

# Big Scale Observations gathered with the help of Client Side Paradata

Gustav Haraldsen, Øyvind Kleven, Anne Sundvoll

*Statistics Norway*

## 1. Introduction: Bringing it all together

The objective of this paper is to suggest that *client side paradata* from web surveys can serve as a bridge between qualitative methods used during the development of a questionnaire and quantitative quality indicators collected in pilot studies or the actual survey. Our arguments will be based on a discussion of the weak and strong aspects of qualitative and quantitative test methods. We will also include an example from a project where we used cognitive interviewing in the development of the questionnaire, and collected paradata while conducting the actual survey. Finally, we will suggest how other types of paradata can be tailored in order to coincide with small scale observations made during the development of a questionnaire.

## 2. Building the bridge

Qualitative testing methods consist of several techniques used to collect information about how the respondent interprets the survey question, collects relevant information and arrives at an answer. The overall strength of qualitative testing methods is that they collect a variety of information about how different test persons think and respond to survey questions.

If the interviewer is well taught and the test person is properly instructed, the overall experience is that think-aloud techniques, accompanied by verbal probes predicts highly valuable information about the process of *question comprehension, information retrieval, judgement and estimation, and response* (Tourangeau, 1984).

One problem associated with cognitive interviews is that the interviewer behaviour might affect what the participants say and what conclusions are drawn (Beatty, 2004). To ensure the satisfactory degree of data quality, the interviewer should be familiar with the “current best practises” (e.g. Snijkers, 2002) to avoid undesirable interviewing practices. The most important problem, however, is that the tests are carried out in small scale. Hence, we do not know if the problems we detect also will apply in full scale. In a survey based on statistical principles, the results can very well be of high quality, even if not all the questions work perfect for all respondents. In fact, this is that may be called the “magic” of statistical surveys. Hence, one of the questions we often struggle with during the development of new surveys, is to distinguish between cognitive problems that should not be ignored and problems that can be overlooked.

What is generally recommended in order to solve this dilemma, is to combine qualitative development and testing methods with quality assessments in representative pilot studies. Generally, however, there is often no money nor time to carry out such a triangulation of planning methods. We think there is another problem with this method as well: The “distance” between what is learned about cognitive problems in the development of a questionnaire and the quality problems detected in the actual survey is so long that it is difficult to establish a solid link between these two observations. This is due to the fact that while cognitive testing focuses on the process quality, quality evaluations of the actual survey focus on the result quality. If test respondents had no problems with the terms and tasks of a survey question, and the answers to

this same question seemed to be of high quality, we argue that this is because this question was easy to answer. If there were problems detected in the cognitive testing and the final quality also seems to be low, we take this as a result of the cognitive problems we have described. We tend, however, to leave out all of these incidents where the quality evaluation of questions does not coincide with qualitative test results. And even when they coincide, the relationship may be questionable.

Traditionally there are three ways of identifying quality problems in surveys. What is considered to be the best method is to compare results from the questionnaire with other sources to the same kind of information. The problem is of course that such a reliable source of information generally does not exist, and that was just why the survey was conducted. A more common method is to use missing units or missing answers as an indicator of response problems. The third, and probably most cost efficient method, is to look for inconsistencies in the response patterns.

None of these indicators yield much insight into the response process that causes the problems, and it might not be problems observed in cognitive testing. Unit nonresponse may be caused by cognitive problems associated with the questions, but can just as likely be reactions to the topic of the survey or be caused by practical problems that hinder those who are addressed to participate. From socio-psychological investigations (see Krosnick 1991, Krosnick and Fabrigar 1997), we know that respondents have a tendency to make a qualified guess rather than leaving a difficult question unanswered. Hence, item nonresponse might also be a poor indicator of cognitive problems. And logical inconsistencies between answers are often not considered to be a cognitive problem for the respondent. Thus, our general argument is that as long as you do not collect process data in the survey, the link between quality indicators and cognitive problems detected in qualitative tests are rather weak. It is this logical gap that we think client side paradata can bridge.

Paradata are data collected in web surveys that describe *how* respondents filled in survey questions, in contrast to *what* they fill in. We distinguish between server side paradata (SSP) which are information concerning *page requests* on the web server, and client side paradata (CSP), which are information about what is going on *within* a web page. Client side paradata describes, with high-precision timestamps, the actions of a respondent, such as clicking response alternatives, changing answers, clicking hyperlinks, scrolling the page, moving the mouse pointer, and interrupting a task (Heerwegh, 2003). These data are collected with the help of a java scripted program.

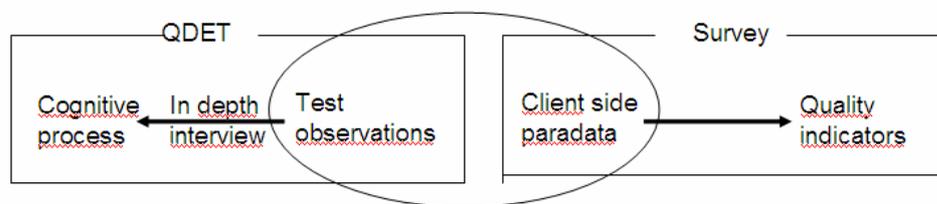
The procedure we generally follow when we carry out cognitive interviews in Statistics Norway is the following:

1. Formalities
2. Warming up for think-aloud session
3. Think-aloud session. The moderator tries to note which questions that cause problems, and what type of problem, but tries to interfere as little as possible
4. Follow up questions based on the observations made during the think-aloud session
5. Planned questions and exercises to test more thoroughly specific questions and problems, selected before the test.
6. Short break while the moderator sums up what he thinks are the main results
7. The moderator presents his summary and asks the test person to clarify, and make additional comments.

We have described this procedure in order to view what we call *cognitive interviewing* in fact is a mixture of observations and in-depth interviewing. With paradata we do not collect the kind of information that we gather in the in-depth interviewing, but data that are similar to the observations made during a cognitive interviewing session. In a previous presentation made to the Quest group in 2003, we demonstrated how programs like Camtasia can be used to record what a test person is doing within a web page in individual tests (Brekke, 2003).

This kind of behavioural data can be collected in a full-scale setting with the help of client side paradata.

In the qualitative tests described above, the follow up questions are used to link observations made during the think-aloud session to the four-fold cognitive typology developed by Tourangeau. In a web survey, paradata can be linked to result quality indicators. The in-depth interview is not repeated in the survey, and the quality evaluation of the final results can not be carried out during the test period. But if we can use client side paradata to identify similar observations to those we experienced during the test, these data can be used to establish a stronger link between the small scale investigations of cognitive process and the big scale evaluation of survey quality. The conceptual model we apply can be drawn like this:



**Figure 1: A conceptual model that links together small scale investigations of cognitive processes with big scale evaluations of survey quality**

In the following example we will try to establish such a link between test results and client side paradata gathered during the development of a customer satisfaction survey, carried out for Statistics Norway in January 2005.

### 3. Qualitative pre-testing of The Statistics Norway's Customer Survey

In the fall of 2004 we carried out a series of cognitive interviews in the development of the customer survey questionnaire. The cognitive pre-testing of the draft questionnaire was run in different stages of the questionnaire development. The test method was based on concurrent think-aloud, accompanied by targeted follow-up probes. The draft questionnaire contained a mixture of behavioural- and attitude questions. In consideration to establish the optimal flow, the questionnaire as a whole was tested. Most of the tests were carried out by using the paper version of the questionnaire, and one test was carried out on the web version.

The moderator instructed and guided the test respondents. The secretary concentrated on observation and producing the report. All the tests were videotaped.

As a conceptual background we used the Forsyth's Questionnaire Appraisal Coding System to map and evaluate the responses (Forsyth et al. 1992). The appraisal system consists of a set of codes that describe question features likely to contribute to response error. The codes are divided into four sections, corresponding to the Tourangeau response model. We use a version of this system where the potential problems relevant in household and organizational questionnaires are

written into the same coding form. In this way it contains both terms that refer to problems the respondents can have as an *individual* and as an *employee* in an organization. This version of Forsyth's coding system is shown in appendix 3. (The terms that only apply to organizational surveys are written in italics). In the following we will use this list of terms in our presentation of findings from the questionnaire testing.

### **3.1. Problems associated with question comprehension**

The customer survey links customers' usage of the Statistics Norway's products/services, and level of satisfaction. The targeted attitude questions represent different indicators of customer satisfaction. The survey administration wanted to develop a general measure of customer satisfaction - a "temperature meter". Hence, several questions were picked from already established European customer surveys, without any special adaptation to the Norwegian society. The pre-testing process indicated quite clearly that the test respondents found it difficult to relate to a set of "general labels". The conclusion drawn from the testing was that some of the questions ideally should be more properly tailored and adjusted, to reflect the customers' relation to the agency in everyday life. Still, a few of these "external" questions were decided kept as they were, without any further change of the wording.

Overall, the test respondents did not express severe difficulties in comprehending the targeted survey questions. However, observations of the test respondents and targeted probes, followed-up by elaboration fostered several examples of potential sources of response error. As a result, several of the initial survey questions were either left out or rephrased.

The comprehension problems identified in the testing process, were mostly due to *ambiguous or vague terms*. One example to address this problem was that the respondents were not familiar with the Statistics Norway's departmental structure - used as response alternatives in the draft questionnaire. The test respondents reported that they had no relation to the suggested classification. Hence, the question was rejected in the revised version of the questionnaire.

By using targeted probes to investigate the survey concepts more carefully, we detected several examples of vague concepts, for instance "information". The cognitive test process indicated more effort should be put to tailor the concepts to the customers' situation, in order to make the questionnaire more applicable. As a result, the questionnaire was extensively revised after the initial tests.

### **3.2. Problems associated with information retrieval from memory**

While question comprehension problems easily are both detected and corrected for by using Qdet test methods, the range of problems associated with *information retrieval* and *judgement*, can be more of a latent and problematic kind. In business surveys we have learned that unsolved retrieval problems easily contribute to the increase of response burden.

By observing the test persons, we saw that some of them were likely to smile indulgently from time to time, and that they easily "blamed it" on the poor wording of the survey questions. However, it soon became clear that the respondents actually struggled with the *task requirements*, posed by the questionnaire. Accordingly, the smiling was attributable to the perceived massiveness of the questionnaire - appearing a bit bureaucratic. In many cases, the test respondents had no relevant information to base their answer on. One example is the assessment of the price level of the products and services bought from Statistics Norway. The test persons reported that even if they had fairly good knowledge about the products, they felt quite a distance

to questions about the amount of money they had spent on the product. In most cases the organization had applied for money for research assignments in beforehand. In this application the price for the product was already accounted for. Hence, the test persons said they were in lack of relevant information about this business, and did not have a relevant basis to make the judgement.

The cognitive testing also shed light on *recall issues*. The test questionnaire was originally designed in a way that the questions about the last contact with Statistics Norway appeared among the first introduction questions. The test persons reported difficulties that easily relate to an *information retrieval problem*. According to the second step of the Tourangeau cognitive model, it somewhat came more easily to the respondents' mind to answer questions about "regular activity" (that is customer relations over a limited time period), than remembering a specific occurrence ("last contact made with Statistics Norway"). Hence, an important outcome of the qualitative testing was the need to add a few questions about regular activity to "warm up" the respondents, before narrowing the reference period to the last contact made with the agency.

### **3.3. Problems associated with judgement and estimation**

In those cases where the test person reported lack of relevant information, the cognitive step of judgement and evaluation was also "challenged". The test persons with a weak relation to the price level had difficulties in making a judgement and formatting an answer. On the other hand, their basis to judge the components of "service" and "professional skills" seemed surprising untroubled.

Some of the test respondents reported that they found it somewhat "annoying" to go through a set of questions that did not feel applicable to their situation as customers. A reason for that might be that the vast majority of the initial test questions were concerning "information products" (e.g. books, publications). In reality, a certain share of the customers is in frequent contact with Statistics Norway in errands of register samples, surveys or customer adapted statistical analysis.

One of the test persons even suggested that in a real survey situation, one might observe a growing tendency of poorly considered responses throughout the questionnaire, due to a motivation drop. We have observed in business survey focus group studies that respondents who feel "mistreated" or somewhat "insulted" by a poorly adapted questionnaire, have a tendency to seek for short cuts - to take guess instead of making the exact calculation, and through this kind of behaviour "pay back" with poor data quality. Hence, it's important to prevent shortcuts, and be alert of possible big scale consequences for the data quality.

### **3.4. Problems associated with formation of the response**

An overall problem exposed by the cognitive testing, was the difficulty in tailoring the response alternatives to fit different types of customers. The process of testing revealed that some of the multiple responses were not properly adapted, and hence the respondents remarked that their "instinctively generated response" was not yet included among the listed responses. By observing the test persons' response pattern, we detected several times that the respondent made a quick glance through the list, and then paused and searched for an adequate response alternative. Sometimes he decided to tick for a predefined response, even though it didn't match the cognitive formatted response. One example to illustrate this is the following: The test respondents who had bought a register sample, were likely to tick for response alternatives including the word "survey". Obviously, there was a need to distinguish "register sample" as a separate response alternative.

As a direct result of the observed response pattern, open responses ("Other, please specify") were

kept in the final questionnaire to make the predefined list of responses appear as exhaustive to the respondents. By choosing such a strategy, you're apparently off with one problem, but you still might end up adding more response burden to the question-and-answer process: Clearly, open responses is time-consuming to the respondent.

#### **4. Paradata observations and response patterns in the Customer Survey**

1250 customers were asked to respond to the customer survey. All respondents received a paper questionnaire, but were also invited to respond using a named Internet address. The overall response rate was 61 %. Of these respondents 47 % filled in the paper questionnaire and 53 % used the web option. Since client side paradata only can be collected in web surveys, it is only in this part of the survey that we are able to link paradata observations to response patterns. Even if the survey design of the paper and web version was very similar, conclusions drawn here may be effects of the web design rather than effects of the questions asked. One most unfortunate effect was that the respondents were obliged to answer each question in the web version, while they were of course free to skip paper questions (see Appendix 1 and 2 for details). Consequently item nonresponse can not be used as a quality indicator in the web questionnaire.

The paradata that were to be collected and how they were to be processed, were determined by the selection and procedures offered by Dirk Heerwegh on his open website <http://perswww.kuleuven.ac.be/~u0034437/public/csp.htm>. From the list of client side paradata offered by Heerwegh we have chosen to present the time it took to fill in the response to each question and the number of response changes made for each individual question. That the respondent spends some extra time on a question or changes answers he has already given are common indicators of problems also in qualitative tests during the development of questionnaires. Hence, these two kinds of client side paradata coincide with observations likely to produce follow up questions in cognitive interviewing. In figure 2 the completion time for each question is given in seconds. In figure 3 the response change rate are given as the percentage of respondents who made changes. In both cases the patterns during the response session are provided in line graphs. The questionnaire was eight pages long and consisted of 40 main questions. Some of these questions were divided into several sub questions. All in all there were 116 response boxes in the questionnaire. As one can see along the x-axis of the graphs, the numbering starts on 1 and ends on 39. The final question is also there, but the label is not shown. Of those questions with a label, there is a mixture of single questions (like 16 and 19) and sub questions (like 21\_1 and 21\_7).

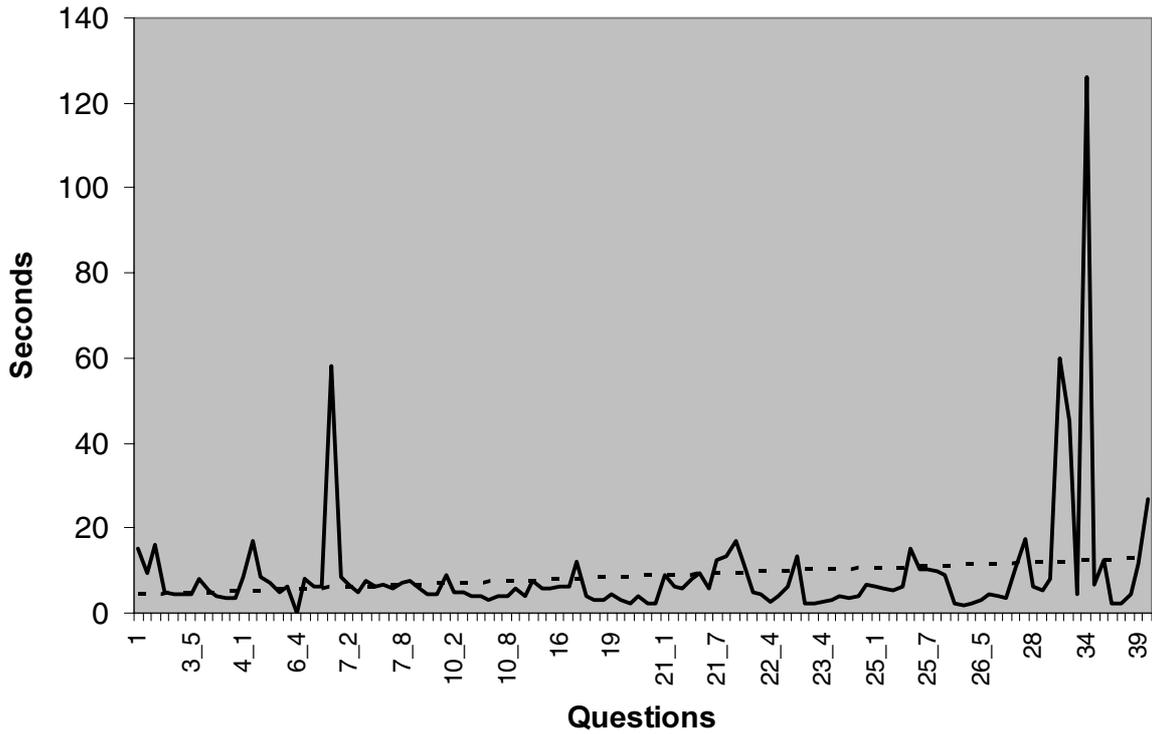


Figure 2: Question Completion Time given in seconds. Trend line (.....)

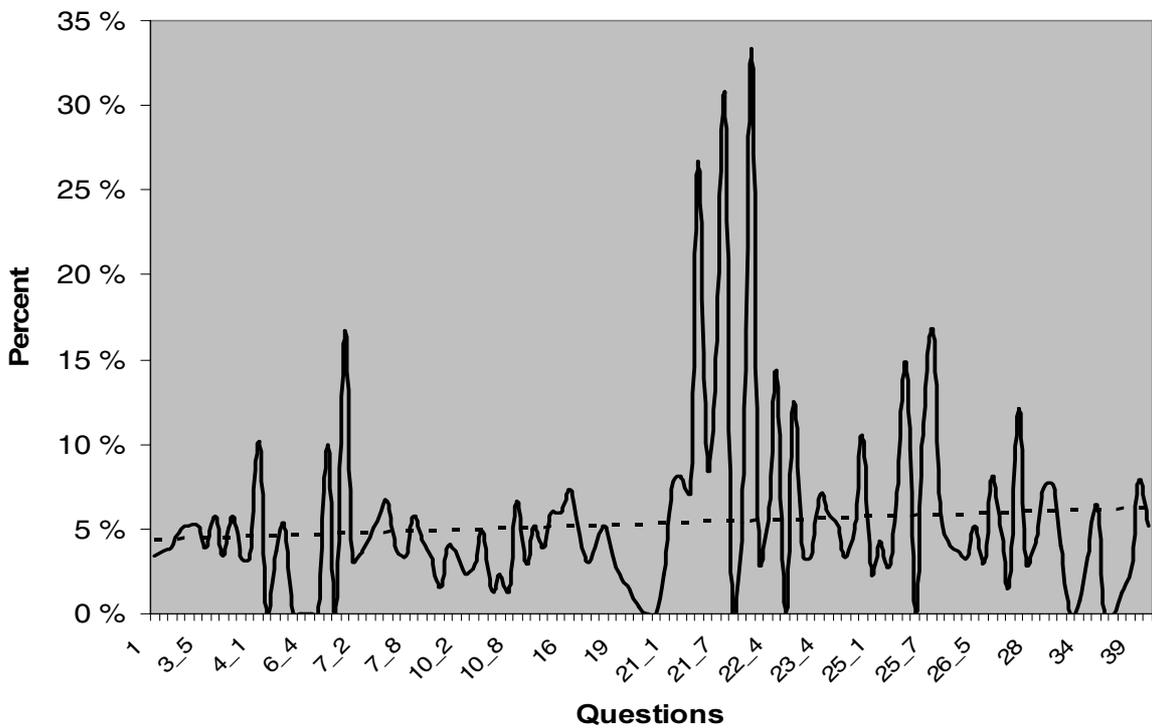


Figure 3: Question Change Rate. Percent who changed their answer. Trend line (.....)

#### 4.1. The general picture of the response process

The average response time for all questions were 15,3 seconds when outliers were included and 8,9 seconds when outliers were excluded. In the graph the outliers are excluded. It took a little bit below 30 minutes for an average respondent to fill in the questionnaire. Our general reactions to these figures are that the questionnaire was quite quickly completed. When one evaluates this, however, one should bear in mind that the questionnaire was divided into sections about different kinds of products and services. For some of these, very few respondents had first hand experience, and were consequently told to skip the evaluation questions. This is for instance true about the seven questions asked to those who had used the statistical database available on the website of Statistics Norway. 1/3 of the web respondents skipped those questions.

One question quite early in the questionnaire (q6\_8) and a few questions in the end (q31 and q34) took noticeably longer time to fill in than the other questions. All these questions are open questions. In general, it is not surprising that open questions take longer time to fill in than questions with fixed response alternatives. Question 6\_8, however, is one out of several examples of questions with an open option for those who did not find the previous fixed alternatives covered their opinion. In question 6, respondents who missed the old paper publications from Statistics Norway were told to specify why they did so. They could chose several reasons given in the first seven response alternatives, or specify other reasons in the last one. For other questions with this structure, it evidently took less time to specify a different opinion from those given in the fixed alternatives. We interpret this as an indication of that the fixed alternatives given did not match very well with the reasons the respondents wanted to communicate.

Between question 31 and 34, a third open question 32 was presented. In question 31, respondents who envisaged that they would use more of our statistics in the future, were asked to specify why. In question 32, those who rather thought that their use of statistics would decline, were given the same kind of follow up question. Hence, those who had an optimistic view of their future use of statistics, also seem to be able or willing to give more reasons for this, than those who were more pessimistic.

The first question in the questionnaire was this: "If you consider all aspects of Statistics Norway, both the products they offer for free and sell, all in all how satisfied or dissatisfied are you with Statistics Norway?" This very general question was posed first because one did not want previous answers to influence on what was meant to be an overall evaluation. On the other hand, the question asks for an overall and difficult evaluation. Only 2 percent of the respondents answered that they didn't know. On the other hand most respondents seemed to spend relatively long time (more than 15 seconds) on answering it.

Another observation that is worth mentioning is that there are several cases where the respondents spend considerably longer time in evaluating the first item in lists of aspects that all should be evaluated along the same scale than the time they spend with the following items on the list. One obvious reason for this is that a timestamp is recorded for every action. And the last action before the evaluation of the first item on a list is recording the last answer on the previous question. During this time, the respondent both needs to read and understand the following question, the first item and the scale along which the items should be evaluated. In fact, the interesting thing is perhaps not that this takes longer time than evaluating the items further down the list, but that the difference is so small. This may indicate that the respondents often read the questions rather superficially. Also it is well known from previous research that the first item on such lists tends to

form a yardstick for the following answers. Therefore, the order of the list should be randomly changed from respondent to respondent. The extra time spent with the first item on the list, probably also reflect that the respondent both consider the question and the response scale.

The first impression from figure 3 is that quite a lot of respondents changed quite a lot of their answers. In some cases, however, even a few changers in an already small group that has been selected by a previous filter question, constitute a high percentage. This is for instance true for question 6, that was only answered by 18 respondents who claimed that they missed the paper publications. Examples of questions that both had many respondents and a high proportion of changers were question 4 (“When you gather information from Statistics Norway, do you most often use our web service or our paper publications?”), question 23\_8 (“When you think about last time you were in contact with Statistics Norway, were you satisfied or dissatisfied with the way the office responded to complaints”) and question 22 (“What kind of product or service did you base your previous customer evaluation on?”). If there is a common denominator for these questions, it is that they ask for judgements.

The average number of changes made by those who changed answers was 1.7, which indicates that most often one and in some cases two changes were made.

In both graphs we have also drawn a trend line. None of these lines indicates that the questionnaire was so long or complicated that the willingness to spend time on the questions changed substantially as the completion went on.

#### **4.2. Linking observations from specific questions to response patterns**

In this part of the analysis we have tried to follow up some of the problem questions from the qualitative testing and observed how they worked in the actual survey. Our approach is to look at the relationship between the client side paradata and the response pattern for these questions. First we have picked examples of questions with a list of evaluation items that should be considered according to a five-point scale that goes from Very satisfied to Very dissatisfied. The first of these is question 26, which was worded like this: “Thinking about last time you bought something from Statistics Norway, were you satisfied or dissatisfied with the following:” Table 1 shows the response distribution, the average response time and the proportion of changers for the different evaluation aspects in the list that followed:

**Table 1. Q26. Thinking about last time you bought something from Statistics Norway, were you satisfied or dissatisfied with the following? Give one answer on each line.**

	Very satisfied	Satisfied	Neither satisfied nor dissatisfied	Dis-satisfied	Very dissatisfied	Don't know	Not applicable	Average response time with and without outliers		Percent who changed their answers
The time it took to get in contact with the right person	42	36	8	4	0	2	8	12	9.1	4.0 %
The service given	<b>45</b>	<b>41</b>	7	1	0	2	3	<b>2.5</b>	<b>2.1</b>	<b>3.7 %</b>
Professional competence	<b>42</b>	<b>38</b>	10	1	0	5	4	<b>2.1</b>	<b>1.7</b>	<b>3.3 %</b>
Statistics Norway's ability to keep deadlines	41	34	12	5	1	1	5	2.6	2.2	5.2 %
The time it took from first contact to delivery	37	33	14	5	3	1	7	3.4	3.0	3.0 %
The price you had to pay	12	31	<b>36</b>	<b>10</b>	<b>5</b>	3	2	<b>7.0</b>	<b>4.4</b>	<b>8.1 %</b>
The product compared with expectations	35	45	13	3	0	2	2	4.3	3.8	4.8 %
Internal coordination of the service given	23	27	13	5	1	16	16	3.8	3.4	1.8 %

#### 4.2.1. The time it takes to read the questions

Question 26 is one of several examples where the respondent spent more time on the wording of the question and the first item than he spent on the next item. If we assumed that it actually did not take significantly more time to give the first evaluation than the second, the figures indicate that it typically took seven seconds (9.1 – 2.1) to read the question and familiarize with the response scale.

Question 26 is the last of five questions using this kind of scale. Figure 3 shows the difference in completion time for item 1 and item 2 in these five questions. The differences vary from 11.3 to 4.6 seconds. There is no trend indicating that the respondents spent less and less time on the questions as they became more familiar with the design or more fed up with the questionnaire. The results rather indicate that it is how difficult the wording or task is that decides the how much longer time it takes to fill in the first line compared with the next. If this is true, the introduction to question 23 was the most difficult, and the introduction to question 10 the easiest to comprehend and respond to. The wording of the questions and first two items in these five questions, ordered according to their apparent difficulty, is given below the graph.

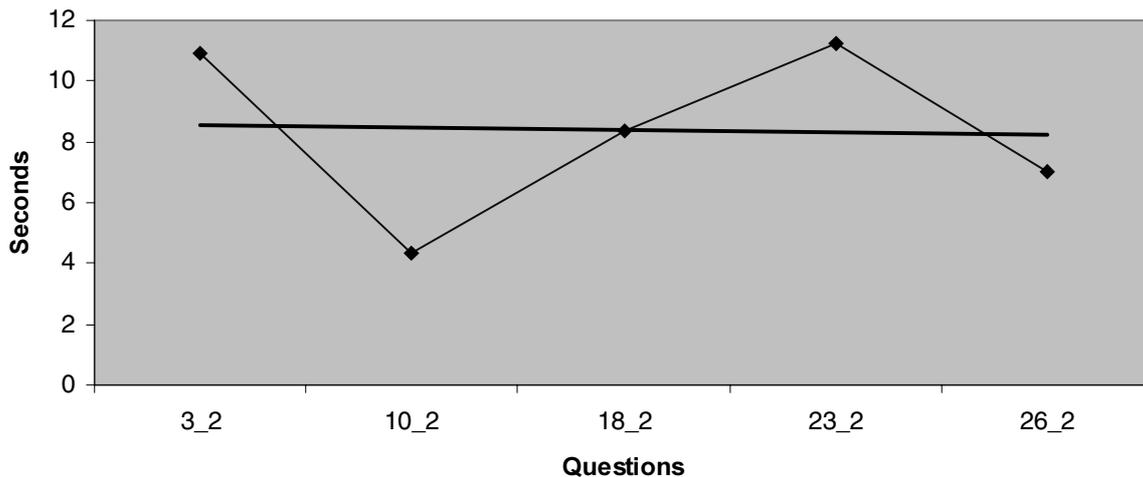


Figure 3: Time difference in completing the first and second item on similar evaluation questions.

**Q23. If you think about the last time you were in contact with Statistics Norway, how satisfied or dissatisfied were you with the following?** Give one answer on each line.

The time it took before you had a response.....  
 The service.....

**Q3. Overall, how satisfied or dissatisfied are you with the following regarding the statistics provided by Statistics Norway?** If you don't have a basis to answer some of the questions below, please tick the Not applicable box. Put only one tick on each line.

How the statistics fulfill your requirements.....  
 How up to date the statistics are .....

**Q18. How satisfied or dissatisfied are you with the statistical database of Statistics Norway?** Give one answer on each line.

The documentation given in "About the statistics".....  
 The number of statistics in the database.....

**Q26. Thinking about last time you bought something from Statistics Norway, where you satisfied or dissatisfied with the following?** Give one answer on each line.

The time it took to get in contact with the right person.....  
 The service given.....

**Q10. How satisfied or dissatisfied are you with the web pages of Statistics Norway evaluating the following aspects?**

Give one answer on each line.  
 How easy it is to find what you want.....  
 How you move from the home page .....

We do not want to speculate too much about what caused the differences in how long time it took to complete the first part of these questions. It seems, however, rather likely that the reason why it took so little time to read and start answering question 10 was that, because they are using the web version of the questionnaire, the respondents we are surveying obviously have first hand experience with our web site. Looking at the other questions from this perspective, it might also be true that those questions that apparently were not so easy to read and start answering, are questions with unclear reference points. This is clearly true for question 3. It is also interesting to notice that it took longer time to start on question 26 than on question 23. These two questions are almost identical, but with two exceptions. The first exception is that question 23 asks about things you have bought, which might be easier to remember than other kinds of contacts. The other difference is that question 23 was posed before question 26. During the cognitive testing we had

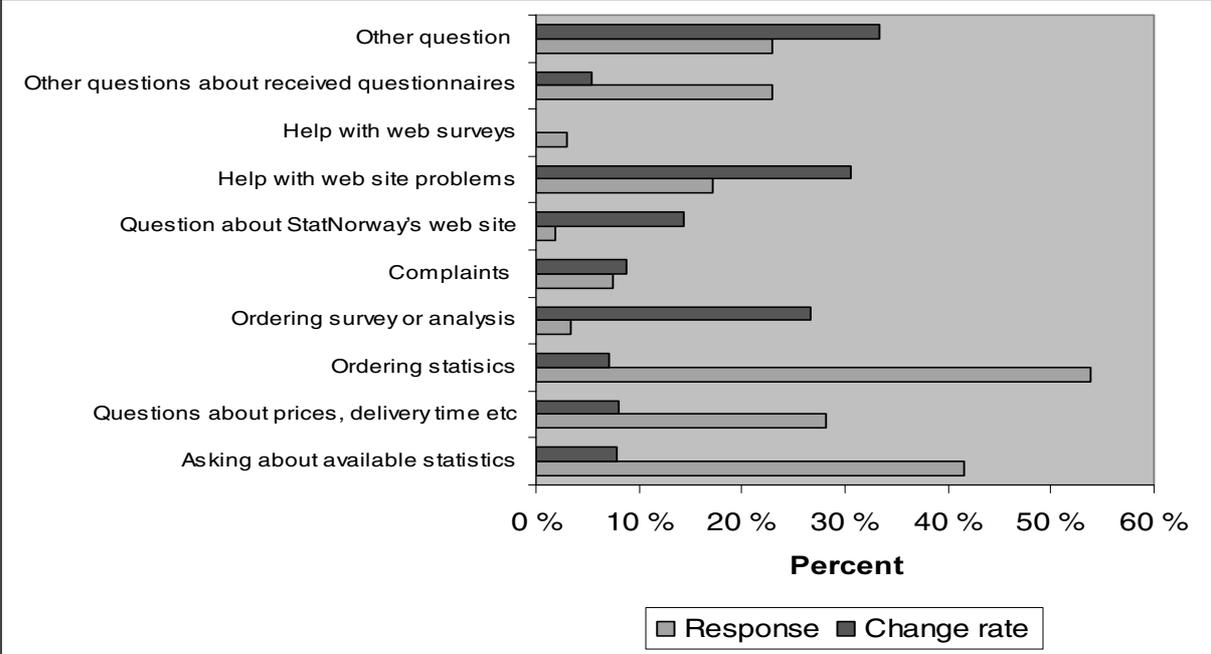
the impression that question 23 was used to set the reference point, and that it was the same reference point that was normally used in question 26 as well.

**4.2.2. The time it takes to be negative or neutral**

What strikes us next in the question 26 table and in many of the other answers given in the questionnaire, is that the respondents generally express high satisfaction with most of the contacts they have had with Statistics Norway and with the products and services that Statistics Norway has delivered. The response pattern described in the table above is typical in this respect. These positive attitudes were also recognized in the cognitive interviewing during the construction of the questionnaire. At that time we were even a little bit anxious that the test person felt obliged to be positive because the tests were run in Statistics Norway. But it rather seems that the statistical institution has an overall positive image among customers. The question which took longer time to answer among the items listed in question 26, was the question about prices, which was the one where the statistical office had the most negative score. In other questions, for instance in question 3, we also noticed that questions with a high proportion of respondents who chose the Neither-nor alternative often had a higher completion time and change rate than other questions. This leads us towards the conclusion that a majority of the questions were quick to answer because they were a repetitive statement of the general confidence people have to Statistics Norway. What took time were those few questions about aspects which the customers had experienced deviated from their general expectation.

**4.2.3. The burden of multiple choice**

Four questions in the questionnaire had multiple choice. Three of these questions stand out with high change rates in figure 3\*. Most changes were made in question 21 which asked for what kind of questions the respondents had posed the last time they had contacted Statistics Norway. Nine fixed suggestions and one open option were presented in this question. In figure 4 the response and change rates for these alternatives are shown as bar graphs.



**Figure 4: Response and change rates in the multiple choice question “What was the topic of your last contact” (with Statistics Norway). Percent.**

\* The fourth was very simple. It asked respondents who had been in contact with Statistics Norway, how the contact was made (by telephone, e-mail, fax, letters or by personal contact).

Multi choice questions like this one can either be read line by line and answered with a tick for “yes” or nothing for “no” or the list of response alternatives can be read before the respondent makes his choice. The first way of reading is the same way that the respondents are asked to read items that should be evaluated according to a common scale, as in question 26. We think, however, that the high change rate indicates that the response alternatives rather are considered vertically. It took about half a minute to evaluate the eight items listed in question 26. Even if it is a little bit more difficult to calculate the time it took to complete question 21, if we ignore the open option, it apparently took three times as long to choose among the nine fixed alternatives given\*.

Another interesting observation to be drawn from the bar chart is that in several cases the change rate is much higher than the proportion of respondents who eventually landed on a certain response alternative. This is particularly true for the suggestion that the last contact could be about survey or analytical commissions. Even if very few respondents eventually chose this alternative, quite a lot of respondents seemed to have considered it. We would guess that many respondents originally missed that it was a question of actually ordering a service and, when they discovered this, changed to a different option.

Also for the two questions suggested about Statistic Norway’s web site, the change rate was higher than the response rate. In this case, one can imagine that quite a lot of respondents found it hard to decide if they only had posed questions about Statistic Norway’s web site or if they had actually brought up a technical problem. The difference between these two options may be difficult to draw.

## **5. Bringing it a step further?**

The client side paradata presented in this paper are offered by Dirk Heerwegh’s web site, and were collected before the theoretical model in figure 1 was developed. Hence, the next steps we want to take in order to investigate how the understanding of cognitive processes can be linked to survey quality indicators, is to carry out a *controlled experiment*. The purpose of qdet methods is to improve the questionnaires before they are used in surveys. For our purpose, however, we would prefer to keep some of the original questions that normally would have been ruled out or improved before they are implemented in the actual survey. Besides this, we think that the weakest part of our analysis was that we did not have good quality indicators to distinguish between questions that gave valid and reliable results and those which did not. Incidentally it is a general problem to come up with good quality indicators in surveys. But in web surveys we even think paradata can be used for this purpose. If the survey is well planned, quality checks can be built into the survey and run in the background as the respondents fill in the answers. The number of quality checks that detect errors can be counted and described with the help of client side paradata. A quality of the questionnaire can be defined as the relation between possible and activated error checks, and the quality of individual questions can be defined as those who do not activate error messages.

Even if we have discovered that client side paradata may be difficult to interpret, and that we have not been able to link the different aspects of our model together yet, we think some of the results are interesting and promising. In the summary at the last QUEST workshop in Mannheim, it was

---

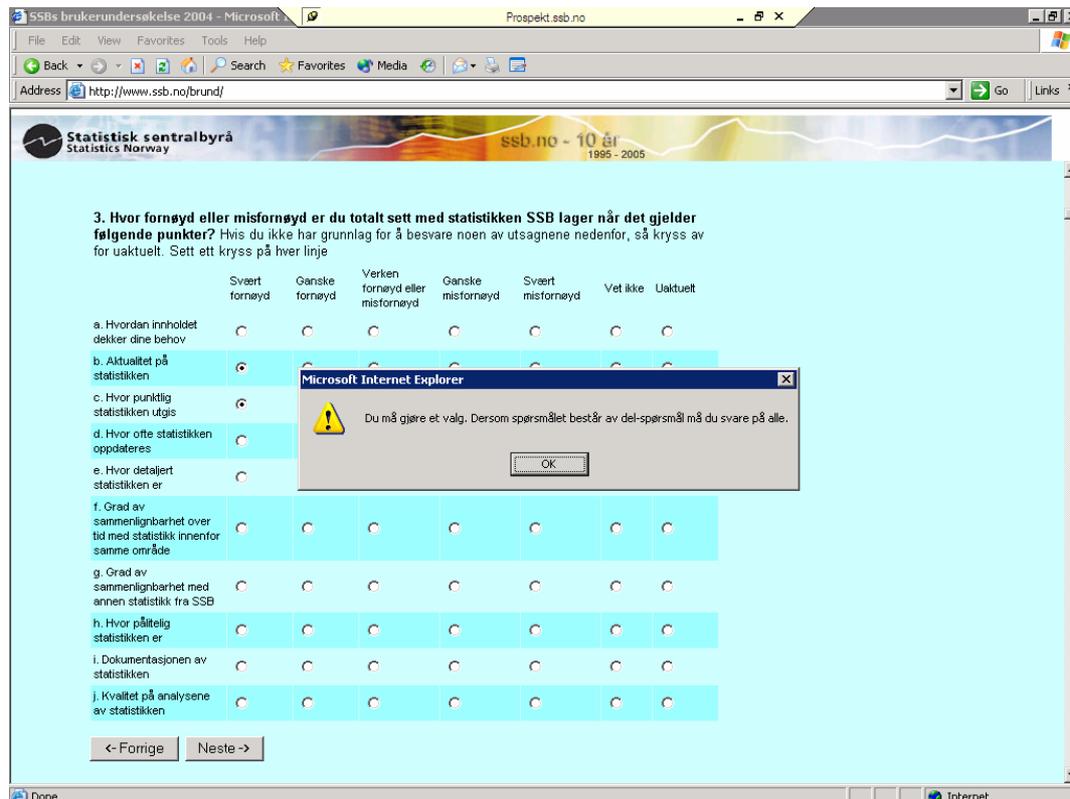
\* 1,46 minutes when outliers are excluded and 2.18 minutes when outliers are included.

pointed out that the methods used cognitive interviewing and other tests of questionnaires seem to detect more comprehension and recall problems than judgement problems and problems with finding an appropriate response category. Client side paradata often seem to point at the same questions and problems revealed in qualitative tests. But in addition to this, we also have a feeling that this kind of observational data perhaps helps us to discover judgement- and response problems that tend to be overlooked in the questionnaire development.

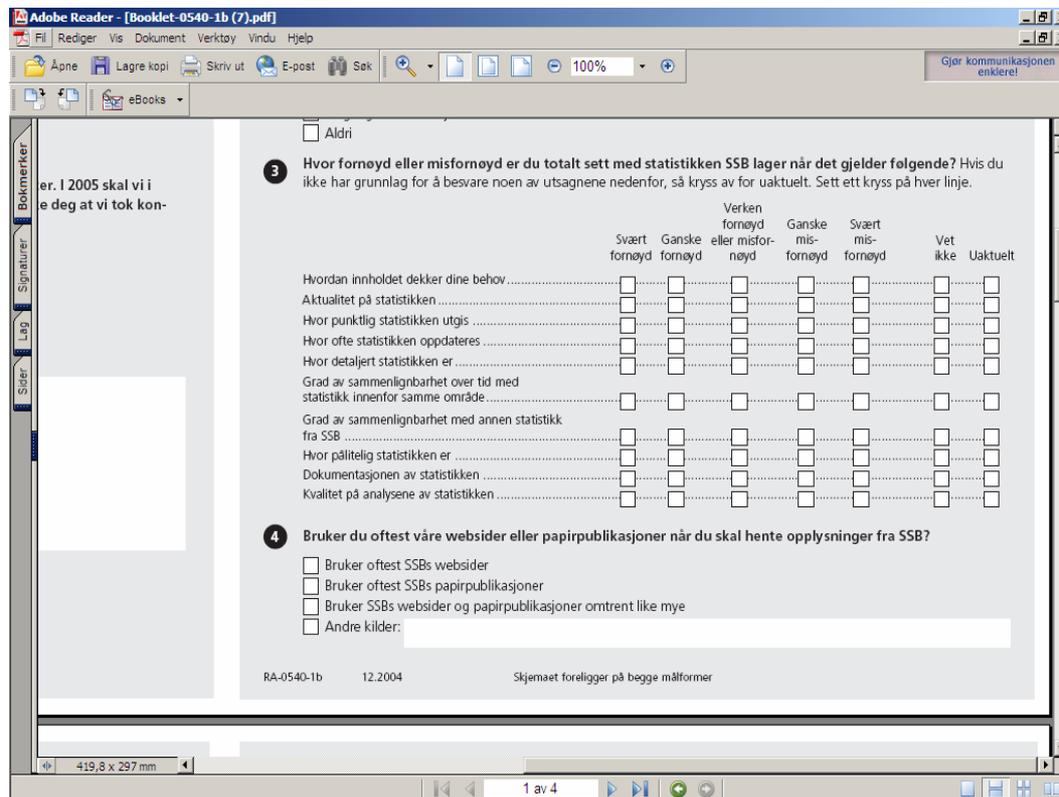
## References

- Beatty, P. (2004) "The Dynamics of Cognitive Interviewing" In S. Presser, J.M. Rothgeb, M. P. Couper, J.T. Lessler, E. Martin, J. Martin & E.Singer (eds.) *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley
- Brekke, Ø (2003) *Using the Computer as Recording Device in CASI-Survey Cognitive Testing*. Presentation at Proceedings of the 4<sup>th</sup> Conference on Questionnaire Evaluation Standards. Mannheim, 21 – 23 October 2003.
- Forsyth, B. , Lessler, J., & Hubbard, M., (1992) "Cognitive evaluation of the questionnaire" In C. Tanur, J. Lessler, and J. Gfroerer (eds.), *Survey Measurement of Drug Use: Methodological Studies*. Rockville, MD: National Institute on Drug Abuse, pp. 12-52
- Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client Side Paradata From a Web Survey. *Social Science Computer Review*, 21, 360-373.
- Krosnick, J. (1991) "Response strategies for coping with the cognitive demands of attitude measures in survey" In *Applied Cognitive Psychology* 5: 213-236.
- Krosnick, J. and Fabrigar, L. (1997) "Designing rating scales for effective measurement in surveys" In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin (eds.) *Survey Measurement and Process Quality*. New York: Wiley
- Snijkers, G. (2002) *Cognitive laboratory experiences: on pretesting, computerised questionnaires and data quality*. Ph.D. dissertation, University of Utrecht.
- Tourangeau, R. (1984) "Cognitive science and survey methods: A cognitive perspective" In T. Jabine, M. Straf, J.Tanur & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press

## Appendix 1: It is not possible for the respondent to skip a question in the web-version:



## Appendix 2: In the paper-version, it is not possible to prohibit the respondent to skip a question:



Appendix 3: Combined version of Forsyths Questionnaire Review Coding Systems for household and organizational surveys.

COMPREHENSION		INFORMATION RETRIEVAL		JUDGMENT		RESPONSE SELECTION
Task Instructions	Question Content	Organization Characteristics	Match: Record and Item	Response Terminology		
- Conflicting instructions	- Complex topic	- <i>Distributed knowledge (multiple sources)</i>	- <i>Mismatch – item and regulatory requirements</i>	- Critical definition(s) missing		
- Inaccurate instructions (Hidden instructions)	- Under-specified topic	- <i>Seasonal or periodic trends</i>	- <i>Mismatch – item and organizational objectives</i>	- Vague term(s)		
- Complicated instructions / Complex syntax	- Topic carried over	- <b>Source identification</b>	- <i>Mismatch – item and variability in record units</i>	- <i>Mismatch to technical language</i>		
- Separate from item / nearby but not embedded in item	- Assumes consistent behaviour	- <i>Assistance to identify sources not provided</i>	- <i>Mismatch – item and system time frames</i>	- <i>Industry-specific terminology</i>		
- Instructions provided too late	- <b>Question Terminology</b>	- <i>Assistance to identify sources not provided</i>				
- Unclear examples	- Critical definition(s) missing	- <i>Source(s) may not be accessible</i>				
- <i>Unclear layout</i>	- Incomplete examples					
- Transition needed	- Ambiguous or vague term(s)					
- <i>Assistance to identify sources not provided</i>	- Multiple definitions					
	- <i>Mismatch to technical language</i>					
	- <i>Industry-specific terminology</i>					
<b>Navigational Instructions</b>	<b>Question Structure</b>	<b>Memory Retrieval</b>	<b>Judgment Process</b>	<b>Response Units</b>		
- Inaccurate instructions (move to wrong place)	- Hidden questions	- Non-routine summary or breakdown required	- <i>Coordination or collaboration necessary</i>	- <i>Mismatch – item and organization units</i>		
- Confusing convention (flow or typographic)	- Complex syntax	- <i>Unspecified level of detail</i>	- Guessing or estimation likely	- Responses use wrong units		
- Complex information	- Implicit assumption	- Shortage of (memory) cues	<b>Task Characteristics</b>	<b>Response Structure</b>		
- Not salient	- Several questions in one	- Unanchored time frame	- <i>Non-routine time frame</i>	- Overlapping categories		
	- Unclear goal		- Complex estimation	- Missing response categories		
	- Q/A mismatch		- Potentially sensitive			
	- Missing question		- Social desirability			
	<b>Time frame</b>	<b>Record Retrieval</b>	- <i>Proprietary information</i>			
- Carry-over time frame	- Records unavailable or don't support estimation	- <i>Record access issues</i>	- <i>Strategic factors</i>			
- Undefined time frame	- Record access issues	- <i>Authority issues</i>				
- Embedded / complex time frame	- Authority issues					
- Abrupt change						
- Problematic length						