in questions where no adequate controls are (can be) implemented.

32. Differences in respondent characteristics may explain some of the difference, but not all. *As we see it, there are reasons to fear that we face a quality reduction throughout several surveys when changing from paper to electronic questionnaires.*

VIII. COLLECTING ELECTRONIC DATA

33. While web questionnaires can be expressed as “electronic collection of data”, transferring data from the respondents systems to SN is “collecting electronic data”. Establishing these concepts usually involve more participants than SN and the respondent. The software vendor will have a central role, and the use of structured metadata and technical standards are crucial.

34. Starting point for collecting data from enterprise systems is that the data wanted actually occurs in their database, and that there is some basic software for administration of business. This software usually is common for several enterprises. But even with same basic software, there will be local adjustments and variants in using this software. When the software vendor develops additional software

to facilitate reporting for a defined survey, these local differences must be locally handled when using the survey software.

35. Updating the basic software may produce changes in the data(-base), in the next run changing the reporting, if the survey software is not revised. If SN changes the content of a survey, that will require revision of the survey specific software. Each survey is served by a number of basic and survey specific software products. Upgrading of software is dependent on local action in every enterprise. In total, this presents a complex data collection system.

36. Editing systems are built to react when one respondent has misunderstood the question or mistyped when answering. Incorrect data in the enterprise system might occur, and will be exposed in usual way. Vendors are interested in implementing controls in their survey specific software, to avoid sending incorrect data, and through that avoid contact from SN in the editing process.

37. Using the traditional editing (control) systems will not necessarily reveal systematic errors created by bugs in software or lack of upgrading at some of the respondents. Control mechanisms to ensure quality data from enterprise systems should include process controls at software vendors and enterprises. So far, the controls implemented are mainly focused on technical/formal matters, to ensure correct transfer of data. The content definitions (and changes) implemented (or not) are so far not under sufficient surveillance in most surveys.

38. Even with all these challenges, we believe that collection of electronic data will expand.

39. As a brief glance into this, we once again look at data for accommodation statistics. 8% (75/990) of the enterprises sent data from their booking systems. Controlling if less than 10% of the rooms are reported to be let out, 12% of the paper questionnaires are marked. The corresponding number for electronic transferred data was 3%. Average price of a room should be between 300NOK and 1000NOK per night. The control picks out 41% of the paper questionnaires, but only 15% of the electronic transferred data.

40. This is not at all a thorough investigation of these data. The wanted conclusion, though, is that when correct implemented, this way of collecting data should give less errors, in addition to the main goal, to reduce the response burden.

IX. ONGOING WORK

41. After launching electronic questionnaires for all business surveys from 2004, Statistics Norway sees clearly that the quality of these is not at the level we want. This is due to limited resources in the developing stage, but also limitations in the web-solutions, the questionnaire designing tools and (worldwide) lack of knowledge on electronic questionnaires. We are working on improvements, putting more effort into chosen surveys, focusing better questionnaires. Experiences gained from this might be transfused into several questionnaires, and give a fundament for developing editing procedures adjusted to the new data collecting instruments.

References:

Auno, Anne Mari; “*Collecting accommodation data via the Internet – the Norwegian experience*”, presented at the 7. International Forum on Tourism Statistics, 9-11 June 2004, Stockholm.

Gustad, Solveig; “*Analyse av revisjonsmetoder i hotellstatistikken*” (Analyses of the editing system of accommodation statistics), Internal document 2007/10 in Statistics Norway (Only in Norwegian).

Haraldsen, Gustav; Dale, Trine; Dalheim, Elisabeth; Strømme, Halvor: “*Mode effects in a Mail plus Internet Designed Census*”, Presented at International Conference on Improving Surveys, Copenhagen 26-28 august 2002.

Haraldsen, Gustav; Kleven, Øyvin; Stålnacke, Margaretha: “*Paradata indications of problems in Web Surveys*”, Powerpoint presentation to European Conference on Quality in Survey Statistics, 24-26 april 2006, Cardiff South Wales.

Kristiansen, Espen: “*Datakvalitet: Web vs ikke-web*”. (Data quality: Web versus not-web). Powerpoint presentation to an internal seminar in Statistics Norway, 2007. (Only in Norwegian.)