PROCEEDINGS

QUEST 2007
STATISTICS CANADA
OTTAWA, ONTARIO
CANADA

APRIL 24th to 26th, 2007

Statistics Canada Statistique Canada

Canada

TABLE OF CONTENTS

Preface

The idea for the QUEST (QUestionnaire Evaluation STandards) workshop grew out of a panel on questionnaire evaluation at the 1995 conference on Survey Methods and Process Quality that took place in Bristol England. The first independent QUEST workshop was held in Orebro, Sweden in 1997. Workshops were then held in London, England (1999); Washington, USA (2001); Mannheim, Germany (2003) and Heerlen, Netherlands (2005). These workshops provided a valuable opportunity to exchange experiences and ideas related to questionnaire design and evaluation. On April 24th, 25th and 26th, 2007 Statistics Canada hosted the sixth QUEST workshop in Ottawa. There were 22 participants representing 12 different organizations and 8 different countries.

The original goals of the QUEST workshop were;

- to discuss what various statistical agencies were doing to test questions
- to discuss what was known about the value and effectiveness of various testing methodologies
- and to move toward some reasonable set of standards for question testing before those questions were used in surveys.

Past workshops have had the goals of promoting an exchange of ideas and experiences amongst questionnaire design researchers and describing professional rules or guidelines for questionnaire evaluation. Each workshop participant is directly involved with the evaluation of survey questionnaires.

In this publication papers presented at the QUEST workshop in Ottawa have been compiled, in the order in which they were presented.

QUEST 2007 WORKSHOP PROGRAM

April 24, 25 and 26, 2007
Statistics Canada (STC)
Simon Goldberg Conference Room

TUESDAY APRIL 24TH

8:30 - 9:00	Registration, coffee and tea	
9:00 - 9:15	Welcome	Paul Kelly
9:15 - 9:20	Opening Remarks	François Maranda, Statistics Canada
9:20 - 10:00	Introduction of attendees	Paul Kelly
10:00 - 10:15	*Break*	
10:15 - 11:45	Round table discussion	All
11:45 - 12:00	Planning Committee QUEST 2009	Paul Kelly
12:00 - 13:30	*Lunch at Statistics Canada's Executive Lounge*	
13:30 - 15:15	**Session 1** Questionnaire design – questions, instructions, definitions, etc.	Facilitator – Debbie Collins 4 papers (Cosenza, Fowler, Godenhjelm, Vis)
15:15 - 15:30	*Break*	
15:30 - 16:45	**Session 2** Questionnaire evaluation methods	Facilitator – Deirdre Giesen 3 papers (Esposito, McGee, Pascale)
18:00	Group Dinner Absenthe Café Corner of Spencer and Holland Avenue	

WEDNESDAY APRIL 25TH

8:30 - 9:00	Morning coffee and tea	
9:00 - 10:30	**Session 3** Cognitive interviewing	Facilitator – Birgit Henningsson 3 papers (Beatty, Collins, Lawrence)
10:30 - 10:45	*Break*	
10:45 - 12:00	**Session 4** Issues in qualitative testing	Facilitator – Paul Beatty 2 papers (Allard, Davidsson)
12:00 - 13:30	*Lunch at Statistics Canada's Executive Lounge*	
13:30 - 15:00	**Session 5** Testing of CASI questionnaires	Facilitator – Rachel Vis 3 papers (Dale/Hole, Notnaes, Potaka)
15:00 - 15:15	*Break*	
15:15 - 16:30	Round table discussion Preliminary Steps in the Direction of a Research Agenda for QUEST's Second Decade	All (lead by Jim Esposito)
18:00	Planning Committee Meeting	

THURSDAY APRIL 26TH

8:30 - 8:45	Morning coffee and tea	
8:45 - 10:30	**Session 6** Testing of business and establishment surveys	Facilitator – Paul Kelly 5 papers (Blanke, Henningsson, Levesque, Giesen, Willimack)
10:30 - 10:45	*Break*	
10:45 - 12:00	Final session, overview, plans for 2009	All (lead by Jack Fowler)
12:00 - 13:30	*Lunch at Statistics Canada's Executive Lounge*	

How Much is Too Much?:
How adding information to a question may narrow, instead of expand,
A Respondent's understanding

Carol Cosenza
Center for Survey Research
University of Massachusetts Boston

As researchers, we often struggle with what to include, and not include, in each question. We wonder whether a phrase should be defined or if examples would help clarify the intent of the question. We hope that by providing more information in the question, the respondents will better understand their task in answering the question and eventually we will have better data. At some point though, we have to stop and think about whether the additional information is actually helping the respondent or whether it is more of a distraction.

With funding from the United States National Center for Health Statistics (NCHS), the Center for Survey Research (CSR) at the University of Massachusetts Boston undertook two randomized telephone surveys focusing on methodological issues of question design. We used a split-ballot design in order to test alternate forms of the same question. This paper will examine data collected from these series of question experiments, looking specifically at the effects of adding words or phrases that were intended to be helpful but actually narrowed the focus of the question in the eyes of the respondent.

The Respondents

Table 1 shows the characteristics of the respondents.

	Wave 1 (Spring 2005)	Wave 2 (Spring & Fall 2006)
Number of Respondents	454	425
Gender	39.6% male 60.4% female	37.9% male 62.1% female
Education	36.3% high school or less 26.4% some college 36.6% college or more	29.2% high school or less 29.6% some college 40.0% college or more
Age	mean age= 51.3 years old 49.0% 18 - 49 years old 27.0% 50 - 64 24.0% 65+	mean age= 53.7 years old 40.1% 18 - 49 years old 32.3% 50 - 64 27.7% 65+

The Experiments

Adding Examples - Experiment 1 (Skin Growths)

One of the ways researchers try to make questions easier for respondents to answer is by providing examples. This is especially true in questions that contain abstract nouns or verbs - words that describe a class of more specific items. For example, words like "vegetable" or "exercise" can be thought of as broad categories that can include many different items. While it appears obvious that, for example, playing tennis is definitely exercising, there is a possibility that a respondent may not think about all the different forms exercising can take (for example, whether "walking" should be counted). So, examples are written into the question to show respondents how broad and diverse the concept being asked about may be.

In one test, we asked respondents about growths on their skin.

Table 2: The Question Wording: Experiment 1

VERSION 1: In the last 12 months, did you see or talk to a doctor about a tumor, cyst or growth of the skin? **VERSION 2:** In the last 12 months, did you see or talk to a doctor about any kind of growth on your skin?

Our original prediction was that Alternative 1 would have more positive responses, since there seemed a real possibility that respondents might not think of tumors or cysts as growths on the skin.

Table 3: The Results: Experiment 1

	Version 1 (tumor, cyst or growth)	Version 2 (any growth)
YES - had growth	8.7% (n=13)	18.2% (n=27)

p=.015

Results were surprising. The addition of the examples greatly reduced the number of positive responses. One possibility was that by including "tumor and cyst", Version 1 may appear to be asking about something more serious than the Version 2 which used the vaguer term "any kind of growth". Another possibility is that respondents focused only on the examples given and were basically answering that they did not talk to a doctor about a tumor or cyst (and actually ignoring the more general "growth on the skin"). Regardless of the reason, we found out that more than twice as many people answered that they had spoken to a doctor about "any kind of growth" on

their skin than "a tumor, cyst or growth" on the skin - we had narrowed, rather than expanded what respondents were thinking about.
When looking at respondent-interviewer interaction, we found that Version 2 resulted in many more problems for the respondents. Perhaps this was because of the lack of examples or definition. We found that 9.3% of respondents who were asked Version 2 required some assistance from interviewer, in contrast to only 2% who needed help from Version 1. Most of this assistance consisted of requests for clarification and probing by the interviewer in order to get an adequate answer.

Adding Examples - Experiment 2 (Problems breathing)

We tested another question, this one about breathing problems, where one version had examples listed and the other did not. We worried that not all respondents would think about asthma, pneumonia, or chronic lung disease as "problems with coughing and breathing." Our solution was to add those examples into the alternative version of the question. We assumed that by adding examples, we would increase the number of positive responses.

Table 4: The Question Wording: Experiment 2

VERSION 1: During the past 12 months, did you see or talk with a doctor about any problems with coughing or breathing? **VERSION 2:** During the past 12 months, did you see or talk with a doctor about any problems with coughing or breathing, including asthma, pneumonia, or chronic lung disease?

Table 5: The Results: Experiment 2

	Version 1 (coughing or breathing problems)	Version 2 (longer version with examples)
YES - had problems	19.9% (n=43)	16.7% (n=35)

p=ns

Although it is not a significant difference, more people answered that they saw a doctor for breathing problems when asked about breathing problems alone, rather than as part of a list that includes more serious health conditions. By providing the examples, we could have been inadvertently setting a context for what the respondent thinks we are asking about. Just as we are very conscious of what ranges we use in our answer choices when we ask questions about activities like watching television or drinking alcohol, we need to consider how the words we are using might appear to be clues to the respondent about what is going on in the question and in the survey.

When we behavior coded these questions, we found that very few respondents had problems or asked for clarification or help. As in Experiment 1, fewer respondents asked for help with Version 2 (the version with the examples.)

Adding Cues - Experiment 3 (Talking with a doctor)

Recall is a key step in the cognitive process. There are many ways that researchers try to aid respondents in this. In addition to providing examples, sometimes we add cues to aid in the recall. If, for example, we wanted to know how many times a person ate cheese in a week, we could ask respondents the question and then ask about different situations where cheese might be eaten -- for example, on a cracker, as a snack, as part of a sandwich, or on a salad. In Experiment 3, we were interested in how often respondents communicated with their doctor - either in person or on the phone. We wondered whether respondents would think of telephone calls if they were not specifically mentioned in the question. We asked one version of this question which included the phrase "talk on the telephone" and another that did not mention phone calls.

Table 6: The Question Wording: Experiment 3

VERSION 1: In the past 12 months, how many times have you seen or talked on the telephone with a {DOCTOR} about your health? **VERSION 2:** In the past 12 months, how many times have you seen or talked with a {DOCTOR} about your health?

As expected, we found that there was a wide range of answers - from 0 times to more than 50 times a year, with a mean of 3.34 times for Version 1 (where telephone was specifically mention) and 3.72 times for Version 2.

Table 7: The Results: Experiment 3

	n	Range	Mean
Version 1 (phone)	350	0 - 60 times	3.34 (SD = 6.31)
Version 2 (no phone)	341	0 - 51 times	3.72 (SD = 4.93)

In each version, there were several responses that were outliers, totally skewing the results. We decided to top-code the results at 12 times (which would translate to about once a month). The means and standard error of means are in Table 8 below. As shown, there was a significant difference in the means of the different versions, with people answering they had less communication when the question included the phrase "telephone".

Table 8: The Results (top-coded at 12 times): Experiment 3

	Mean	Standard Error of Mean
Version 1 (phone)	2.71 (SD=3.10)	.166
Version 2 (no phone)	3.36 (SD = 3.27)	.177

(p=.008)

The original intent was to add "telephone" to the question to help jog the memory of people who may not have seen their doctor in person. However, this test shows we created an unintentional result. It appears that respondents focus on the word "telephone" and not on the entire question which is about all communication. This hypothesis was validated by cognitive testing done by NCHS prior to this split ballot project. It found that respondents did indeed focus on the word "telephone" and mistakenly counted less communication because they were not including in-person contact.

Although behavior coding was extremely similar for the two versions, Version 2 was slightly easier for respondents to answer than Version 1.

Conclusions

It is hard to know whether the extra "stuff" we add to questions helps broaden a respondent's thinking or narrows it. We have seen in these three experiments that adding additional examples and cues may not always be the best way to help respondents. In fact, it could harm the data by allowing the respondent to make assumptions about the meaning of the question that was never intended.

We test questions to help us better understand the data we get from the answers respondents give. Cognitive testing, behavior coding and other evaluative techniques can provide us with clues about what is going on during the question-and-answer process. However, only split ballot experiments can help us figure out what happens to the data when we change the words in a question. Split ballot experiments help us figure out if those definitions or examples we add or delete actually impacts what's most important - the results. And, as we've seen from these experiments, results aren't always what you expect.

More Data on When Two Questions Are Better Than One

Jack Fowler

Introduction

The principle of asking one question at a time is fairly well established as a fundamental principle for good question design. When multiple questions are asked in the same question, respondents often have to choose to address one part of the question and ignore the others. We have previously presented data showing that decomposing a complex question into two or more questions seems to produce more valid data. This paper reports results from further testing, comparing results from alternative versions of questions that were designed to meet the same question objectives.

Methods

The purpose of this methodological study was to examine the effects of different approaches to asking questions about complex constructs. A set of questions used in surveys sponsored by the Center for Disease Control were identified as having characteristics that might pose cognitive challenges for respondents. In most cases, that conclusion was based on cognitive testing.

The investigators then designed alternative questions that they thought would address the cognitive issues that were identified. The original and alternative versions of the questions were integrated into parallel survey instruments.

A sample of telephone numbers drawn from clusters of numbers that included listed telephone numbers was drawn from exchanges serving the Continental United States. Interviewers working from a central telephone facility called the numbers and attempted to complete an interview with an adult 18 or older. When an eligible respondent was recruited, he or she was randomly assigned to one of the versions of the survey instrument. The interviews took less than15 minutes to complete.

Because the purpose of the study was to compare answers to alternative questions, rather than to make population estimates, efforts to enlist cooperation were modest. Interviewers made no more than three calls to numbers that were not answered; they did not attempt to convert initial refusals. At the beginning of the data collection, they were free to interview any available adult, but at the end of data collection, in order to help balance the gender of respondents, they tried to interview a household male whenever they could.

A total of 425 respondents were interviewed. Of those, 62 per cent were female; 14% were younger than 35; 27 per cent were 65 or older; 40 per cent had graduated from college; 89 per cent reported their race to be white.

This analysis will focus on two series of questions in which it appeared that the test question was asking two or more questions at once. The alternative version broke the questions down into two or more questions.

Test 1. During the past 6 months, …were you are anyone else …**injured or poisoned** seriously enough that you or they got medical advice or treatment?.

Problem: Question is asking about both poisoning and injury in a single question, which may distract respondents and lead to underreporting, as they try to simplify the cognitive complexity of the task and focus on one problem and not the other.

Alternatives: A. During the past 6 months, ..were you or anyone else…**injured** seriously enough…?

_____ B. During the past 6 months, …were you or anyone else…**poisoned** seriously enough…?

Hypothesis: Two questions will yield more reports of injury and poisoning than the single combined question.

Test 2. **To lower your risk of heart problems or stroke,** has a doctor or other health professional advised you to:

a. Cut down on salt or sodium in your food?

b. Eat fewer high fat foods?

c. Get more exercise?

d. Control your weight or lose weight?

Problem: These clearly constitute two questions at once. One question was whether or not a doctor advised you to "cut down on salt" and the other is whether or not it was to lower your risk of heart problems or stroke. The hypothesis is that some respondents would attend to only one part of the question. Most likely, they would report whether or not they had been advised to do these things and ignore the issue of whether or not it was specifically to "lower your risk of heart problems or stroke". This would lead to over reporting of the specific construct that the question is trying to measure.

Alternative: A. Has a doctor or other health professional ever advised you to cut down on salt…?

B. (IF YES) Did the doctor…recommend this for your general health or specifically to lower your risk of heart problems or stroke.

Hypothesis: The dual question will yield fewer reports than the single "double-barrelled" question.

Results

Test 1. Table 1 presents the results of the first test. The combined question asked about injuries and poisoning in one question, while the "separated" asked about them separately. It can be seen that when the combined question is asked, injuries or poisonings serious enough to be treated by a doctor were reported at the rate of 13 per 100 households. When the questions were separated and asked individually, injuries and poisonings were reported at the rate of 21 per 100 households.

TABLE 1: POISONINGS OR INJURIES REQUIRING MEDICAL CARE BY QUESTION TYPE

QUESTION TYPE	RATE PER 100 HOUSEHOLDS
Combined question	13/100
Separated questions	21/100

$P < .05$

Test 2. Table 2 presents the basic comparison between the number of people who reported that their doctors had advised them to take various health behaviors to reduce their risk of heart attack or stroke. The "combined questions" included the concept of reducing risk into the question while the "separated questions" asked first about whether a behavior had been advised, then followed up with a question about whether it was to reduce the risk of heart attack or stroke.

TABLE 2: PER CENT WHO REPORT DOCTOR ADVISING VARIOUS HEALTHY BEHAVIORS TO REDUCE HEART/STROKE RISKS BY QUESTION TYPE

HEALTHY BEHAVIOR	SINGLE QUESTION	SEPARATED QUESTIONS
Reduce salt	22%	10%
Fewer high-fat foods	45%	17%
Get more exercise	50%	9%
Control or lose weight	39%	5%

It is quite clear that the hypothesis that respondents do not attend to the part of the question about the reason for advising these behaviors is supported. In each case, significantly fewer people reported being advised to do these behaviors specifically reduce their heart or stroke risks when the questions were separated than when they are combined. Further evidence about what is happening is that for the last three health behaviors on the list, the majority of respondents to the separated questions reported that the reason for the advice was not specifically to reduce heart or stroke risk; it was for general health. Only salt reduction was said most often to be specifically tied to heart risks.

TABLE 3: REASON REPORTED FOR DOCTOR ADVISING VARIOUS HEALTHY BEHAVIORS

	REASON DOCTOR ADVISED	
HEALTHY BEHAVIOR	GENERAL HEALTH	REDUCE HEART/STROKE RISKS*
Reduce salt	41%	59%
Fewer high-fat foods	51%	49%
Get more excercise	71%	29%
Control or lose weight	70%	30%

*Includes those who said "both" for general health and to reduce heart/stroke risks

However, the picture is a little less clear when the data from Table 4 are added. The initial question in the series of questions that constitute version 2 asked about getting advice from doctors about the various behaviors without specifying the reason. Given the reasoning above and the results in Table 3, we would expect many more people would have reported getting advice when the reason was not specified than when the reason was specifically to reduce heart or stroke risk. Table 4 shows that this is not the case.

TABLE 4: PERCENT REPORTING DCOTOR ADVISING VARIOUS HEALTHY BEHAVIORS BY WHETHER REDUCTION OF HEART/STROKE RISK WAS OR WAS NOT SPECIFIED AS THE REASON FOR THE ADVICE

HEALTHY BEHAVIOR	HEART/STROKE RISK REDUCTION SPECIFIED IN QUESTION	NO REASON MENTIONED IN QUESTION
Reduce salt	22%	29%
Fewer high-fat foods	45%	44%
Get more exercise	50%	43%
Control or lose weight	39%	30%

Although none of the differences in Table 4 is statistically significant, for three of the four behaviors, the rates at which respondents reported getting advice from a doctor were actually higher, not lower, when reduction of heart or stroke risks was specified.

Discussion

The general principle that consistently has emerged from our studies of question wording seems to be well supported by these data: when multiple concepts are included in a question, respondents attend to only a subset of the issues. Cognitively, they ignore some parts of the question, so that they are answering a simplified and different question. Respondents cannot answer a question about poisoning and injury at the same time. More than half the time, respondents were ignoring the issue of whether or not their doctors' advice was specifically related to reducing their risk of heart problems or stroke. To obtain meaningful and accurate data, questions that require respondents to attend to more than one issue at a time should be decomposed into two or more questions.

Having said that, we are left with the puzzle about why respondents to whom heart risks were not mentioned did not report more advice. Logically, the answers to the question including risk reduction should be a subset of all the reports of advice people received. Table 3 tells us that much advice about health behavior is not linked specifically to risk reduction.

The simplest hypothesis is that respondents whose question included risk reduction just ignored it completely—a result that is consistent with the results in Table 3. In essence, we probably can conclude that the two questions, with or without the mention of risk reduction, are the same question so far as respondents are concerned.

Another possibility is that the mention of reducing risks of heart problems or strokes stimulated recall of conversations about healthy behavior. We know that providing meaningful cues in

questions can sometimes improve recall. So, while those asked the question that included risk reduction should have reported fewer events, this tendency may have been somewhat offset by the stimulus provided to them that helped them recall more such conversations.
Whether or not there was some recall benefit from the inclusion of the concept of risk reduction will have to await another study. However, overall these data, add to the growing body of evidence that trying to ask two questions at once is simply bad question design. When there are two issues that must be addressed in order to answer a question, ask two questions.

Location, scope and amount of definitions and instructions defining the quality of survey response

Petri Godenhjelm

Every year, farmers receive several questionnaires and have to fill them in with data that are needed for the production of statistics describing agriculture. TIKE, the Information Centre of the Ministry of Agriculture and Forestry in Finland, asked Statistics Finland to test two questionnaires during autumn 2006 and spring 2007. The questionnaire of the crop yield inquiry was tested in connection with the normal data collection for this survey whereas the questionnaire of the farm structure survey was tested in spring prior to the actual, regular data collection for the survey. More than one-half of all farms, or over 35,000 farms, in Finland provide data for the farm structure survey which examines farmers' working hours spent on agricultural work and secondary business activities on farms.

The questionnaire of the farm structure survey was tested by examining its completion in the situation when the farmer had to deal with it for the very first time. The data collection design also includes a telephone interview with which Statistics Finland's interviewers then eventually collect the actual survey data. Collecting of the data by telephone has not always run smoothly because the farmers have to go and check the data from their own book-keeping records. However, the telephone interviewing has kept the response rate high and the interviewers have been able to explain unclear items on the questionnaire.

The farmers for the testing were selected by size of farm, age of farmer and area and in the farm structure survey also according to whether the farmer was engaged in secondary business activity. Farmers' attitudes to the testing were more positive than usual and up to four out of five of them agreed to the test interview. All interviews were tape recorded and during some interviews Statistics Finland's researcher accompanied the responsible researcher from TIKE. During the testing, the farmers thought aloud about their answers while filling in the questionnaire. In addition, they were asked extra questions about certain parts of the questionnaire that had been agreed in advance as being of specific interest.

Visual look matters

In recent research, a separate, fifth phase involving observing and focusing attention to the information on the questionnaire has been inserted to the beginning of the conventional testing model comprised of comprehension, retrieval, judgement and response. Before a respondent goes through the different phases, his/her attention is drawn to the visual information that influences the way he/she answers or whether he/she answers at all to a presented question. (Dillman 2007.)

The interviews of the farmers showed clearly how numbering, layout and symbols guided navigation on the questionnaire and, at the same time, also comprehension of the questions. The questionnaire contained matrix-format questions which the respondents interpreted on the basis of the information in the columns. Understandably, a column cannot contain much information. Although definitions were given in separate boxes above the matrix of answers, they were not nearly always noticed.

However, because instructions contain information that is relevant to answering, the researcher should consider thoroughly what information is the most important to communicate to ensure that the received answers will be comparable. At this point the researcher must make a strategic choice about what, how and how much information should be communicated to the respondent. The contents, volume and positioning of questionnaire instructions may influence considerably both the quantity and quality of obtained answers. Meaning of concepts is communicated not only in questions, but in definitions and instructions as well.

Indeed, questionnaires that are self-administered by respondents often contain matrix-like structures where one row may contain several column headings and columns which, together with cases in rows, actually make up the asked question. Zukerberg and Lee (1998) observed in their studies that respondents study column headings quite closely. They used this observation as a justification for locating instructions next to column headings as far as possible. This way, instructions can be brought closer to the cognitive task presented to the respondent.

Some of the encountered problems arose from the questionnaire's concepts, and their location and quantity on it. Vagueness in the navigation path, meaning here the intended order in which answering should progress, caused slight problems to answering. For example, yield of grass fodder was often given for the whole year although it was asked about by crop on the questionnaire. "Oh, I see, this should've been broken down, it was never like this before, first crop and second crop, then I have to smudge this once more…should I put it down as second crop, because we never got one, or as third crop because that's when it was harvested, second crop came to nothing, it wasn't worth gathering…I'll put it down as third crop." It was suggested that the navigation path should be made clearer at this point of the questionnaire, so that the respondents would notice the relevant information and understand immediately that they should answer crop by crop.

Calculation problems and instructions

This survey also contained problems connected with calculation. These could arise from a farmer using his own, perhaps even inconsistent unit for calculating crop volumes, which could be a cubic metre or bale depending on the method use for preserving fodder on the farm. However, the questionnaire asked for the amount of fodder in kilograms. It was not always clear how bales or cubic metres should be converted to kilograms. To solve this during the telephone interviewing phase, the statistical interviewers used conversion tables from which amounts could be established in kilograms irrespective of whether the farmer had the information as numbers of bales or cubic metres. However, these conversion tables were not available to the farmers in the answering situation, which caused problems and increased response burden. "If I now had to think in hundreds of kilos, I'd have to count all those bales, find out which came from which field…it's a bit of a guesstimate, putting kilos down, so it'd have to be number of bales…but then this second crop went straight to fodder, as it was so poor this year, light I mean, I'd say a quarter came form the second crop. How much might a bale of dry hay weigh, should probably ask the old farmer."

I myself had both the interviewing instructions and the conversion tables with me when testing the questionnaire. If calculating numbers of kilograms had caused problems, I asked the

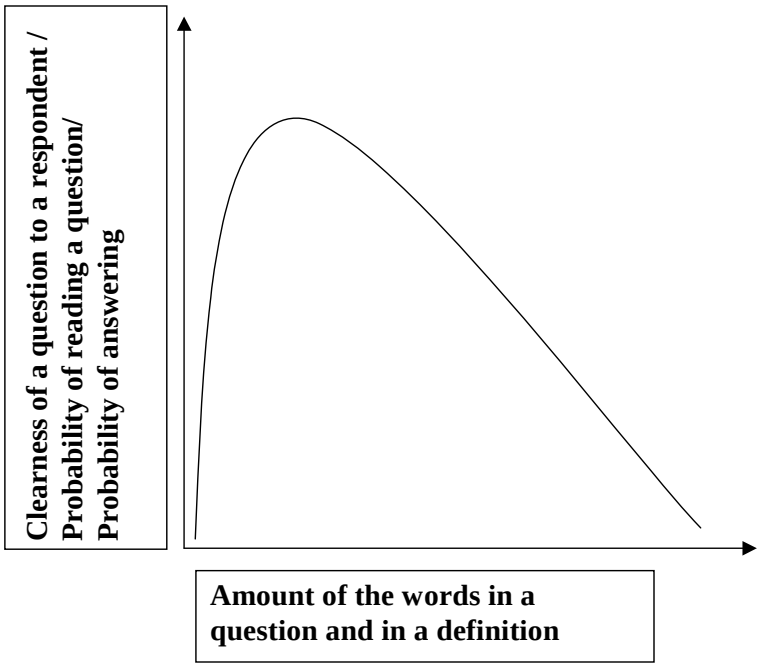
interviewee's opinion about the conversion tables after the interview. All farmers thought that it would be useful to have the conversion tables to hand when filling in the questionnaire.

Volume and positioning of instruction must be optimised in inquiries

Endeavours to design a questionnaire that meets the conceptual aims set for it may on the other hand lead to making compromises in respect of other guidelines on good questionnaire design. Trying to make a question as clear as possible by either adding words to it or to its additional instructions can make the question or its instructions so long that respondents cannot be bothered to read them thoroughly or ignore them altogether. (Zukerberg and Lee 1998.) According to Conrad and Schober (2005) clarification of all questions is not even necessary because it not only takes time in interviews but also space on questionnaires.

This means that the designer of a questionnaire must make far-reaching decisions. On the one hand, he/she must make sure that the essential content of the question is communicated, but on the other, ensure that the question and the related instructions are not read superficially or bypassed altogether. How definitions and instructions are used and placed have important implications to the comparability of responses and survey quality. The adjacent figure sketches the questionnaire designer's position in this kind of a decision-making situation. The peak of the curve describes the ideal situation.

Figure 1 Questionnaire designer's dilemma



References:

Conrad, F. G., Schober, M. F. (2005): Promoting Uniform Question Understanding in Today's and Tomorrow's Surveys. Journal of Official Statistics Vol. 21, No. 2.

Dillman, D. A. (2007): Mail and Internet Surveys. The Tailored Design Method, 2007. Update With New Internet, Visual, and Mixed-Mode Guide. Jon Wiley & Sons Inc. SanFrancisco.

Zukerberg, A., Lee, M. (1998): Better Formatting For Lower Response Burden. U.S. Bureau of the Census.

Cognitive Pre-Tests on How to Present "Don't Know" Help Texts in Web Surveys, Results from the First Test

Rachel Vis-Visschers[1]
Statistics Netherlands, Division of Methodology and Quality

1. Introduction

Primary data collection is expensive. Researchers are continuously searching for more cost efficient ways of data collection. To this end, Statistics Netherlands (SN) has decided that in 2009 most social surveys will be executed in a mixed-mode design, in which web interviewing will be an important mode.
Mixed-mode data collection is seen as more cost efficient, yet there are also some drawbacks to this way of collecting data: the "mode effects". To maintain data quality and to prevent data discontinuities the mode effects have to be minimized. (Van den Brakel et al, 2006; Gilljam and Granberg, 1993; Presser and Schuman, 1989; DeRouvray and Couper, 2002)
Mode effects can occur in all steps of the statistical process; e.g. different non-response in different modes, different answers to similar questions due to oral or visual presentation of the questions. There should be research into the different possible mode effects, so that the effects can be either prevented or corrected.
SN's Questionnaire Laboratory will focus on preventing mode effects by studying mixed-mode questionnaire design. The Questionnaire Laboratory (Q-lab) will execute several pre-tests in 2007. The aim of these pre-tests will be to come up with question formats that work for multiple-mode surveys.
In this paper we will discuss the first results of a pre-test in which we investigated the best way to present the answer category "don't know" (DK) and the help texts in web surveys so that the results are comparable to results in CAPI/CATI-modes.

2. The Test

The test took place in the period January until April 2007.

2.1 The Aim

In this laboratory test we wanted to:

- find the most efficient way to present DK on screen, so that use of the answer option in the web questionnaire is comparable to that in a CAPI/CATI setting.
- find the most efficient way to present help texts (i.e. texts to explain the content of a question, not texts to explain how to use the questionnaire) on screen, so that use of the function in the web questionnaire is comparable to the use in a CAPI/CATI setting.

[1] With special thanks to Cees van Berkel, Dirkjan Beukenhorst, Ralph Dolmans, Deirdre Giesen, Frans Kerssemakers, Jos Logister, Frank Meessen and Marco Puts.

2.2 The answer category "don't know"

For questionnaires that have to be filled out by respondents on their own, like web questionnaires, it is more difficult to asses whether the questions are understood as they are intended and it is a challenge to keep the respondents motivated enough to finish the questionnaire, but also to keep them from giving satisficing answers. In interviewer administrated surveys the interviewers can help the respondents answer the questions correctly and they also motivate the respondents to finish the questionnaire and not to choose "don't know" too often.

At SN the interviewers are instructed to present the DK-option only if stated explicitly in the questionnaire text. Still a respondent can refuse to answer, or can answer "don't know", but the interviewers can probe whether the respondent really cannot give an answer. Only if a respondent indicates that he really does not know the answer they are allowed to accept this response.

It is a challenge to try and "translate" this interviewer instruction and the influence of the interviewer to a web questionnaire. The way in which DK is presented in different modes can have an effect on the data from the different modes.

In this laboratory test four different "translations" were explored:

1. DK always on screen as an answer category.
   The "don't know" option will be presented with the other answer categories for all questions. From previous experiments we know that this increases the chance that the option will be chosen (e.g. Van den Brakel et al, 2006). For this laboratory test we expect that the amount of DK's will be highest in this variant.

2. DK is never presented on screen.
   In this variant "don't know" is never an option. The respondents are forced to select an answer, even if they do not know, in order to continue with the questionnaire. This will possibly result in non-attitudes or pseudo attitudes. We expect that respondents can get irritated. In a lab setting it will be highly unlikely, but in reality it will be very probable that a respondent will quit before finishing the questionnaire.

3. DK is hidden. If the respondent tries to skip the question a warning appears that all questions have to be answered, and the DK-option is presented under the other answer categories. The procedure is explained at the beginning of the questionnaire.

   This option expects a deliberate action from the respondent. It resembles the procedure with an interviewer where the respondent first has to indicate that he does not know an answer after which the interviewer tries to motivate the respondent to give another answer only to accept DK as a last alternative.

   In this test we want to find out whether a respondent is aware of this hidden option. It is possible that even though there is an explanation at the beginning of the questionnaire that a respondent does not read it or forgets about it in the heat of the moment. Other questions are, how do respondents react to the warning that they cannot skip a question? Do they see the DK-option appear after they have tried to skip a question?

4. DK is presented less visually prominent. At bottom left side of every screen a DK-button is presented.

   The idea behind this variant is that it is always possible to answer don't know, as it is in an interviewer administrated survey, but answering it is made a bit less inviting because the option is put away from the other response categories. A similar option was tested in DeRouvray and Couper (2002), where DK was presented together with the other answers, but in a lighter colour. This experiment had some drawbacks and they advise that it should be tested differently.

   In this test we want to find out whether this variant is indeed "less visually prominent", and does it really resemble an interviewer administrated survey.

Beforehand we know that explicitly presenting DK and not presenting it at all (options 1 and 2) are not optimal. We expect that the best way will be somewhere in between, i.e. options 3 or 4.

2.3 The help texts

As was said before, in interviewer administrated surveys the interviewer can assist the respondent in how to understand or interpret the questions. The interviewer has several clues at hand to help a respondent. There are the standardized help texts or instructions in the questionnaire, and there is the interviewer training, but there is also the "general knowledge" of the interviewer personally. Interviewers are instructed not to help a respondent overly much, in order to prevent them from steering the respondent in a certain direction.
How can this all be translated into a web questionnaire? It is fairly impossible to include the background knowledge of all interviewers in a web questionnaire. So this is left untouched in this test. The standardized help texts and instruction and the instruction from the interviewer training can be put literally in the questionnaire. Still how should it be presented on the screen?

For this test two alternatives were developed:

1. Clickable button. The help texts can be accessed by clicking on the ?-button (Figure 1). The button is placed directly under the question, before the answer categories.

2. Mouse-over button. The help texts can be accessed by moving the mouse over the ?-button. It is not necessary to click on the button. The button is placed directly under the question and before the answer categories.

During the test we want to find out whether the respondents consult the help texts and whether they express a preference for either one of the variants.

Figure 1. The button for the help texts.

2.4 The Questionnaire

The test called for a challenging questionnaire. We wanted to measure the use of the DK-option and the help texts, so the questions should be composed of difficult questions to which one really

would not know the answer or for which one needs an instruction. Since the results from the test would result in recommendations for SN's surveys we only used questions from existing SN survey questions. The question consisted of eight sections (Table 1) and there were different versions of the questionnaire all built in Blaise IS (Table 2). In the appendix several screen shots of the questionnaire are presented.

Table 1. The different section of the questionnaire

Section	**Aim**
1. Opening:	Introduction to the questionnaire and instruction.
2. Background information:	Gathering background information from respondent.
3. Occupation and education:	Gathering background information from respondent and offering help texts.
4. Accommodation (questions on type of housing, number of rooms etc.):	Offering help texts.
5. Neighbourhood satisfaction (attitude and opinion questions on safety, neighbours etc.):	"Don't know" questions.
6. Environmental behaviour (attitude and opinion questions):	"Don't know" questions.
7. Holiday (questions on last holiday):	Offering help texts.
8. Conclusion	Closing the questionnaire.

Table 2. The different versions of the questionnaire.

Version	**"Don't know"**	**Help texts**
A	Always on screen	Clickable button
B	Always on screen	Mouse-over button
C	Never on screen	Clickable button
D	Never on screen	Mouse-over button
E	Hidden	Clickable button
F	Hidden	Mouse-over button
G	DK-button	Clickable button
H	DK-button	Mouse-over button

2.5 The Respondents

In the most ideal situation the test respondents should represent the whole of Dutch society. Since it is only a small test this is not feasible, but we can try to recruit a divers group based on gender, age and computer experience.

For the test a sample of 120 persons living in towns near the office of SN was drawn. The recruitment letters were addressed to these persons. In the letters the aim of the test was explained and a participation fee of € 20 was announced. A short introduction to SN and the work of the Q-lab were included. The respondents could respond by sending back the application form (post-paid). After two weeks all non-respondents were sent a reminder.

Recruiting test respondents proved very difficult. We sent 120 recruitment letters, but only 12 people responded. Of these twelve, two did not want to participate. Of the ten who wanted to

participate, nine actually did. In table 3 some background information on the respondents is presented.

Table 3. Some background information of the respondents.

R	Gender	Age	Education	Employed
1	F	62	Vocational Education	No
2	M	57	Vocational Education	No
3	F	31	Secondary School	Yes
4	F	58	Higher Vocational Education	Yes
5	F	50	Secondary School	Yes
6	F	18	Secondary School	No
7	F	54	Secondary School	Yes
8	F	42	Secondary School	No
9	F	37	Higher Vocational Education	Yes

2.6 The Method

The tests were held at the office of SN in Heerlen. The interviews were executed according to a fixed protocol. The test respondents were invited to come to the office and were asked to fill in a questionnaire on the computer. Each respondent was assigned a specific version of the questionnaire. The respondents were asked to think out loud and the interviewer observed the respondent while filling in the questionnaire, this was followed by an in-depth interview with follow-up questions. At the end of the interview we explained the reason for the test and we showed the different versions of DK and the help texts. As a final question we asked whether the respondents had a preference for one of the test versions. Everything was recorded on video. The respondents' answers and other findings were recorded in MS Excel. Later on the spread sheets were used for the report.

3. Results

3.1 The use of "don't know"

- Most respondents indicate that they did not have the wish to answer "don't know", because they know the answers to the questions.
- One respondent explicitly said that she preferred any answer above "don't know".
- Two persons had wanted to answer "don't know", but did not. In one case the respondent, who was assigned to the hidden DK option, did not think it was an option, since it was not on the screen. She did not try to skip the question and thus did not find out that DK appeared then. She did not remember the announcement at the beginning of the questionnaire. In the other case the respondent, assigned to the DK button, did not see the button. In the heat of the moment she had "forgotten" about it.

3.2 The use of the help texts

- Most respondents indicate that they did not have the need for extra help.
- Only one person did find the "?"-button on her own.

- Five of the nine test persons only found the button after a hint from the interviewer.
- The respondents indicate that the button is not clear: Two respondents said that they had seen the question mark, but had not thought about its purpose. One person was confused by the place of the button. To her it seemed as if the instruction would be about a part of the question that she did understand, thus she did not open the instruction.
- There is no difference between the use of the clickable button and the mouse-over button.

3.3 Other results

Though we were not testing navigational issues explicitly, we found the following:

- Six people started by asking the interviewer how to move to the next page.
- Most respondents used the mouse for selecting an answer and [enter] to move on.
- Only a minority used the navigation buttons.
- Most respondents indicate that the navigation buttons are too far away from the questions.

4. Conclusion

The results we found were not the results we were aiming at. We did not really find much about the use of the DK-options and the use of the help texts. What we did find is:

- Recruiting test respondents only by means of recruitment letters does not work.
- This questionnaire was not "challenging" enough, there was no need for the respondents to answer "don't know" or to consult the help texts.
- The question mark on the help text button is not self explanatory. If respondents search extra information they do not "see" the ?-button.
- Most remarks that were made concerned navigational issues. The navigation buttons are placed too far away from the question and thus the respondents prefer to use the [enter]-key.

Considering the effort it had taken to recruit nine test respondents and the results we gathered, we concluded that we would not continue this test, even though we initially wanted to test 30 respondents.

5. Future

At this moment, July-August 2007, the Q-lab is repeating the test. For this test we have:

- developed a new challenging questionnaire, i.e. we have selected difficult questions and we did not always use questions from SN surveys;
- applied a different layout, e.g. with the navigational buttons closer to the questions and instead of the question mark we put the text "Instruction" on the help text button;
- employed a different way of recruiting test respondents, e.g. recruitment letter followed by a telephone call.

References

Presser, S. and H. Schuman. 1989. The management of a middle position in attitude surveys. In Survey Research Methods, edited by E. Singer and S. Presser, 108-123. University of Chicago.

Gilljam, M. and D. Granberg. 1993. Should we take don't know for an answer? In Public Opinion Quarterly 57: 348-357.

DeRouvray, C. and Couper, M. 2002. Designing a Strategy for Reducing "No Opinion" Responses in Web-Based Surveys. In Social Science Computer Review, Vol. 20 No.1, Spring 2002: 3-9

Van den Brakel, J., Vis-Visschers, R. and Schmeets, J. 2006. An Experiment with Data Collection Modes and Incentives in the Dutch Family and Fertility Survey for Young Moroccans and Turks. In Field Methods, Vol. 18, No. 3, August 2006: 321-334.

Appendix. Screenshots of the questionnaire

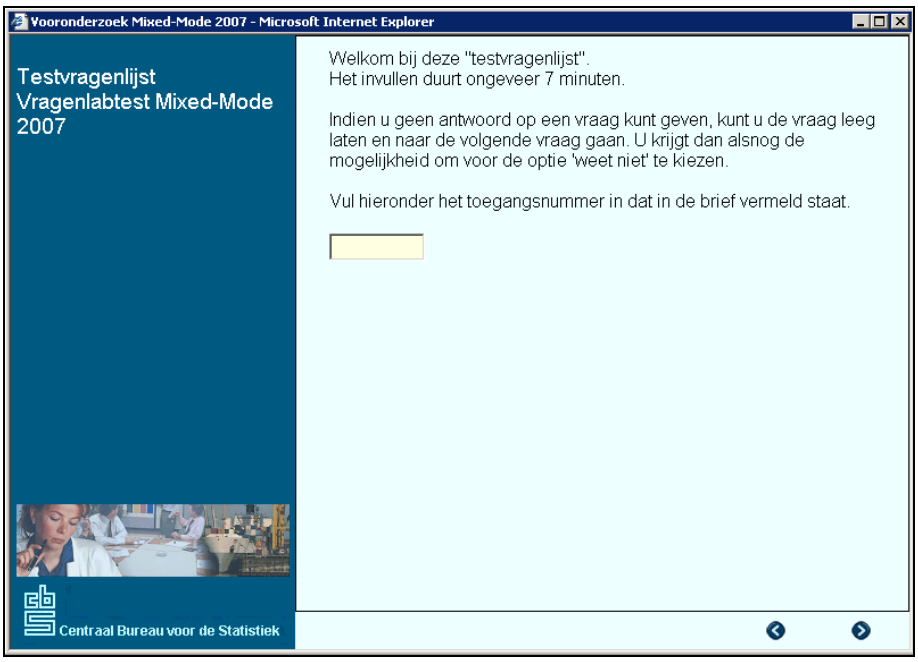


Figure 1. The introduction page. For the hidden DK-option there is the explanation that the respondent can answer "don't know" by first trying to skip the question.

Figure 2. In this screen the DK-option (in Dutch "Weet niet") is presented on screen. Also the clickable help text button can be seen under the question text. In figure 3 the help text is unfolded.

Figure 3. In this screen there is no DK-option at all. The clickable help text button has been activated and the help text is unfolded.

Figure 4. In this screen the DK-button (in Dutch “Weet niet”) is visible in the menu bar below left. Also the help text is made visible by moving the mouse pointer over the mouse-over button.

The Utility of Metadata in Questionnaire Redesign and Evaluation Research

James L. Esposito
U.S. Bureau of Labor Statistics [2]

A. Introduction

A1. Objectives. The objectives of this paper are to accomplish the following: (1) to draw attention to the concept of metadata by noting its scope and relevance; and (2) to describe a case study involving the measurement of labor force status that illustrates the utility of metadata in evaluating and redesigning questionnaire items. Survey methodologists draw upon conceptual metadata (e.g., definitions of key terms; item-specific objectives) and operational metadata (e.g., interviewer manuals; classification algorithms) not only when developing or redesigning survey questionnaires, but also when conducting presurvey and postsurvey evaluation research. And information generated in the course of conducting evaluation research (e.g., qualitative and/or quantitative data gathered during interviewer debriefings, behavior coding, et cetera) also constitutes metadata. It is difficult to imagine how survey practitioners could design/redesign questionnaires intelligently without recourse to survey-related metadata.

A2. Metadata: What is it? Statistics Canada (2006) defines/describes metadata as follows:

"Metadata provide information on all data published by Statistics Canada in order to inform users of the features that affect their quality and **fitness for use**. The information includes the definitions of the variables and description of their classification schemes, the description of the methodology used in collecting, processing and analyzing the data, and information on the accuracy of the data."

This definition captures the essence of the concept in general terms. For our purposes here, let us define *metadata* more specifically as any information (verbal or numeric or code, qualitative or quantitative) that provides context for understanding survey-generated data, and that includes but is not limited to the following: (1) ethnographic observations/information regarding the domain-of-interest; (2) specification of measurement objectives and domain-specific concepts; (3) question wordings, objectives and ancillary item-specific instructions; (4) details regarding data-collection modes; (5) instructional materials provided to interviewers and/or respondents; (6) documentation of presurvey and postsurvey evaluation research; and (7) survey-specific classification algorithms and imputation procedures. [Note: For more formal discussions of metadata and metadata registries, see Dippo and Sundgren, 2000; and Gillman, 2004.]

B. Case Study: The Measurement of Labor Force Status in Two U.S. Surveys

The focus of the case study described below has to do with the measurement of *labor force status* (i.e., the state of being employed, unemployed or not-in-the-labor-force) in two major demographic surveys conducted in the United States. This case study draws on recent research conducted during the redesign of the American Community Survey [**ACS**] and on earlier research conducted during the redesign of the Current Population Survey [**CPS**]. As noted above, the

[2] **Disclaimer:** The views, opinions and perspective expressed in this paper are those of the author and do not necessarily reflect those of the Bureau of Labor Statistics or the Bureau of the Census. The author is deeply committed to survey research that is collaborative, nonpartisan and socially constructive in all respects.

purpose of the case study is to illustrate the utility of metadata in evaluating and redesigning survey questions/questionnaires. I hope to achieve this goal by establishing connections between CPS metadata (e.g., prior evaluation research), ACS-relevant metadata (e.g., subsequent evaluation research) and ACS redesign activities and decisions (i.e., the approach taken in using evaluation research to modify question wording and sequencing in the redesigned ACS labor force series).

B1. The Current Population Survey [**CPS**]. The CPS, which is sponsored jointly by the Bureau of Labor Statistics [**BLS**] and the Bureau of the Census, is the principal/official source of household-based labor force data about the American non-institutional population. Each month, Census Bureau interviewers administer the CPS to approximately 60,000 households; several weeks later, usually on the first Friday of the month, the BLS Commissioner draws on CPS data to report the nation's monthly unemployment rate and other important labor force data.

CPS Labor Force Concepts: Work and Employment. The CPS is used to classify eligible household members into one of three labor force categories: employed, unemployed or not-in-the-labor-force. In classifying persons as *employed*, it is important to understand that the CPS measures **work**, not jobs. The concepts of *work* and *employment* are defined as follows:

"Work includes any activity for wages or salary, for profit or fees, or for payment in kind. One hour or more of such activity constitutes work. Work also includes unpaid activity of at least 15 hours a week on a family business or farm." (U.S. Census Bureau, 1999, page B1-2)

"Employed people are those who, during the reference week (a) did any work at all (for at least 1 hour) as paid employees; worked in their own businesses, professions, or on their own farms; or worked 15 hours or more as unpaid workers in an enterprise operated by a family member or (b) were not working, but who had a job or business from which they were temporarily absent because of vacation, illness, bad weather, childcare problems, maternity or paternity leave, labor-management dispute, job training, or other family or personal reasons whether or not they were paid for the time off or were seeking other jobs. Each employed person is counted only once, even if he or she holds more than one job." (U.S. Census Bureau, 2006a, page 5-2)

With regard to efforts to assess measurement error, explicit definitions like those provided above for the concepts of work and employment (i.e., conceptual metadata)—and other relevant concepts not defined here (e.g., job; payment in kind; reference week)—are of absolutely critical importance. Critical in what sense? Critical in the sense that it would be very difficult (if not impossible) for survey practitioners to determine the extent to which respondents understand specific survey concepts and questions if subject-matter specialists (e.g., survey sponsors and others knowledgeable about the domain-of-interest) did not explicitly define the concepts embedded in their survey questions. To the extent that a disparity exists, it is the mismatch or gap between sponsor-specified definitions of key survey concepts and respondent-centered understandings of those concepts that represents the most basic (though not the only) source of measurement error. *Ceteris paribus*, the greater the disparity between sponsor-versus-respondent perspectives, the greater the magnitude of measurement error.

The CPS Redesign. A multiple-phase redesign of the CPS was undertaken in the early 1990s that involved changes both to the content of the labor force questionnaire and to the methodology for collecting data (i.e., from pencil-and-paper to computer-assisted data collection). During the course of the three-phase redesign, a multiple-method evaluation plan was used (e.g., behavior coding, interviewer and respondent debriefings, split-ballot testing) to compare the existing CPS questionnaire to various design alternatives. Given the importance of the CPS as a statistical measure of the "health" of the economy, this redesign generated a substantial amount of interest

and metadata (e.g., Bregger and Dippo, 1993; Cohany, Polivka and Rothgeb, 1994; Polivka and Miller, 1998; Rothgeb and Cohany, 1992).

The CPS "At-Work-Last-Week" Questions. The redesigned CPS relies on about 16 questions to generate estimates for its three major labor force categories (employed, unemployed or not-in-the-labor force) and for various other subcategories (e.g., unemployed *looking*; unemployed *layoff*)—see Butani, Alexander and Esposito, 1999/2000 (pp. 108-109). That said however, most of the burden for classifying a person as *employed* is carried by the *at-work-last-week* questions [see **Table 1**, items (4) and (5)]; moreover, given the relationship among the three principal categories, the classification accuracy of the remaining two labor force categories (unemployed and not-in-the-labor force) depends to a large extent on how well persons are classified as employed—the largest of the three categories. For these reasons, and in order to simplify the discussion as much as possible, the primary focus of this case study will be the measurement of work (i.e., the *at-work-last-week* questions). [Note: The *at-work-last-week* questions used in the CPS *prior to 1994* can be found in Table 1, items (1) and (2).]

Given the brevity and apparent simplicity of the *at-work-last-week* questions [items (4) and (5)], it was somewhat surprising to learn from CPS interviewers that these two questions generate considerable confusion when initially posed to some survey respondents. For example, during the final evaluation phase of the CPS redesign (1992-1993), these questions were rated by interviewers as the *most problematic* items on the redesigned questionnaire. Interviewers consistently reported that respondents would reply to this question by saying, "Just my job", or asking, "Do you mean my regular job?" Other evaluation data collected via postsurvey respondent-debriefing questions indicated that a small percentage of paid work (1.6%) was being missed, though that percentage was not substantively different relative to the percentage for the corresponding *at-work-last-week* question on the control questionnaire.[3] Still other evaluation data (e.g., behavior-coding data and response-distribution analyses) indicated that the problems experienced by some respondents in understanding the *at-work-last-week* question did **not** represent a serious data-quality issue, largely because of the high likelihood of interviewer mediation and "repair work". For example, whenever a respondent would answer "just my job" or "do you mean my regular job", experienced interviewers would simply enter "yes" for this item and then move on to the next series of questions.

The methodological research conducted during the CPS redesign was documented extensively by researchers at the BLS and the Census Bureau (e.g., Campanelli, Martin and Rothgeb, 1991; Esposito and Hess, 1992; Esposito and Rothgeb, 1997; Martin and Polivka, 1992 and 1995; Rothgeb, Campanelli, Polivka and Esposito, 1991). These publications, which represent evaluation metadata, are important for understanding and communicating the potential for error that exists when measuring and differentiating the American labor force using CPS questions and the CPS data-collection methodology. It stands to reason then, in view of the information such documentation provides, that any survey organization that chooses to integrate a subset of CPS labor force questions into one of their surveys and that chooses to use a different data-collection

[3] The first of two debriefing questions used to assess missed *paid work* was worded as follows: "In addition to people who have regular jobs, we are also interested in people who may only work a few hours per week. LAST WEEK, did [name] do any work at all, even for as little as one hour?" The utility of this question in identifying instances of missed *paid work* should help to explain our decision to use very similar wording when revising the *at-work-last week* question in the redesigned ACS labor force series [see Table 1, item (9).]

methodology, runs the risk of incurring unacceptable levels of measurement error in its attempt to assess labor force status.

Table 1. "At-Work-Last-Week" Question Wording on Various Survey Questionnaires

Current Population Survey [1990; paper questionnaire]

(1) What were you doing most of last week: working, keeping house, going to school or something else? [Interviewers proceed to item (2) for responses other than "working".]

(2) Did you do any work at all LAST WEEK, not counting work around the house? (Interviewer Note: If farm or business in the household, ask about unpaid work).

Decennial Census [1990 long form; this wording was used for cognitive testing in 1995]

(3) LAST WEEK, did this person work at any time?
- Yes—Mark this box if the person worked full time or part time. (Count part-time work such as delivering papers, or helping without pay in a family business or farm. Also count active duty in Armed Forces.]
- No—Mark this box if this person did not work, or only did school work, volunteer work or his or her own school work.

Redesigned Current Population Survey [1994 to present] ******

[No business in the household]

(4) LAST WEEK, did you do ANY work for pay?

[With business in the household]

(5) LAST WEEK, did you do any work for either pay or profit?

Decennial Census [2000; long form] and the **American Community Survey** [2000-2007]

[Computer-assisted version (CATI and CAPI)]

(6) LAST WEEK, did you do ANY work for either pay or profit?
[Read if necessary: Include any work even if (she/he/you) worked only 1 hour, or helped without pay in a family business or farm for 15 hours or more, or (was/were) on active duty in the Armed Forces.]

[Self-administered questionnaire (SAQ) version]

(7) LAST WEEK, did this person do ANY work for either pay or profit? *Mark (X) the "Yes" box even if the person worked only 1 hour, or helped without pay in the family business or farm for 15 hours or more, or was on active duty in the Armed Forces.*

National Content Test [2006] and the **Redesigned American Community Survey** [2008]

(8) **28a:** LAST WEEK, did this person work for pay at a job (or business)?
- Yes—SKIP to question 29
- No—Did not work (or retired)

[**Note:** If "no" to question 28a, the respondent is supposed to read and answer 28b.]

(9) **28b:** LAST WEEK, did this person do ANY work for pay, even for as little as one hour?

**** Note:** In the CPS, prior to asking the *at-work-last-week question*, the reference period is explicitly defined for respondents as follows: "I am going to ask a few questions about work-related activities LAST WEEK. By last week, I mean the week beginning Sunday, January xx, and ending on Saturday, January xx." It is not clear what instructions interviewers and respondents receive regarding the reference week in the other surveys listed above.

Impact of CPS Redesign (1990-1993) on the Decennial Census (2000). The "long form" of the Census Bureau's decennial census gathers data on a broad range of demographic topics (e.g., population, housing, disability status, labor force status, educational attainment, health insurance). With regard to the collection of labor force data, the decennial census follows the CPS; that is to say, it incorporates a subset of the full CPS labor force series, it adheres to CPS labor force concepts, and it assigns individuals to one of the same three labor force categories as does the CPS. So when the CPS was redesigned in the early 1990s, this redesign work had implications for the decennial-census labor force series. As a result, in the mid-1990s, as part of its planning for the 1996 National Content Test (also know as the 2000 Census Test), the Census Bureau's Center for Survey Methods Research [CSMR] was authorized to conduct some cognitive testing on a modified version of the 1990 labor force series and other topical series (Bates et al., 1995/2007). With regard to the *at-work-last-week* question asked in 1990 [see Table 1, item (3)], researchers noted that, in the 18 opportunities they had to observe respondent behavior, respondents appeared to have little difficulty in answering this item; however, they did not always read the instructional information that was provided after the "yes" and "no" response options. CSMR researchers recommended that the question be reworded as follows:

"LAST WEEK, did this person do ANY work for pay? Mark "yes" even if the person worked only 1 hour. Also, mark "yes" if the person is on active duty in the Armed Forces" (Bates et al., 1995/2007, p. 21).

The researchers also noted that it was the preference of a high-level BLS program manager "… that the wording of the question be as close to the CPS wording as possible" [see Table 1, items (4) and (5)]. It is not really clear what evaluation work was conducted on this item (and the rest of labor force series) after 1995 and prior to 2000. It appears that the *at-work-last-week* question may have been modified to include the phrase "pay *or profit*" and that response-specific instructional material may have been enhanced *prior to* the 1996 National Content Test, but I could find no documentation to support these speculations. A subsequent Census Bureau report intimates that the labor force series was evaluated successfully during this field test (Luckett-Clark et al., 2003, p. 79), but again I was unable to find any specific documentation along these lines or any mention of split-ballot research comparing the revised decennial labor force series and the redesigned CPS labor force series during the period 1996-1999.

B2. The American Community Survey [**ACS**]. The ACS, which has been designed to replace the "long form" of the decennial census (and which is very similar to it in both content and structure), was developed over a series of stages and achieved full implementation in 2005 (U.S. Census Bureau, 2006b, Chapter 2). The ACS is the largest continuous survey conducted in the United States and has a sample of 250,000 residential addresses every month. Data are collected continuously via three modes: (1) A self-administered, mail-out/mail-back questionnaire (**SAQ**), which accounts for about two-thirds of the ACS data actually collected from respondents; (2) centralized telephone interviewing (**CATI**); and (3) personal-visit/computer-assisted personal interviewing (**CAPI**). For any given monthly sample of addresses (e.g., May 2007), ACS data are collected in three stages: By SAQ during the first month (May 2007); by CATI during the second month (June 2007); and by CAPI during the third month (July 2007). Given time and cost constraints, it is not possible to gather ACS data from every sampled household in any given data-collection cycle. In view of this reality, the cases that remain after completing the second stage of interviewing are sampled at a rate of approximately 1:3 before the final personal-visit

stage commences. [For more details on ACS sampling, data collection and capture, see U.S. Census Bureau, 2006b, Chapters 4 and 7.]

ACS Content. Like its predecessor, the decennial census long form, the ACS gathers data on a broad range of demographic topics (e.g., population, housing, disability status, labor force status, educational attainment, health insurance). At present (2007), the ACS draws on the content of *ten* CPS questions to construct its *six-item* labor force series, and this abbreviated/reconfigured series of items is used to generate estimates for three labor force categories: employed, unemployed, not-in-the-labor-force. Given substantial differences in data-collection methodology (e.g., mode of administration; imputation procedures), one would not expect the two surveys to produce statistically equivalent labor-force estimates.

ACS At-Work-Last-Week Questions. The ACS *at-work-last-week* question asked by interviewers during CATI and CAPI interviews [Table 1, item (6)] is similar in wording to the CPS *at-work-last-week* questions [see items (4) and (5)]; however, in contrast to the CPS items, both ACS questions include a *read-as-necessary* statement [see items (6) and (7)]. The read-as-necessary statement is particularly important in the self-administration (SAQ) mode, because it contains conceptual information that is critical for understanding and answering the question correctly and because there is no interviewer available to provide this information should respondents be motivated to request clarification as to question intent or meaning. Very highly motivated respondents have the option of calling an "800" helpline for assistance; alternatively, they can read an instructions booklet that is enclosed with the ACS data-collection packet.

In generating ACS estimates of labor force status (and of other demographic variables), the Census Bureau draws on mail survey responses (SAQ), telephone responses (CATI) and personal-visit responses (using CAPI), as noted previously. The approximate weighted response percentages for the three modes are 51%, 9% and 38%, respectively; unit nonresponse accounts for the remaining 2% (U.S. Census Bureau, 2006b; p. 7-3). So while mail-out/mail-back/SAQ questionnaires account for nearly two-thirds of the questionnaires completed and returned by respondents, this mode actually contributes about 51% of the data used in generating estimates; and, again, this is because of the sampling that takes place prior to the third/personal-visit phase of ACS data collection and the subsequent weighting of data by mode. In contrast to the ACS, about a quarter of CPS interviews are conducted by personal visit (via CAPI) and the remaining interviews are collected by telephone (via CATI or CAPI, but mostly the latter); there are **no** self-administered questionnaires in the CPS. The CPS nonresponse rate is approximately 8%.

B3. Manifest Differences between the CPS and the ACS. As noted in the prior two subsections, there are substantial differences between the CPS and the ACS (e.g., number and content of labor force questions; mode of administration), and these differences have implications for labor force classification. At a government-sponsored conference in late 1999, Butani, Alexander and Esposito (1999/2000) presented a paper that briefly described and discussed many of the more important differences between the two surveys and came to the following conclusion regarding the comparability of labor force estimates:

"While it will **not** be possible to assign the target person to one of the seven CPS labor force categories using the [six labor force] questions on the ACS, it will be possible to assign persons to one of the three major [labor force] categories (employed, unemployed, not in labor force). Given the different sets of questions asked (and ignoring for the time being other important considerations, like mode effects), at issue here is whether the two surveys are capable of producing equivalent estimates with respect to these three major [labor force] categories. In our view,

considering question content **alone**, the two sets of [labor force] questions (ACS vs. CPS) are capable of producing fairly similar estimates of major labor force categories. One key estimate, the unemployment rate, will **not** be equivalent—but the discrepancy may not be that large. In the CPS, most (but certainly not all) persons are classified as unemployed on the basis of answers provided to a sequence of four questions (LK, LKM1, LKAVL, and LKAVR). The ACS asks the equivalent of three of these questions. The only question not specifically asked in the ACS is [LKM1: "What are all the things you have done to find work during the last four weeks?"]. [Note: *Field testing would need to be conducted in order to determine actual CPS-versus-ACS differences for each major labor force category (employed, unemployed, not in labor force*).]" (Butani, Alexander and Esposito, 1999/2000, pp. 106-107; italics added for emphasis)

Regarding differences in the mode of data collection, and alluding to prior research conducted during the CPS redesign, the authors specifically identified the *at-work-last-week* question as a potential source of measurement error:

"Given that interviewers will not be on hand to answer questions respondents might have when completing the [mail versions of the] ACS, one might presume that data quality for the self-administered ACS cases will be inferior to that collected by interviewers for both the ACS and the CPS. The following example illustrates a *potential data-quality problem*. During the quality assessment phase of the redesign of the CPS, interviewers consistently reported that some respondents were experiencing confusion on how to respond to CPS item WK: "LAST WEEK, did you do ANY work for (either) pay (or profit)?" A common response to this question was: "Just my job." Apparently this happens fairly frequently, but it is not a serious data quality issue for the CPS because interviewers know to code such answers as a "yes" response and move on. It is not known how respondents will deal with this confusion in the self-administered context of the ACS. Almost certainly, some percentage of respondents will answer "no" to [the ACS work question] (assuming that this question is asking about work *other than* that associated with their jobs) and, as a consequence, will be classified as either unemployed or not in labor force by the ACS. *This is a potentially serious data-quality problem*." (Butani, Alexander and Esposito, 1999/2000, page 107, italics added for emphasis)

These lengthy quotations are reproduced here because they represent domain-appropriate metadata (i.e., explicit statements/predictions germane to asking CPS labor force questions on the ACS) that essentially anticipate and describe, *in part*, where labor force estimates for the ACS might be expected to diverge from CPS estimates, and why. The evaluation work upon which these statements/predictions were grounded is equally relevant as metadata (i.e., research reports published during and after the CPS redesign). Yet, somewhat inexplicably, the paper by Butani, Alexander and Esposito is not referenced in any of the Census Bureau publications that appear in the reference section of the present paper; nor are there any citations to the relevant empirical research that was conducted during the early 1990s.[4]

B4. Research Comparing Labor Force Estimates: The CPS/Census-2000 Match Study. In an effort to evaluate the quality of data being gathered from labor force questions on the decennial census questionnaire [see Table 1, items (6) and (7)], and in view of the status of the CPS as the official source of employment and unemployment data, the Census Bureau undertook quantitative research that compared labor force estimates generated by households that had completed both the decennial census and the CPS over a four-month period in early 2000 (Palumbo and Siegel, 2004; also see, Adams, 2004). This research indicates that, relative to the CPS, the decennial census underestimates employment and overstates the percentages of the other two labor force categories (see **Table 2**). While these differences may appear small, each

[4] When checking a citation index, I could find the Butani, **Alexander** and Esposito paper cited only *once* in any article authored by a Census Bureau professional, and that was a very interesting paper on "rolling samples" written by Charles Alexander (Alexander, 2001)—the brilliant statistician (an **co-author** of the paper in question) who was the principal architect and advocate of the ACS at the Census Bureau before his untimely death in 2002.

percentage point represents about two million individuals. Recalling now that the labor force items asked on the decennial census (2000) and the ACS (2000-2007) are identical, this research suggests that, relative to the CPS, the ACS is also likely to underestimate employment and overstate the percentages of the other two labor force categories—and subsequent research by Vroman (2003) provides support for this presupposition.

Table 2. Labor Force Estimates from the CPS/Census Match Study [1,2]

Labor Force Category	CPS	Decennial	Decennial minus CPS
Employed	64.14 %	62.31 %	- 1.83 %
Unemployed	2.65 %	3.37 %	+ 0.72 %
NILF	32.79 %	34.01 %	+ 1.22 %

Source: Adams (2004; data above are drawn from the margins of her Table 2, p. 3238).

Note 1. Combined-month sample: February through May, 2000, specific rotations. Total number of cases with final labor force codes on both surveys: 13,249. Weighted sample: 207,875,749.

Note 2. For calendar year 2000, the wording and structure of labor force items used in the decennial census and the ACS questionnaire were virtually identical.

The disparity between CPS labor force estimates, on the one hand, and decennial-census (and ACS) estimates, on the other, raised concerns at the Census Bureau. The CPS-Census Match Study had quantified the differences and had identified a number of possible factors that might be responsible for the disparities (e.g., imputation methods; differences in question wording; mode of administration; reference periods), but this study was not designed to address more qualitative issues, like how a respondent's cognitive and motivational processes might be mediating responses to somewhat different labor force questions being asked under manifestly different administrative conditions. In recognition of such limitations, the principal authors of the match study offered a set of recommendations for improving ACS data, the following specifically addressing the need for additional research of a qualitative nature:

"Research aimed at improving the accuracy of the American Community Survey [labor force] data through questionnaire improvements must include a large component of cognitive/behavioral research to develop new questions or approaches prior to pre-testing them [in the field]. This evaluation suggests that the effects of shortcomings in the [labor force] questions may be too subtle to detect in [field-based] pre-tests alone." (Palumbo and Siegel, 2004, Executive Summary, p. ix) [**Note:** The parenthetical material regarding field-based testing was added by the present author in an effort to possibly clarify meaning.]

And,

"Attempts to revise the American Community Survey [labor force] questions should proceed by evolutionary or incremental means. The evaluation results suggest that the existing questions, in spite of their likely flaws, likely have many virtues as well." (Palumbo and Siegel, 2004, Executive Summary, p. ix)

In response to such recommendations presumably, and under the auspices of the Office of Management and Budget, the Census Bureau initiated a broad-based research effort to redesign various ACS question sets, including the labor force series; this research was done collaboratively with the BLS (labor force series) and other Federal agencies (other question sets).

B.5. Evaluation of the ACS Labor Force Series. In 2004, researchers at the Census Bureau and the BLS formed a *labor-force-series working group* [hereafter simply, the *working group*] and collaborated on a small-scale evaluation the ACS labor force series; both subject-matter experts and survey practitioners were involved in this process. The evaluation methods included expert reviews, behavior coding, focus groups (with ACS interviewers) and cognitive interviews. Expert reviews were conducted by four individuals, three of whom were involved in the redesign of the CPS back in the early 1990s. While the reviews covered all of the items in the ACS labor force series, much of the content of those reviews (and of the discussions that followed) focused on the *at-work-last-week* questions. Drawing largely on evaluation data from the CPS redesign, and in view of manifest differences between the CPS and ACS (e.g., question wording and mode of administration), one of the expert reviewers (Esposito, 2004a) proposed that the basic *at-work-last-week* question be asked in two parts: The first part of the question would ask if the target person had worked at a job for pay last week and, if not, the second part would ask about marginal work (i.e., paid work for as little as one hour). A final decision by the *working group* as to their recommendations for revising the *at-work-last-week* question (and other items in the ACS labor force series) would not be made until other evaluation work had been completed.

Behavior coding was conducted at two telephone centers staffed by Census Bureau interviewers; this evaluation work focused on the CATI version of the ACS *at-work-last-week* question [Table 1, item (6)]. A total of 51 household interviews were monitored and a total of 104 ACS person interviews were coded (Esposito, 2004b). Though interviewers appeared to experience some difficulties reading the *at-work-last-week* question as worded (e.g., percentage of exact question readings, 78%; percentage of major changes in question wording, 10%) relative to comparative findings reported years earlier during the CPS redesign (e.g., exact readings in the 94-100% range; Esposito and Rothgeb, 1997), there were extenuating circumstances; for instance, ACS interviewers were sometimes aware of the respondent's employment status on the basis of responses to disability questions that *preceded* the labor force series. On the other side of the exchange, even though respondents provided a high percentage of answers that could be coded as adequate (98%), about twenty percent of these responses were something other than a straightforward "yes" or "no" answer (e.g., "For pay, yes."; "Just his regular job."; "No, currently unemployed."). Moreover, during the monitoring and coding process it was observed that the read-if-necessary statement, which contains important instructional and conceptual information, was *never read*.

During the same week that behavior coding was conducted, two groups of ACS interviewers (fifteen in total) were debriefed using a focus group format; one of the groups consisted of bilingual interviewers exclusively (Esposito, 2004c). During the focus groups, interviewers were asked to identify those labor force items that respondents had experienced difficulty answering adequately; once they had done so, and prior to having a detailed discussion of the items so identified, interviewers were asked to rate problematic items on a five-point scale (1 to 5)—the more difficult for respondents to answer adequately, the higher the rating. The bilingual group of interviewers rated the *at-work-last-week* question as the most difficult item in the labor force series (rating of 2.83); the other group rated the item as the second most difficult item in the series (rating of 2.89). During the ensuing discussions, interviewers mentioned a variety of problems with the *at-work-last-week* question. For example, several interviewers reported having respondents who answered: "Well, yeah, I worked (in a testy tone)"; or "I had a job". Other respondents, most likely multiple-job holders or self-employed persons, also experienced some confusion (e.g., "Did you mean, other than my regular job?"). Several bilingual interviewers

mentioned that some of their Latino respondents would answer "no" to the *at-work-last-week* question, presumably confused by the wording "pay *or profit*." Certain interviewers, anticipating these sorts of problems, simply asked respondents if they had a job, either in lieu of asking the scripted *at-work-last-week* question or after asking it. And many interviewers stated candidly that they rarely, if ever, read the read-if-necessary statement to respondents, not really knowing for sure when that information might be needed.

After reviewing memoranda that summarized the opinions of expert reviewers and findings from behavior coding and interviewer debriefings, the *working group* met to discuss options for modifying the *at-work-last-week* question (and other items in the ACS labor force series). Based on their review of these evaluation materials, they recommended that the *at-work-last-week* question be asked in two parts; the first item in the two-item set would ask about working *at a job* for pay and the second item, if needed, would ask about marginal work of an hour or more during the reference week [see Table 1, items (8) and (9)]. The *working group* made recommendations for modifying the wording of other items in the labor force series, but these other changes were relatively minor. With the revised question set in hand, the modified labor force series was then evaluated by Census Bureau researchers who conducted a total of 40 cognitive interviews on the redesigned series (Rothgeb, 2007). The outcome of that cognitive research was very favorable. The two-item approach to asking about work last week appeared to function as intended, with little evidence of confusion or misunderstanding on the part of research participants.

B6. The 2006 ACS Content Test. In this final chapter of the case study, I summarize briefly the major findings of the 2006 ACS Content Test, which was a large split-panel field test incorporating modifications to all of the substantive ACS question series, including the labor force series (Holder and Raglin, 2007). The test commenced in January 2006, ended in March 2006, and involved approximately 62,900 residential addresses equally split between the control survey (i.e., the then-current ACS questionnaire) and the test survey (i.e., the redesigned ACS questionnaire). The principal findings of this test are presented in **Table 3** and they indicate that the test/redesigned labor force series captured a significantly higher percentage of employed persons and a significantly lower percentage of persons not-in-the-labor than did the control/existing series; the difference observed in the estimate for the unemployed category was not statistically significant. With the exception of the latter result, these are the outcomes that the *working group* had sought to achieve in making changes to the ACS labor force series. Moreover, other evaluation data collected during the field test indicated that the level of response error associated with the employed and not-in-labor-force categories was lower for the redesigned labor force series relative to the control series, suggesting that the estimates from the redesigned series are more accurate.

Table 3. Labor Force Estimates from 2006 National Content Test

Labor Force Category	Control ACS	Test ACS	Test minus Control
Employed	62.8 %	65.7 %	+ 2.9 % **
Unemployed	4.1 %	3.6 %	- 0.5 %
NILF	33.1 %	30.7 %	- 2.4 % **

Source: Holder and Raglin (2007; see their Table 2, p. 12). Chi-square test significant at the 0.0024 level.
Note. Approximately 62,000 residential addresses were sampled with half assigned to the test panel and half to the control panel.

C. Closing Remarks

The anomaly that served as the stimulus for this paper was first observed during the early stages of the CPS redesign when interviewers asserted that the *at-work-last-week* questions—the shortest and most-straightforward questions on the CPS questionnaire—were giving some survey respondents a difficult time. In fact, many interviewers had reported that these questions were the most problematic items on the entire survey (Esposito and Hess, 1992; Esposito and Rothgeb, 1997). How could that be, I wondered. There are no unusually difficult or technical words to cause comprehension problems (e.g., work, pay, profit). The domain-of-interest, *employment*, is familiar to almost everyone. The reference period is short (one week) and precisely defined. And the question itself is succinct and situated very early in the labor force series. How could any respondent struggle with such a simple question? To this query, there appears to be no obvious or simple answer (but see the **addendum** for some speculations). This curious anomaly had become embedded in my memory and had remained there in a semi-dormant state until I was asked to collaborate on a conference paper juxtaposing two high-profile government surveys, the Current Population Survey and the American Community Survey (Butani, Alexander and Esposito, 1999/2000). When I learned almost five years later (circa 2004) that the two surveys were generating discrepant labor force estimates, I had a very strong "intuition" of what the principal source of the disparities might be, given prior work on the CPS redesign and the significant mode differences between the two surveys. What surprised me at that time was that the ACS sponsor seemed unaware of the existence of a possible connection between metadata produced during the CPS redesign (e.g., evaluation reports and other published works) and discrepancies that had been found in the labor force estimates for the two surveys. Yes, the Census Bureau had done some impressive quantitative research (e.g., the CPS/Census-2000 Match Study)—*after* a subset of the redesigned CPS labor force items had been incorporated into the census-2000 questionnaire—and had formulated some compelling hypotheses as to what the most likely problem sources might be (see Luckett-Clark et al., 2003, Appendix 1; Palumbo and Siegel, 2004, pp. vii-ix and pp. 45-48). But they had apparently overlooked (or discounted) the obvious: The extensive literature from the CPS redesign—and other relevant metadata. Had program managers reviewed this literature carefully and field-tested the imported CPS labor force items *prior to* Census 2000, they might have learned in the mid-to-late 1990s what they later learned in 2004 (CPS/Census-2000 Match Study) and in 2006 (National Content Test): That a solitary *at-work-last-week* question would not yield acceptable labor force estimates—and that a two-item approach possibly could. A missed opportunity, to be sure, and one with unfortunate consequences: Eight years of less-than-optimal labor force data (ACS, 2000-2007) and an anticipated series break in 2008. If ever a case study was potentially instructive as a means of illustrating the importance and the utility of metadata, this one seemed to qualify.

So what have we learned from this experience? The following lessons seem worthy of note: First, it would appear that our best hope for minimizing measurement error in survey redesign work is to conduct a thorough and critical review of relevant metadata, and to use such reviews as the basis for making subsequent design-and-evaluation decisions. Second, as one key aspect of the often long and labor-intensive process of evaluating and redesigning questionnaire items, quantitative research, alone, is not sufficient—there needs to be a balance between qualitative and quantitative research. Third, it is risky to export question sets from one survey questionnaire to another under the assumption that such commerce will yield estimates in the recipient survey that approximate the estimates of the parent survey; and this seems especially true when questions from the parent survey have been modified in various ways and when the two surveys differ

significantly in terms of data-collection methodology (e.g., mode of administration). To state the obvious, the equivalence of survey estimates cannot be assumed; comparability has to be established empirically—and one reliable method of doing so would be to conduct pre-production, split-ballot field tests. And lastly, even survey questions that appear simple and straightforward from the perspective of a design specialist or a survey sponsor *may not be viewed as such* when heard or read by survey respondents. When unexpected difficulties with specific questionnaire items are observed during an evaluation phase, they should always be documented—even when other findings suggest that such anomalies are harmless. One never knows when such metadata might be useful in minimizing measurement error in another survey context. At the very least, you might have stumbled upon a puzzle that other behavioral scientists and survey practitioners might find both stimulating and challenging.

ADDENDUM

One of the frequently recited principles of good questionnaire design is the following: *Keep your questions clear and concise*. Okay then, how about this: "Last week, did you do any work for pay?" That's pretty clear, and concise, too—only nine words. The long version contains twelve words: "Last week, did you do any work for either pay or profit?" So why do you suppose so many (CPS and ACS) respondents have difficulty answering "yes" or "no" to this question? [They *do* have difficulty, trust me. I found it hard to believe, too.] Feel free to take a few minutes to reflect upon this curious finding before reading further.

After about fifteen years of pondering this anomaly (on-and-off, of course), I suspect the answer may lie in one or more of the Gricean maxims—the two that appear relevant in this particular case are the *Maxims on Quantity* (Grice 1975):

(1) "Make your contribution as informative as is required (for the current purposes of the exchange)."
(2) "Do not make your contribution more informative than is required." (Grice, 1975, p. 45) [5]

Let's consider a hypothetical example that focuses on the self-administered version of the ACS *at-work-last-week* question [see Table 1, item (7)]. Assume the respondent has a full-time job, a high-school education, and a long-standing distaste for surveys that run on for six pages or more.

- *Respondent reads to himself (and superficially scans the italicized information that follows the question):* LAST WEEK, did this person do ANY work for either pay or profit?
- *Respondent thinks to himself:* How should I answer this [#!&?@] question? It's doesn't mention a **job** and probably would if that's what they wanted to know. And it specifically says "*work* for pay or profit", so it must mean doing work on the side. OK, just check the "**no**" box.

Analysis (part a.): From a design perspective, the question wording and the accompanying italicized instructions communicate information to the respondent that is needed to measure the

[5] Needless to say, within the context of survey design and administration, a questionnaire design team would have to know a great deal about its respondents to be able to design survey questions that satisfy both of these maxims simultaneously; violations are inevitable for certain individuals. In the case of interviewer-administered surveys, the presence of trained and experienced interviewers helps to compensate for what the design team *does not know* about individual respondents in that interviewers can mediate when there is some misunderstanding or design flaw. Such is not the case in self-administered questionnaires and, as a result, this administration mode is more problematic.

concept of work and to answer the *at-work-last-week question* accurately; that is, according to ACS (and CPS) specifications. Not being an economist, the respondent interprets the question from his lay perspective, observing that the question does not mention a "job"; he concludes that the question is probably asking about work-on-the-side. There is no interviewer available from whom to request clarification and the likelihood of the respondent reading the enclosed instructional booklet or calling an "800" helpline for assistance is low. The respondent decides to mark "no" in the answer box of the *at-work-last-week* question and then moves on to other questionnaire items. [Before concluding that this is a bogus scenario, recall how CPS respondents often reply to this question (e.g., "Just my job") and how CPS interviewers proceed in such cases (i.e., they typically enter "yes" and move on without further discussion).]

Assessment (part a.): By not making the question "as informative as possible (for the current purposes of the exchange)," the questionnaire designer appears to have violated Grice's *first maxim on quantity*. Given that most people who work do so in the context of their jobs, to satisfy maxim (1), some reference to a "job" appears to be needed in the wording of the *at-work-last-week* question, even though the objective of this question is **not** to determine if a person had a job but rather to determine more simply if the person worked during the reference week.

Analysis (part b.): The respondent notes that the question specifically states "work for pay or profit", which would seem superfluous to a person with a full time job: Who in their right mind works all those hours for free? Again, the respondent concludes that this question is probably asking about work-on-the-side. This seems consistent with the thought processes noted above and reinforces an answer of "no" to the question.

Assessment (part b.): Even though the questionnaire designer has captured the work concept precisely, she/he appears to have violated Grice's *second maxim on quantity* by providing more information than is required ("work *for pay or profit*"). Almost everyone who works at a job (or business) does so for pay (or profit); so to add these words, *for pay or profit*, seems redundant and only serves to confuse *some* respondents. Again, when there is no interviewer around to clarify intent and no strong motivation on the part of the respondent to determine what the intent might be, the likelihood of measurement error increases dramatically.

Resolution: One design option for addressing the issues raised by this scenario—the design option selected for the redesigned ACS labor force series—is to ask the *at-work-last week* question as a two-item set. The first question [Table 1, item (8)] specifically mentions "job", "work for pay" and "business". The second question in the set [item (9)], is asked only of respondents who answer "no" to the first question and is designed to capture those who work for pay for "as little as one hour" during the reference week, a group that is often not identified as employed and therefore misclassified as unemployed or not-in-the-labor-force. While not a perfect solution (from a Gricean perspective), the results of the 2006 National Content Test (Table 3) suggest that this two-question approach is more effective than the one-question approach in measuring employment and in assessing labor force status, overall.

REFERENCES

Adams, Tamara (2004). Examining Differences in the Labor Status in the Current Population Survey and the Census 2000. *Proceedings of the Section on Survey Research Methods, American Statistical Association.* Alexandria, VA: American Statistical Association, pp. 3233-3240.

Alexander, C.H. (2001). Still Rolling: Leslie Kish's "Rolling Samples" and the American Community Survey. *Proceedings of the Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective.* Ottawa, Ontario: Statistics Canada.

Bates, N., Bogen, K., DeMaio, T., Gerber, E., Hess, J., Jenkins, C., Martin, E., Moore, J., Rothgeb, J., and von Thurn, D. (1995/2007). Cognitive Interview Results and Recommendations for the 1996 National Content Test. *Study Series Report (Survey Methodology: 2007-12)*. Washington, DC: CSMR, U.S. Census Bureau.

Bregger, J. and Dippo, C.S. (1993). Overhauling the Current Population Survey: Why is it necessary to change? *Monthly Labor Review*, 116(9), pp. 3-9.

Butani, S., Alexander, C. and Esposito, J. (1999/2000). Using the American Community Survey to Enhance the Current Population Survey: Opportunities and Issues. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Statistical Policy Working Paper 29*, pp. 102-111.

Campanelli, P.C., Martin, E.A., and Rothgeb, J.M. (1991). The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data. *The Statistician*, 40, pp. 253-264.

Cohany, S. R., Polivka, A. E., & Rothgeb, J. M. (1994). Revisions in the Current Population Survey effective January 1994. *Employment and Earnings*, 41-2, pp. 13-37.

Dippo, C.S. and Sundgren, B. (2000). The Role of Metadata in Statistics. *Proceedings of the Second International Conference on Establishment Surveys.* Alexandria, VA: American Statistical Association, pp. 909-918.

Esposito, J.L. (2004a). ACS Employment-Status Series: Expert Review. Memorandum from Esposito to Palumbo [12 October 2004].

Esposito, J.L. (2004b). Preliminary Behavior Coding Data [HTC]: ACS Employment Status Items. Memorandum from Esposito to Rothgeb [16 October 2004].

Esposito, J.L. (2004c). Preliminary Interviewer Debriefing Findings/Data [HTC and TTC]: ACS Employment Status Items. Memorandum from Esposito to Rothgeb [18 October 2004].

Esposito, J.L. and Rothgeb, J.M. (1997). Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In *Survey Measurement and Process Quality*. L. Lyberg, P. Biemer, M. Collins, E. deLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.). New York: Wiley, pp. 541-571.

Esposito, J.L., and Hess, J. (1992). The Use of Interviewer Debriefings to Identify Problematic Questions on Alternate Questionnaires. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Gillman, D. (2004). ISO/IEC 11179: Framework for a Metadata Registry. *Joint UNECE/ Eurostat/OECD Work Session on Statistical Metadata (METIS).* Geneva, Switzerland.

Grice, H.P. (1975). Logic and Conversation. In P. Cole and J.L. Morgan (editors), *Syntax and Semantics.* New York: Academic Press, pp. 41-58.

Holder, K. and Raglin, D. (2007). Evaluation Report Covering Employment Status. Final Report. *2006 American Community Survey Content Test Report, P.6.a.* Washington, DC: U.S. Census Bureau.

Luckett-Clark, S., Iceland, J., Palumbo, T., Posey, K, and Weismantle, M. (2003). Comparing Employment, Income and Poverty: Census 2000 and the Current Population Survey. *A Census-2000 Evaluation Report.* Washington, DC: U.S. Census Bureau.

Martin, E. (1987). Some Conceptual Problems in the Current Population Survey. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association.* Alexandria, VA: American Statistical Association, pp. 420-424.

Martin, E., and Polivka, A.E. (1992). The Effect of Questionnaire Redesign on Conceptual Problems in the Current Population Survey. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association.* Alexandria, VA: American Statistical Association, pp. 655-660.

Martin, E.A. and Polivka, A.E. (1995). Diagnostics for Redesigning Survey Questionnaires: Measuring Work in the Current Population Survey. *Public Opinion Quarterly*, 59, pp. 547-67.

Palumbo, T. and Siegel, P. (2004). Accuracy of Data for Employment Status as Measured by the CPS-Census 2000 Match. *Census 2000 Evaluation Report B.7.* Washington, DC: U.S. Census Bureau.

Polivka, A.E., and Miller, S.M. (1998). The CPS after the Redesign: Refocusing the Economic Lens. In *Labor Statistics Measurement Issues*, NBER Studies in Income and Wealth, Volume 60, University of Chicago Press, pp. 249-286.

Rothgeb, J.M. (2007). ACS Labor Force Questions: Results from Cognitive Testing. *SRD Research Report Series (Survey Methodology #2007—16).* Washington, DC: U.S. Census Bureau (Statistical Research Division).

Rothgeb, J. M., & Cohany, S. R. (1992). The Revised CPS questionnaire: Differences between the Current and Proposed Questionnaires. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association.* Alexandria, VA: American Statistical Association, pp. 649-654.

Rothgeb, J., Campanelli, P.C., Polivka, A.E. and Esposito, J.L. (1991). Determining Which Questions Are Best: Methodologies for Evaluating Survey Questions. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association.* Alexandria, VA: American Statistical Association, pp. 46-55.

Statistics Canada (2006). About Definitions Data Sources and Methods [web metadata found at http:/www.statcan.ca/english/concepts/background.htm]. Ottawa, Ontario: Statistics Canada.

U.S. Census Bureau (1999). Current Population Survey Interviewing Manual [CPS-250]. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau (2006a). Current Population Survey: Design and Methodology. Technical Paper 66. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau (2006b). *Design and Methodology*, *American Community Survey*. Washington, DC: U.S. Government Printing Office.

Vroman, W. (2003). Comparing Labor Market Indicators from the CPS and ACS. Unpublished draft report. QUEST2007 // Paper 073007

How do we assess whether we are improving instrument design? Using multiple methods to evaluate whether a re-designed travel record was 'better' than the existing one

Alice McGee

This paper focuses on how to determine whether attempts to improve instrument design has been successful. In addressing this issue it draws on a recent project, undertaken within the Question Design and Testing (QDT) Hub within the Survey Methods Unit at NatCen, to demonstrate how multiple evaluation methods were used to find whether a re-designed Travel Record was 'better' than the existing one.

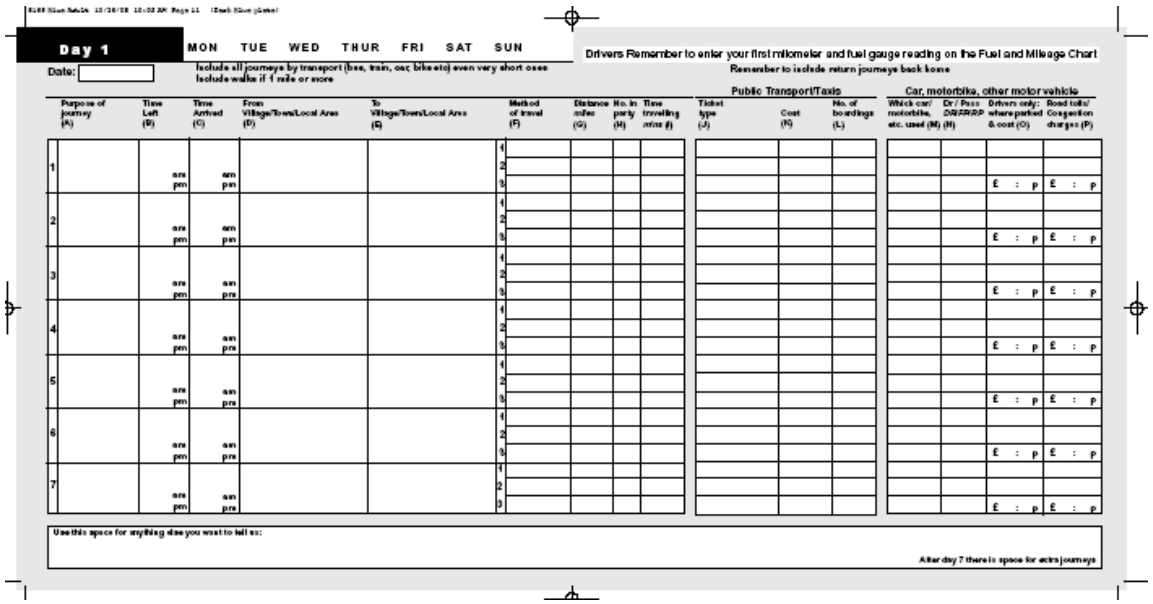
In this paper, I provide background information on the Travel Record used as part of the National Travel Survey (NTS) and outline the rationale for its re-design. Next I discuss the range of evaluation methods, both quantitative and qualitative, that were triangulated in assessing the existing Record and its proposed replacement. Lastly, I consider how well these evaluation methods worked in complementing each other to provide sufficient evidence to indicate that the new Record was indeed an improvement on the existing one and thus should replace it.

National Travel Survey and seven-day Travel Record

The QDT Hub was commissioned to evaluate the existing Travel Record used as part of the National Travel Survey (NTS). The NTS is a large, continuous, national cross-sectional survey that collects detailed information about people's travel patterns and behaviour, and is commissioned by the Department for Transport (DfT). Information is first collected at the household level via a face-to-face CAPI interview; and following this each member of the household[6] is asked to complete a seven-day Travel Record. Figure 1 illustrates a recording page from the existing Record.

[6] Both adults and children are asked to complete a Travel Record. The child version differed slightly to the adult version in that it did not ask for details on car parking and charges and included a question on one day about time spent in the street.

Figure 1 Existing Travel Record recording page



The Travel Record adopts a 'row approach' where information about each journey is entered horizontally across the page. Information in columns A-E is entered at the journey level (e.g. purpose of journey, time left and time arrived) and columns F-P contain details about each stage of that journey (e.g. the distance travelled and the time spent travelling). A stage could be walking, taking the bus or taking the tube for example, as long as each stage is part of that journey (i.e. travelling to a certain place, for example to work).

Rationale for re-design

There were two key concerns that led to DfT's decision to review the existing design of the NTS Travel Record: (a) worries surrounding data quality; and (b) international evidence supporting alternative Travel Record designs.

Taking first the issue of data quality, DfT had various concerns about the data yielded from the Travel Record, for instance the proportions of 'missing' data for particular data items were found to be high. Furthermore, anecdotal evidence from interviewers suggested that respondents could become confused at certain data items, increasing the possibility of measurement error (i.e. where there is a gap between the 'true' value and the response obtained).

Secondly taking international evidence, a review of the ways in which travel behaviour data are collected suggested that there may be an alternative format in which to display the Record that would make it easier for respondents to navigate and fill in. Essentially this would involve altering the existing NTS Travel Record from a 'row-based' format (see Figure 1 above) to a 'columnar' format (where information about each journey is collected in a column)[7].

Thus, DfT commission this piece of work and a three-stage research programme was proposed:

[7] Axhausen K (1995), Travel Diaries: An annotated catalogue, 2nd edition (draft)

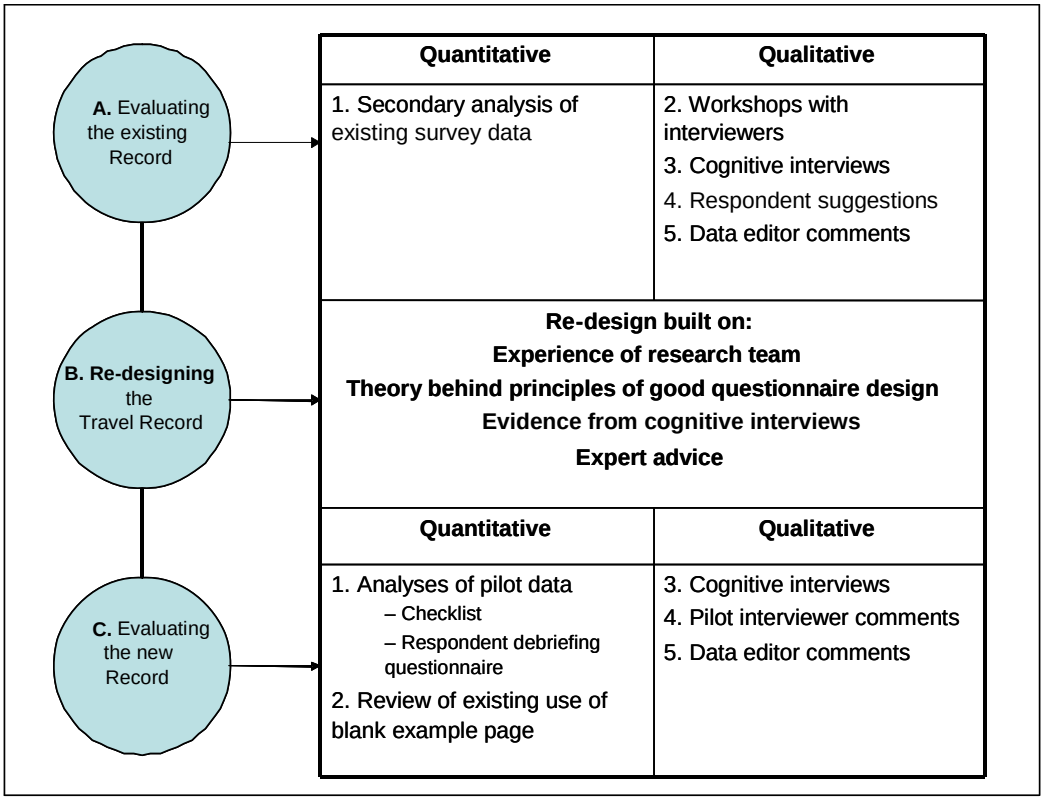
- Stage A involved an evaluation of the existing Record using a number of complementary methods;
- Stage B would involve a re-design of the Travel Record, based on (i) existing theoretical knowledge in terms of general principles for good questionnaire design; and (ii) the findings at Stage A and;
- Stage C would involve the subsequent evaluation of the new Record, again using a range of different methods, to assess whether it was 'better' than the original.

It was agreed that Stages B and C would only take place if evidence from Stage A indicated that the format of the existing Travel Record was problematic and therefore that changing the format might resolve some of these problems.

Evaluation methods

Figure 2 below shows the range of evaluation methods used for each of the three stages of the research process.

Figure 2 Evaluation methods



Stages of the research process

Let us consider each of the stages of the research process, shown in Figure 2, in more detail.

A. Evaluating the existing Travel Record

This stage involved evaluating the (a) layout and (b) content of the existing NTS Travel Record. The existing Travel Record was evaluated using **two** main methods:

- **quantitatively**, through secondary analysis of Travel Record survey data; and
- **qualitatively**, mainly via workshops with NTS interviewers and cognitive interviews with respondents.

Below a little more detail is given on each of the evaluation methods used. The activities took place sequentially, in the order shown below, with findings from earlier activities feeding into later ones.

Quantitative methods

A1 - Secondary analysis of NTS Travel Record survey data

This analysis identified data items that had high item non-response and those that required a 'large' amount of editing which might indicate problems with the existing layout of the Travel Record. We also examined whether the Travel Record was less likely to be completed by certain population groups. The findings of this analysis fed into the formulation of research questions to be addressed by the interviewer workshops and cognitive testing described below.

Qualitative methods

A2 - Workshops with NTS interviewers

Three workshops were held with NTS interviewers who worked in different geographical locations and had varying levels of experience of working on the survey. A topic guide was devised prior to the workshops. Interviewers were asked to discuss the kinds of errors respondents made when completing the Travel Record; any reformatting of information they (the interviewers) did (i.e. breaking down composite journeys into single ones, adding in missed journeys or adding in additional information about journeys) before returning the Records to Head Office; and suggestions for improving the Record.

A3 - Discussions with the NTS Travel Record editors

The NTS has a team of dedicated editors, based in-house, who check the returned Records, make any necessary 'corrections' and resolve discrepancies in accordance with an agreed set of procedures. Discussions with the editors were helpful in identifying (further) errors and problems that they had to deal with.

A4 - Cognitive interviews with respondents

Thirty one cognitive interviews were conducted with respondents who (retrospectively) completed two days of the Travel Record[8]. Respondents were asked to 'think out loud' as they filled in the Travel Record, this allowed overt problems to be identified and internal debates the respondent had to be aired. Once the respondent had filled in the Travel Record the interviewer followed up with some retrospective probes. These interviews were systematically analysed using a content analysis approach. The evidence generated from these interviews helped to uncover, and to understand how respondents went about filling in the Record and whether the **format** of the existing Record caused them any problems. The findings of this Stage One evaluation of the existing travel record are detailed in a published report[9].

B. Re-designing the Travel Record

In this section I set out the approach we took to re-designing the Travel Record. There were two underpinning elements to the re-design: (a) existing literature in the form of a theoretical basis for how people approach filling in self-completion documents; and (b) empirical evidence obtained at Stage A.

How people approach filling in forms: Readers and Skimmers

First taking, existing literature, Jenkins et al[10] suggest that people fall into two categories when completing a form. 'Readers' are those that read through all or most of the material or instructions provided whereas 'Skimmers' read only as much as they think is required to complete the task. Skimmers employ 'shortcut heuristics' which are mental shortcuts or 'rules of thumb' that allow people to make inferences or decisions quickly and with reduced effort[11]. These heuristics have been used to explain how people draw conclusions in social settings but the same rule can be applied to a task such as filling in a Record. The desire to take a shortcut forms an important part of the cognitive process people go through when making decisions about each part of this Record.

Using Jenkins et al's 'Reader/Skimmer' analogy Readers are, on the whole, able to work through a badly designed questionnaire or form but Skimmers are more likely to make mistakes if they are unable to understand quickly, and easily, the information presented to them. If the task becomes difficult Skimmers may become frustrated with the effort required and so give up on the task altogether. We looked for evidence of these behaviours when analysing the Stage One cognitive interview data, classifying respondents as readers or skimmers. Our typology in fact

---

[8] Cognitive respondents completed (retrospectively) two days of the Travel Record, usually the two days prior to interview. It was not felt by the research team to be feasible to replicate the actual survey process of respondents completing the Record for seven days prospectively and then conducting the cognitive interview. This was because it was thought that respondents' recall of how they went about completing the record and the details of any problems they encountered would deteriorate over that time.

[9] **NTS Travel Record Review Stage 1** (2006) Alice MCGEE, Michelle GRAY and Debbie COLLINS Published by the Department for Transport, web only
http://www.dft.gov.uk/pgr/statistics/datatablespublications/personal/methodology/ntsrecords/ntstravelrecord1?version=1

[10] Jenkins, C.R., Ciochetto, S. and Davies, W. (1992) *'Results of cognitive research on the public school 1991-92 field test questionnaire for the schools and staffing survey'*. Unpublished, in Collins, D. and White, A. (1995) 'Making the next Census form more respondent-friendly' in Survey Methodology Bulletin July 1995, No 37, OPCS.

[11] Kahneman, D. and Tversky, A., (1973) '*On the Psychology of Prediction'* in Psychological Review Vol 80, No 4, American Psychological Association.

consisted of three categories: (1) Readers, (2) Skimmer 1s (who, on coming across difficulty or 'getting stuck', were prepared to look for help) and (3) Skimmer 2s (who gave up at the first hurdle and did not attempt to locate instructions).

Implications for re-designing the Travel Record

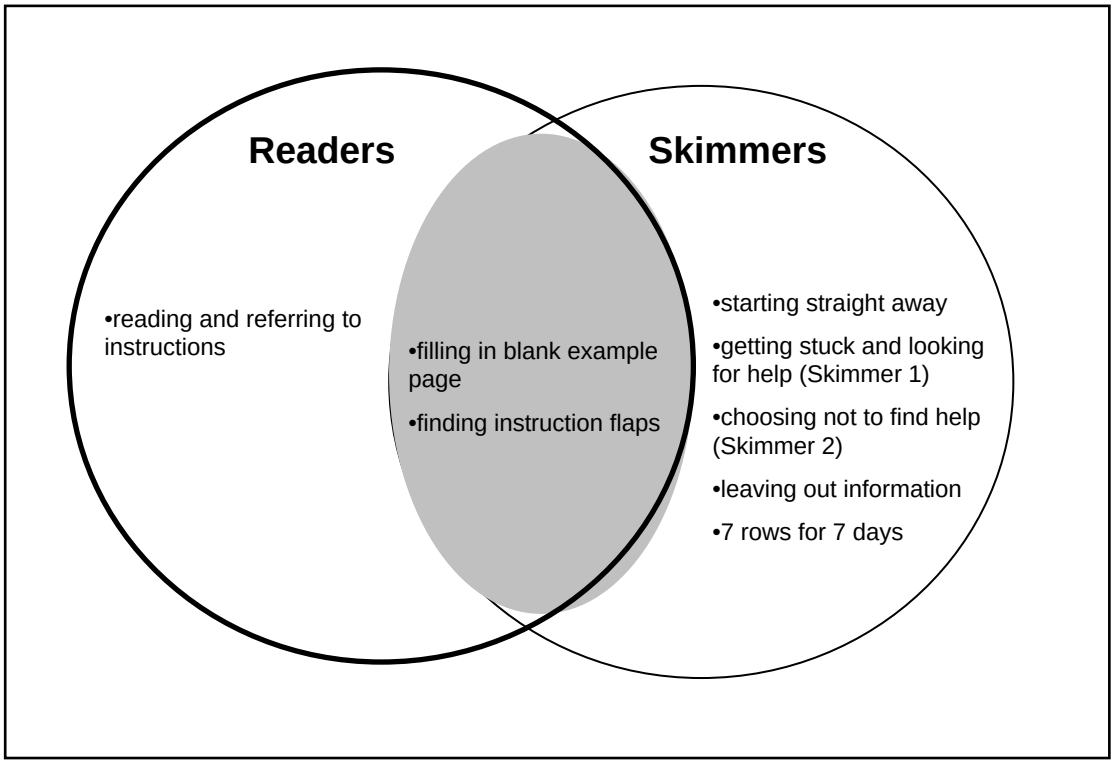
The ways respondents process information had implications for our re-design. Analysis of the Stage A cognitive interview data uncovered two types of problems respondents experienced. These were: (a) **comprehension** problems (that is, not understanding specific concepts within the Travel Record, for instance how to approach filling in the number of miles travelled at each stage of a journey); and (b) **presentational** problems (e.g. not being able to find instructions and finding the headings at the top of columns confusing (see figure 1)).

Applying our 'Reader/Skimmer' typology to the cognitive interview data from Stage A indicated that:

- Readers experienced a narrower range of problems when completing the Record than Skimmers and that these were principally **comprehension** problems.
- Skimmers, in contrast, experienced a wider range of problems, including **presentational** problems

The figure below outlines the 'Reader/Skimmer' typology and indicates the main strategies respondents employed in completing the Record and which group or groups of people adopted each one.

Figure 3 Strategies for filling in the Travel Record



In terms of finding solutions to these problems, there were, in the view of the research team, some fairly straightforward solutions to many of the problems experienced by Readers and those Skimmers who looked for help upon coming across difficulty or 'getting stuck' in trying to complete the Record. The real challenge for the re-design was going to be in finding ways to encourage Skimmers who gave up at the first hurdle to complete the Record.

Principles of good questionnaire design

Jenkins and Dillman[12] put forward a model attempting to overcome some of these problems, the principles of which we have used in our re-design. The model consists of two components: (1) encouraging respondents to follow a prescribed path through the questionnaire or designing "navigational guides", and (2) taking care with language when designing questions and answers (or in our case column headings) and the sequence in which respondents are expected to process them, termed as "good information organisation".

The re-design aimed to make the task of completing the Travel Record an easier and more straightforward one for all respondents, both Readers and Skimmers alike, taking into account the varying difficulties faced by each group. We used Jenkins and Dillmans' model to produce two overarching aims from our own re-design of the Travel Record. In general we attempted to help respondents:

- find their way around the Record by providing navigational guides; and
- understand the task by improving the way in which information was organised and displayed.

These aims focused on resolving the range of **presentational** problems uncovered at Stage A. The re-design did not aim to address some of the problems related to **comprehension** (the more conceptual problems) identified at Stage A because some of our initial recommendations were considered too radical in that they had wider implications for the comparability of survey findings over time. These issues would need to be carefully considered before more substantial changes to the Record could be made, and as such could not be addressed within the timescale for this project.

Main changes implemented following Stage One

Table 1 shows the main changes that were recommended and implemented to the design of the Travel Record following Stage A. It should be noted that we did not recommend the adoption of a columnar format. The findings from Stage A highlighted that the problems respondents had in filling in the Record were related to specific items within the Record and the level of detailed information required rather than the format of the Record. It was therefore decided to retain the existing row approach rather than moving to a columnar one.

[12] Jenkins, C.R. and Dillman, D.A. (1995) *'Towards a theory of self-administered questionnaire design'* in Lyberg et al (1997) *'Survey Measurement and Process Quality'* Wiley.

Table 1 Main changes to Travel Record Implemented following Stage One[13]

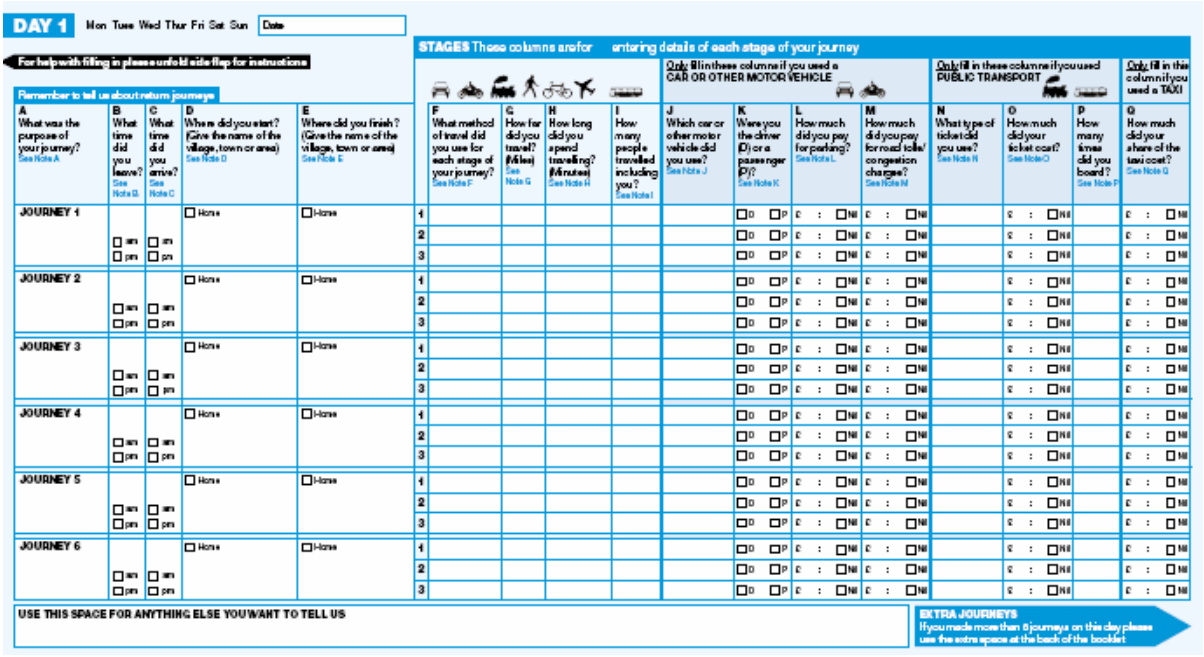
Front cover • ***Colouring and layout more user-friendly***	*Example page* • ***Example hand written instead of in type font*** • Example journey not based in London as respondents could feel it did not apply to them
Instructions • ***Instructions brought together so easier to find and in one place (on one flap at the front of the Record)*** • ***Visual guides to lead the way to instructions introduced (e.g. arrows)*** • Clearer signposting to instructions (e.g. 'See Note A' as opposed to simply '(A)'. • Pictorial images introduced to clarify which forms of transport to include (these were intended to help respondents understand the task more quickly and easily than if using large blocks of text) • Instruction wording made clearer and more concise	*Main diary* • Tick boxes introduced (as opposed to writing in) to make task quicker and easier (e.g. tick a box at column D if a journey started at Home or tick D or P for Driver or Passenger at column K) • Column headings posed as questions to help respondents better understand what they are being asked to provide (e.g. 'Purpose of journey' now reads 'What was the purpose of your journey?' and 'Method of travel' is now asked as 'What method of travel did you use for each stage of your journey'.[14]

The 'new' Record's design sought to address the problems found with the existing record, applying Jenkins and Dillman's principles for good questionnaire design. Figure 4 illustrates the new design for the recording page that was tested at Stage C.

[13] The Stage One report contains more detailed information on the recommendations made following this stage.
NTS Travel Record Review Stage 1 (2006) Alice MCGEE, Michelle GRAY and Debbie COLLINS Published by the Department for Transport, web only
http://www.dft.gov.uk/pgr/statistics/datatablespublications/personal/methodology/ntsrecords/ntstravelrecord1?version=1

[14] This was probably our strongest change – Stage A showed that both types of skimmers would head directly to the place where they were first required to enter information. This meant the column headings needed to be clear enough to understand on their own, if the respondent was unwilling to look anywhere else for instruction).

Figure 4 Re-designed Travel Record recording page



C. Evaluating the new Record

The aims for evaluating the new Travel Record were two fold: (a) to find out how respondents went about filling it in and the problems they experienced using a variety of evaluation methods; and (b) to discover whether there was any evidence to indicate that the quality of the information collected had improved.

The newly designed Travel Record was evaluated using **two** main methods:

- **quantitatively**, a pilot test of 100 respondents was conducted using the re-designed Record and the extent of errors made by respondents during this test were compared against a sample of 100 existing Records. Data from respondent debriefing questions asked as part of the pilot interview were also analysed; and
- **qualitatively**, principally using cognitive interviewing methods, but also incorporating interviewer and editor debriefing comments.

Below we discuss the evaluation methods used during this stage in a little more detail.

Quantitative methods

C1 - Review of pilot data

A systematic review **of respondent errors** was conducted, using a specially developed code frame (or checklist). The checklist was systematically applied to a sample of 100 current Records and to 100 of the new Records completed by pilot respondents. A separate checklist was used for each Record. The checklist was used to record different types of respondent error. Primarily these errors related to respondents providing insufficient information, or not providing any information

at particular columns. Problems identified were systematically recorded using a bar-gate system, allowing the total number of occurrences of a particular type of error to be quantified. Levels of respondent error were then compared between the pilot Records and sample of existing Records.

C2 - Respondent debriefing questionnaires

The pilot interviewers administered a debriefing questionnaire to all respondents following the seven-day record keeping period. The questionnaire asked respondents how easy or difficult they found the travel record overall, whether they read or referred to the example page and instructions, how easy these were to find and how helpful they were, and whether there were any parts of the record that were difficult or confusing. These data were keyed and systematic analysis conducted.

C3 - Review of the use of the blank example page

A systematic **review of the use the blank example page** contained in the existing Travel Record was undertaken. This was done to help inform the decision about whether the page should be retained, by assessing how often it was completed[15].

Qualitative methods

C4 - Cognitive interviews

Thirty-two cognitive interviews were conducted using the same format as at Stage A, to allow a comparison between the two round of interviewing. Again, these interviews were systematically analysed using a content analysis approach.

C5 - Pilot interviewer comments

The ten pilot interviewers were contacted by telephone following the pilot test and asked a series of questions about how successful the new diary had been in the field and whether they had any suggestions for how it could be improved. These telephone interviews were written up into a short report.

C6 - Comments from the NTS data editors

The same team of dedicated data editors were asked to examine a selection of pilot Travel Records, provide feedback on how they thought the re-designed Record had worked during this test, and give any further recommendations for final revisions.

[15] Interviewers are encouraged to work through a practical example with respondents when placing the Travel Records.

Hypotheses

Three hypotheses were tested during the analysis of Stage C. These were:

1. The cognitive interview data would identify a narrower range of problems with the 'new' Travel Record than the existing one.
2. The cognitive interview data would show problems related to information organisation and navigation to be diminished in the new Record but that the more conceptual problems would still remain.
3. The systematic review of respondent error undertaken would show the new Record to have lower error rates (particularly for navigational elements) than the existing Record.

In addressing these hypotheses all of the different sources of evidence outlined in this paper were reviewed and triangulated.

Findings

The findings from this evaluation proved all three of our hypotheses to be correct. Triangulating the different types of evidence showed the new Record to be an improvement over the existing one. The cognitive interview findings showed the task of completing the Record to be better understood. Those who were prepared to look for help with filling in the Record were able to find it more easily (e.g. locating the notes on the instruction flap). Those who were not prepared to consult notes found the task more straightforward as the information at the specific places they recorded information was better organised and displayed. Supporting this further, the Record was piloted among 100 respondents and yielded data of a higher quality than a comparison sample of existing Records. Further details on the re-design (Stage B) and findings from Stage C are published on the DfT website.[16]

Following the evaluation a few small changes were made to the Record, shown in Table 2 below, and it was strongly recommended that this version of the newly designed Record was implemented in the 2007 survey. This recommendation was taken forward.

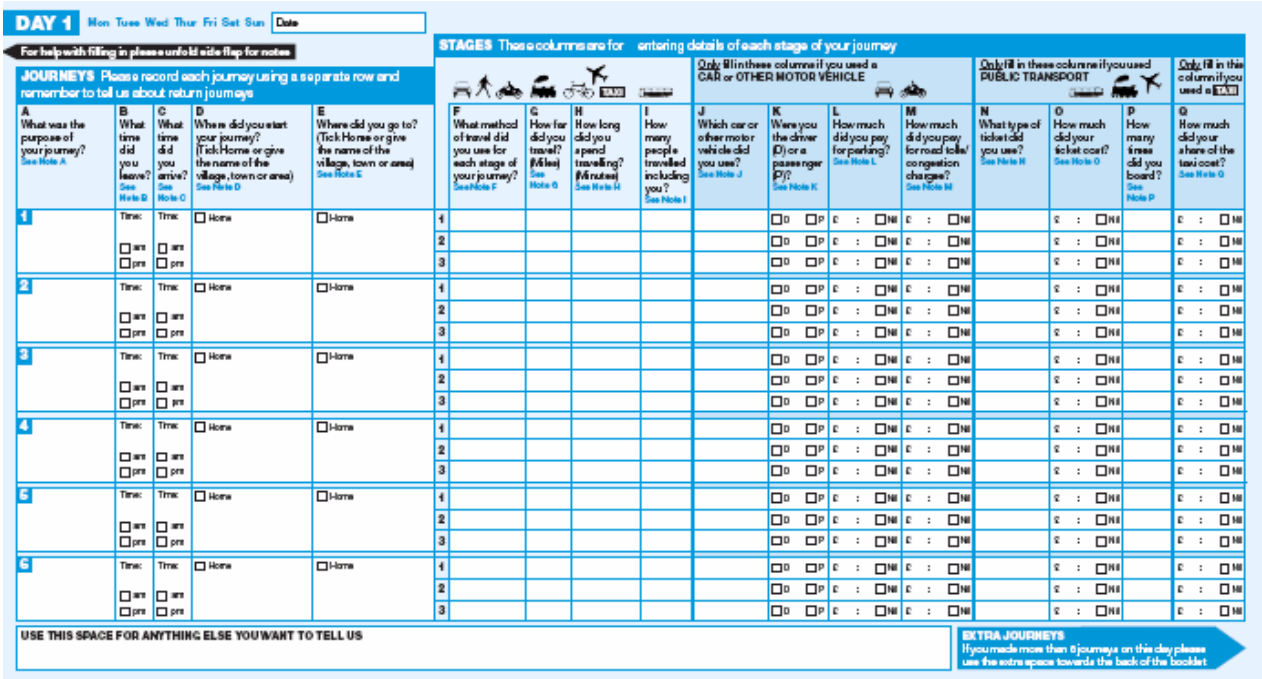
Table 2 Main changes implemented following Stage C

- Colour scheme reversed for example pages (grey based) to make clear where examples are used;
- Subheading box introduced at the top of A-E to make the concept of a journey clearer;
- Journey numbering placed in a separate box at the start of each row;
- 'Time:' added at columns B and C to encourage respondents to enter a figure as well as ticking an am/pm box;
- Taxi symbol added to the row of transport images.

The final Travel Record recording page is shown in Figure 5 below.

[16] **NTS Travel Record Review Stage 2** (2006) Alice MCGEE, Michelle GRAY, Fiona ANDREWS, Robin LEGARD, Natasha WOOD and Debbie COLLINS Published by the Department for Transport, web only http://www.dft.gov.uk/pgr/statistics/datatablespublications/personal/methodology/ntsrecords/ntstravelrecord2?version=1

Figure 5 Final Travel Record recording page



Conclusions

This paper has shown how multiple evaluation methods were brought together to evaluate whether a re-designed survey instrument was 'better' than the existing one. The mixture of both quantitative and qualitative evaluation methods complemented and supported each other in providing a strong evidence base to clearly demonstrate that the changes made were indeed 'improvements' and that the new Travel Record should indeed replace the existing one.

Using Behavior Coding to Evaluate the Effectiveness of Dependent Interviewing

Joanne Pascale, US Census Bureau and
Alice McGee, National Centre for Social Research

Abstract

Dependent interviewing (DI) is used in many longitudinal surveys to "feed forward" data from one wave to the next. Though it is a promising technique which has been demonstrated to enhance data quality in certain respects, relatively little is known about how it is actually administered in the field. This research seeks to address this issue through behavior coding. Various styles of DI were employed in the English Longitudinal Study on Aging (ELSA) in January, 2006, and recordings were made of pilot field interviews. These recordings were analysed to determine whether the questions (particularly the DI aspects) were administered appropriately and to explore the respondent's reaction to the fed-forward data. Of particular interest was whether respondents confirmed or challenged the previously-reported information, whether the prior wave data came into play when respondents were providing their current-wave answers, and how any discrepancies were negotiated by the interviewer and respondent. Also of interest was to examine the effectiveness of various styles of DI. For example, in some cases the prior wave data was brought forward and respondents were asked to explicitly confirm it; in other cases the previous data was read and respondents were asked if the situation was still the same. Results indicate varying levels of compliance in terms of initial question-reading, and suggest that some styles of DI may be more effective than others.

Keywords: Dependent interviewing, longitudinal surveys, panel surveys, behavior coding

1 Introduction[17]

In recent years there has been increased interest in and use of "dependent interviewing" (or DI) in longitudinal surveys. DI (also known as "previously reported data" or PRD) is a technique whereby data collected from one wave are carried forward into the next wave in order to tailor question wording and skip patterns. For example, if at Wave 1 a respondent reported working for Employer X, one version of a Wave 2 DI question would read: "Last time you said you worked for Employer X. Are you still working for Employer X?" This is in contrast to an "independent" (that is, non-DI) method whereby at Wave 2 the respondent would simply be asked "from scratch" for the name of the employer. A related implementation of DI is to route respondents around detailed questions if the "state" from one wave to another has not changed. For example, a detailed set of questions about Employer X may be asked in Wave 1, and if at Wave 2 the respondent reports they are still working for the same employer, those details need to be collected a second time.

[17]This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on methodological issues are those of the authors and not necessarily those of the US Census Bureau or the National Centre for Social Research.

The proliferation of automated surveys has contributed to the increased interest in DI, since the technique can be difficult and cumbersome to implement in a paper/pencil questionnaire. Another factor contributing to the interest in DI is its potential to enhance data quality in a number of ways. Generally DI can make for a smoother, smarter, more efficient interview by reminding respondents of their previous answers and allowing them to simply report whether anything has changed since then. Rigorous research evidence "proving in" this potential is beginning to emerge. For example, there is consistent evidence that DI reduces spurious change, particularly in employment characteristics (Polivka and Rothgeb, 1993; Jackle and Lynn, 2004), and there is strong evidence that it significantly reduces (though does not eliminate) seam bias (Moore et al, 2006). And there is qualitative evidence that respondents want and expect DI (Pascale and Mayer, 2004). In the summer of 2006 a major conference was organized to assess the Astate of the art@ of DI and several papers demonstrated specific benefits of DI. An edited monograph book of selected papers is to be published by John Wiley and Sons in 2008 (http://www.iser.essex.ac.uk/ulsc/mols2006/).

What the literature seems to lack up to now, however, is evidence of how DI is actually implemented in the field. The current research set out to address this gap. In particular we used behavior coding to examine whether interviewers read questions as worded, focusing especially on the dependent words and phrases within the questions, and we examine respondents' reactions to the dependent phrases - that is, whether they affirm or dispute the previously-collected data, and whether providing this information seems to help or hinder the reporting task. Finally we examine whether these behaviors seem to vary at all by "style" of DI - that is, the particular way that the previously-collected information is fed back to the respondent. The vehicle we use for this research is the English Longitudinal Study on Aging (ELSA), carried out by the National Centre for Social Research (called "NatCen") in collaboration with University College London and the Institute of Fiscal Studies.

2 Methods

2.1 ELSA: the Survey Vehicle

ELSA is a study of people aged 50 and over and their younger partners. The study explores the dynamics of health and disability, family structure, public program participation, economic circumstances, and retirement. The first wave was administered in 2002 with 12,100 respondents, and follow-up interviews have been conducted every two years to measure changes in health, social and economic circumstances. Dependent interviewing was embedded in the Wave 2 instrument but due to budget and schedule constraints, little evaluation was done prior to its implementation. Analysis of Wave 2 data, however, raised some concerns about the effect of DI. For example, roughly 20% of respondents who reported high blood pressure at Wave 1 claimed they no longer had the condition at Wave 2. Due in part to this finding, the current research project was undertaken to generally assess the implementation of DI techniques in the field. Behavior coding was chosen as the evaluation method in order to carefully assess interviewer-respondent interactions, and to measure the extent to which the questions were being administered as written.

2.2 Field Interviewing and Recording

The pilot phase of Wave 3 ELSA was conducted over a 4 week period in January, 2006, with 17 NatCen field interviewers from different areas around the United Kingdom. All interviews were conducted face-to-face using a computer-assisted paper instrument (CAPI). The questionnaire included questions on a number of topics: household and individual demographics, health status, income and assets. A total of 121 individuals participated in this pilot; 101 had participated in previous waves, while 20 were added as part of a refresher sample.

Interviews were recorded using Computer Audio Recorded Interviewing (CARI), a software application that allows field interviews to be recorded directly onto computer laptops as digital sound files. A consent question asking respondents for permission to record the interview was embedded into the beginning of the questionnaire, and if respondents did not consent the recorder was not switched on. In total 13 of the 121 individual interviews (11%) were not recorded due to a lack of consent, and another 4 cases were unusable due to corrupted sound files, with a net result of 104 productive recorded individual interviews. Thirty four of the individual interviews took place in single-person households and the other 70 took place in multi-person households.

2.3 Dependent Interviewing Question Wording

Dependent interviewing was embedded in the instrument across three different topic areas: demographics, health conditions and vehicle ownership (see Figure 1). In the health condition section there were three broad categories of illnesses: eye, cardiovascular conditions, and chronic conditions. Within each of these broad categories there were multiple specific illnesses asked about. For example under eye conditions there were four illnesses (such as glaucoma and cataracts). Items 4 and 5 were repeated for each illness or condition the respondent had reported in the prior wave.

Figure 1: Question Wording of Items Using Dependent Interviewing

A. Demographics

1. Does NAME still live here?
2. Can I just check, is NAME=s date of birth DOB?
3. Our records show that when we last interviewed you, you had a child called NAME, whose date of birth is DOB. Are these details correct?

B. Health Conditions

4. Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [EYE/CVD/CHRONIC CONDITION].
5. Do you still have [EYE/CVD/CHRONIC CONDITION]?

C. Vehicle Ownership

6.	Last time we saw you, you told us that you were the main user of a [MAKE OF VEHICLE], with a [LETTER] registration. Do you still have that vehicle?

Five different styles of DI were used across these three topic areas, but as was mentioned earlier, no particular research guided those design decisions. As Figure 1 indicates, each of the six items employed a slightly different style of DI. The first two items in the demographics section do provide previously-reported data but do not explicitly mention having gathered this data in the previous interview. Rather, the past data is simply presented and the respondent is asked to verify it. The third item explicitly states that the data was collected last time and the respondent is asked if the information is correct. Unlike the demographics questions, the health questions were separated into two distinct items. The first (4) was simply a statement, informing the respondent of a particular illness they reported during the previous interview. It was meant to be read as a statement and the respondent was not asked or expected to provide a response to this statement. The second item (5) then asked whether the respondent still had the illness or condition. And finally for vehicle ownership the routine was somewhat similar to the health conditions questions; first a statement was read that informed the respondent of what they reported last time, and then a question was asked to determine if that condition still existed (that is: do you still own the vehicle?). The difference was that for the vehicle item the statement on the past condition and the question ("still have it") were wrapped into one single item, while in the health section there were two distinct items -- the statement and then the "still?" question.

2.4 Behavior Coding

In order to develop the code frame for behavior coding, we first listened to several tapes to get a general feel for the flow of the interview, the frequency and nature of non-standard interviewer behavior, and respondent's reactions to the questions. We determined that the "first exchange" - the interviewer's initial reading of the question and the respondent's first utterance in response to that B were sufficiently rich for analysis and thus developed a code frame to capture only these behaviors, as well as a final outcome. Within these three behaviors (interviewer's initial question-reading, respondent's initial response, and outcome), we started out with a fairly standard code frame and adapted it based on the content of the tapes and our particular interest in learning about the functioning of the feed-forward phrases embedded within the questions (see Figure 2). For interviewer behavior we used three main code categories: (1) question was read as worded or with only a minor change that did not change the meaning of the question (2) question was read with a "major change" that changed or could change the meaning of the question and (3) the question was omitted. Within the major change code we developed two DI-specific codes. On the tapes it was rather common to hear interviewers changing a statement into a question. For example, in the health section the statement: "Our records show that last time you reported X condition" became a question because interviewers often added "Is that right?" or used an intonation and a pause to turn the statement into a question. In other cases a question became a statement. For example in the demographics section the question "Does NAME still live here?" was modified to "And NAME still lives here." Since these were the most frequently-observed problems we created dedicated codes for them.

Figure 2: Behavior Codes

A. Interviewer Codes

S: Standard; read as worded or with a minor change that did not change the meaning

MC1: Fed-forward statement was read as a question (e.g.: "Last time you told us you had high blood pressure. Is that correct?")

MC2: Fed-forward question was read as a statement (e.g.: "And your date of birth was 25th May 1933.")

MC3: Any other change that did or could change the meaning of the question

O: Omission

I/O: Recording was inaudible or the behavior does not fit into one of the above codes

B. Respondent Codes

AA: Adequate; acknowledged or did not dispute the fed-forward data

AD: Adequate; disputed or challenged the fed-forward data

CL: Request for clarification

IA: Inadequate answer or elaboration

DK: Don't know

R: Refused

I/O: Recording was inaudible or the behavior does not fit into one of the above codes

C. Outcome Codes

AA: Adequate; final response fit one of the given response categories

IA: Inadequate; final response did not fit any of the response categories

DK: Don't know

R: Refused

I/O: Recording was inaudible or the behavior does not fit into one of the above codes

Respondent codes were fairly standard, again with the exception of DI-specific codes. An "adequate" code meant that the respondent's initial utterance fit one of the response categories. We adapted this code to capture whether respondents affirmed or disputed the fed-forward data. There were also codes for a request for clarification and a rereading of the question, and a general "inadequate" code, meaning the respondent's answer did not fit any of the given response categories. Outcome codes were simply "adequate" and "inadequate".

3 Results

Findings will first be presented for each topic area, then themes across topics will be discussed. Regarding the outcome code, adequate answers were obtained in the vast majority of cases (upwards of 90% of the time) and there was little variation across items so those results are not shown.

3.1 Demographic Items

In the demographics section, first regarding interviewer behavior, there was fairly wide variation in the extent to which interviewers adhered to standardized technique, ranging from 40-79%, depending on the item (See Table 1A).

Table 1A: Interviewer Behavior for Demographic Items

Item	Base (n)	Interviewer Behavior Code (in percent)				
		Read as worded	Q read as statement	Other major change	Omitted	Other
LIVE: Does NAME still live here?	120	40	33	4	18	5
DOB: Can I just check, is NAME's date of birth DOB?	107	57	37	1	1	4
CHILD: Our records show that when we last interviewed you, you had a child called NAME, whose date of birth is DOB. Are these details correct?	84	79	8	11	0	2

As was predicted from our earlier (unsystematic) listening of the tapes, for the most part when interviewers diverged from the script they turned the question into a statement (e.g.: "Is NAME's date of birth January 1?" would become "And NAME's date of birth is January 1."). This behavior occurred 33-37% of the time for the first two items (LIVE and DOB) and only 8% of the time for the last item (CHILD). This may not be too surprising considering the nature of the items. Answers to the first two items may seem obvious - particularly at Wave 3 - and interviewers may have been somewhat reluctant to ask a question with an obvious answer. Indeed LIVE was omitted altogether 18% of the time, and this could be because the interviewer was talking to the person referenced in the question. The third item, on the other hand, asks about someone else in the household (a child), the information is rather specific (name and date of birth) and the actual question ("are these details correct?") may not seem to have an obvious answer. That is, it may seem like a more "legitimate" question to ask than asking a person, in what appears to be their home, "Do you still live here?" This could explain why this last item was read as worded so frequently - 79% of the time.

Turning to respondent behavior, on the whole respondents provided a codeable answer straightaway more than 80% of the time (see Table 1B). They rarely disputed the fed-forward data (up to only 5% of the time), and most of these disputes stemmed from keying errors in the name or date of birth previously recorded.

Table 1B: Respondent Behavior for Demographic Items

Item	Base (n)	Respondent Behavior Code (in percent)				
		Adequate; affirmed FF	Adequate; disputed FF	Clarifica-tion	Inade-quate	Other
LIVE: Does NAME still live here?	91	81	1	1	4	12
DOB: Can I just check, is NAME's date of birth DOB?	102	91	0	0	0	9
CHILD: Our records show that when we last interviewed you, you had a child called NAME, whose date of birth is DOB. Are these details correct?	84	89	5	1	1	4

3.2 Health Items

As noted above, the health questions were asked in two parts. First a statement about the condition reported during the prior wave was read, and then a question was asked to determine whether the condition still existed. Overall levels of "exact reading" of these items were moderate - ranging from 41-76% but generally in the low 60s (see Table 2A). When interviewers diverged from the script they tended to turn the statement into a question (20-38% of the time) by adding something along the lines of "Is that correct?" to the end of the statement. Interviewers would then often omit the actual question "Do you still have it?" altogether - 13-18% of the time. The implications are important here, because it means the respondent is getting a fundamentally different question, specifically "Is it correct that you reported this condition last time?"versus "Do you still have this condition now?"

Another problem was when the actual question "Do you still [have condition X]?" was read, interviewers often read it as a statement rather than a question: "And you still have it." - 3-16% of the time. This has serious implications for data quality as well, because the respondent is not being given the opportunity to think about whether they really do still have the condition; they are just being told they do.

Table 2A: Interviewer Behavior for Health Items

Item	Base (n)	Interviewer Behavior Code (in percent)					
		Read as worded	Q read as statement	Statement read as Q	Other major change	Omitted	Other
LAST-EYE: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	21	62	na	38	0	0	0
STILL-EYE: Do you still have [condition]	19	63	16	na	5	16	0
LAST-CVD: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	100	63	na	20	17	0	0
STILL-CVD: Do you still have [condition]	79	76	3	na	8	13	1
LAST-CHRON: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	59	41	na	34	17	5	3
STILL-CHRON: Do you still have [condition]	51	61	14	na	2	18	6

Regarding respondent behavior, there were fairly high levels of adequate behavior - over 90% for both CVD and chronic conditions - and 72% for eye conditions (however the base here was only 21 cases). Respondents disputed prior wave data for a variety of reasons. Some said they used to have the condition but no longer do, and this is essentially how the questionnaire was expected to operate. But in other cases the fed-forward data were problematic; respondents either denied that they'd reported the condition at the prior wave, or they disagreed with the characterization of the illness. For example, in one case an illness was recorded as cancer in the prior wave and when asked about it in the next wave the respondent said it wasn't cancer. He wasn't sure what the diagnosis was but said it was not cancer. In another case a respondent reported memory impairment at the prior wave but this particular condition was grouped in with other related illnesses in the instrument ("dementia, senility or memory impairment"). When the DI question appeared on the screen the interviewer only read "dementia" and the respondent refuted it. Only when the respondent went back and read the full question, with all three conditions, did the respondent affirm that he had a memory impairment.

Table 2B: Respondent Behavior for Health Items

Item	Base (n)	Respondent Behavior Code (in percent)					
		Adequate; affirmed FF	Adequate; disputed FF	Ade-quate*	Inade-quate	Clarifica-tion	Other
LAST-EYE: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	21	62	10	[72]	10	5	15
STILL-EYE: Do you still have [condition]	19	na	na	94	0	0	6
LAST-CVD: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	100	87	5	[93]	4	0	4
STILL-CVD: Do you still have [condition]	80	na	na	69	24	0	8
LAST-CHRON: Our records show that when we last interviewed you in January 2004, you said you had had (or been told by a doctor that you had had) [condition]	53	85	4	[89]	2	0	9
STILL-CHRON: Do you still have [condition]	35	na	na	89	6	0	6

* For "LAST-XX" items this column shows the sum of "Adequate; affirmed FF" and "Adequate; disputed FF"

3.3 Vehicle Item

The vehicle item was similar to the health items - first providing a statement about what was recorded in the prior wave and then asking a question about whether the situation is still the same. A key difference, however, was that rather than presenting the statement and question as two distinct items on two different screens, they were rolled into one item. Across all items in the questionnaire the vehicle item had the highest level of interviewers reading the question as worded at 82% (see Table 3A). The problems identified in the health section - interviewers turning the statement into a question, or the question into a statement, or omitting the question -- did not turn up very often here, perhaps because the style of DI was different. Specifically, interviewers did not have to have read a statement about the prior report but could move directly into the question: "Do you still have this vehicle?" By not displaying the statement on the prior wave data as a distinct item, interviewers may have been less tempted to turn that statement into a question by asking, for example, "Is it correct that you reported this vehicle last time?" The result was that the intended question - whether the vehicle was still owned - was being asked, rather than an unintended question ("Did you report owning this vehicle last time?"). However, among

the non-standard behaviors there were still several instances of interviewers (8% of the time) turning the question into a statement: "And you still own xx vehicle." This could be a result of interviewers having seen the vehicle in question on their way to the doorstep.

Table 3A: Interviewer Behavior for Vehicle Item

Item	Base (n)	Interviewer Behavior Code (in percent)				
		Read as worded	Q read as statement	Other major change	Omitted	Other
VEHICLE: Last time we saw you, you told us that you were the main user of a [MAKE OF VEHICLE], with a [LETTER] registration. Do you still have that vehicle?	51	82	8	2	4	4

Respondent behavior here was similar to the health section. Respondents provided a codeable answer straightaway 79% of the time (see Table 3B). They rarely disputed the fed-forward data (6% of the time), and most of these disputes stemmed from keying errors in the fed-forward registration information.

Table 3B: Respondent Behavior for Vehicle Item

Item	Base (n)	Respondent Behavior Code (in percent)				
		Adequate; affirmed FF	Adequate; disputed FF	Inade-quate	Clarifica -tion	Other
VEHICLE: Last time we saw you, you told us that you were the main user of a [MAKE OF VEHICLE], with a [LETTER] registration. Do you still have that vehicle?	51	74	6	10	4	6

4 Summary and Recommendations

The extent to which interviewers adhered to the standardized script varied quite a bit - questions were read as worded 40-82% of the time, depending on the particular item. When interviewers diverged from the script, the way they changed the wording varied by topic area and style of DI which, unfortunately, were confounded because each item had a unique style of DI. In the demographics and vehicle items, for the most part interviewers changed the question into a statement ("Does NAME still live here?" became "And NAME still lives here.") In the health section interviewers read statements about what was reported last time as questions. Rather than simply reading the statement "Last time you reported X condition" interviewers would add "Is that correct?" (which is an ambiguous question) and often omit the question "Do you still have

condition x?" The result was that often the intended question - to determine whether the condition still exists -- was obscured or omitted.

For the most part respondents provided codeable answers on the first exchange 72-94% of the time. It was fairly uncommon for respondents to dispute the fed-forward data (0-10% of the time) but when they did it was for a variety of reasons. Some confirmed the prior wave report but said they no longer have it. Some denied the prior wave report, and some disagreed with the details of the fed-forward data. Note that this first scenario is what we expect to happen in the instrument so it is actually a misnomer to say the respondent "disputed" the earlier report. Respondents here are not disputing what they said earlier, but rather they are confirming their earlier report and then reporting change. However, when the code frame was developed we heard very few instances of respondents disputing the prior data at all; the majority of cases were respondents simply agreeing with the fed-forward data. We therefore failed to recognize that it would have been valuable to create separate codes for agreeing to the fed-forward data and reporting real change versus actually disputing the prior report. Even with the full dataset, however, the frequency with which respondents did not simply agree to the prior wave data was too low for a rich analysis, and a larger dataset would be needed to address this issue.

In terms of recommendations these findings strongly suggest that questionnaire designers should avoid providing statements of prior wave data without an actual question, because interviewers are too tempted to turn these statements into questions, which obscures the question on whether the prior wave situation still exists. If it is important to confirm or verify information reported in a prior wave, this should be done explicitly by adding wording like "Is that correct?" and then mapping out the proper paths to follow if the earlier report was recorded in error.

Our findings from the health conditions section suggest that for certain topic areas it is important to feed back prior wave data in the respondent's own words as much as possible. When respondents' descriptions of their illnesses were obscured by either the instrument or the interviewer grouping the illness with other conditions, respondents no longer recognized the illness they originally reported.

Finally our findings suggest a more general recommendation that the style of DI should be carefully tailored depending on the particular item. For example, for topics unlikely to change from one wave to the next (such as DOB), avoid re-asking questions because interviewers often read them as statements or omit them altogether. For these topics it may be more effective to either explicitly verify the accuracy of the earlier report (as suggested above), or to avoid bringing back the information at all. A hybrid-type approach for a study with several waves would be to verify the accuracy of previously-recorded data in wave 2 and then accept the data as correct and avoid re-affirming it in all later waves.

REFERENCES

Jäckle, A. and Lynn, P. (2004), "Dependent Interviewing and Seam Effects in Work History Data," Working Paper #2005-24 of the Institute for Social and Economic Research, Colchester, UK: University of Essex.

Moore, Jeffrey C., Nancy Bates, Joanne Pascale and Anieken Okon (2006), "Tackling Seam Bias Through Questionnaire Design," Invited paper presented at the conference on Methodology on Longitudinal Surveys, July 12-14, Essex, United Kingdom.

Pascale, J. and T. Mayer (2004). "Alternative Methods for Exploring Confidentiality Issues Related to Dependent Interviewing." Journal of Official Statistics, Volume 20, Number 2, pp 357-377, 2004

Polivka, A. and Rothgeb, J. (1993), "Redesigning the CPS Questionnaire," Monthly Labor Review, September 1993, 10-28.

On Ethics and Integrity in Cognitive Interviewing Practice

Paul Beatty
National Center for Health Statistics

Assuming that proper procedures and confidentiality safeguards are in place, the act of asking questions to willing participants is generally benign. This applies both to large scale data collection and to the pre-testing and development activities that QUEST members take part in. However, recent cognitive interviewing experiences have suggested that it is worth considering a few additional safeguards to prevent (1) undue consequences to research participants, and (2) the potential for collaborators to influence findings when they have a particular stake in the outcome of the testing.

One of the few risks of cognitive interviewing is emotional, i.e., upsetting participants when the discussion touches on sensitive issues. This risk is generally seen as minimal, given that participants are usually aware of the topic in advance. However, while cognitive interviewing draws cues of interviewer detachment from survey interviewing practice, the actual depth and intimacy of discussion may have little resemblance to its survey model. Does this exacerbate discomfort? Perhaps more importantly, when testing draft questions that may prove too controversial for production surveys, does the model of interviewer detachment have the potential to influence attitudes or even behaviors in unforeseen ways?

In addition, there may be times when an outside party has a particular interest in demonstrating that a questionnaire does or does not perform well. If such an individual or institution collaborates in the cognitive interviewing evaluation, they could have several opportunities to influence findings. One such opportunity could occur if the collaborator is involved in identifying study participants. In this presentation I will discuss several situations where an outside party had the means and motive to tweak a questionnaire evaluation project to further an agenda. Such collaborations also have, on occasional, threatened our ability to carry out promises of confidentiality, as will also be discussed.

There is no evidence at this point that cognitive interviewing procedures at NCHS (or elsewhere) have actually led to harm or tainted findings. However, some of the situations described here may serve as a starting point for further discussions regarding what safeguards should be routinely adopted to prevent any future lapses.

"For various reasons, Paul Beatty was unable to disseminate the full paper in the proceedings volume. However, if you would like a copy of the presentation, please contact him directly at pbeatty@cdc.gov."

Analysing and interpreting cognitive interview data: A qualitative approach

Debbie Collins
Question Design and Testing Hub, National Centre for Social Research

The aim of cognitive interviewing is to provide evidence on whether the survey questions under scrutiny are meeting their measurement objectives. This evidence helps us in making decisions about whether and how to revise them. In analysing cognitive interview data we are attempting to unearth evidence of question performance in terms of the problems that the question structure, content and survey context may cause respondents. These data can be seen as being qualitative in nature, in so much as they are respondents' accounts of their thought processes, understanding of the survey response task presented, and the factors that shape their responses. Furthermore, the sampling designs commonly utilised in the cognitive testing of questions are purposive, driven by hypotheses about the sorts of respondent characteristics that are likely to affect responses, and numbers are small. They are not designed (usually) to be statistically representative. This paper is concerned with how we analyse cognitive interview data.

1. Current practice

A cursory review of the literature indicates that there is little written on the subject of how to analyse cognitive interview data. There have been a few papers describing the development and use of standardised coding schemes to analyse cognitive interview data, for example Presser and Blair (1994), Conrad et at (1999) and Willis et al (1999). These coding systems reflect the 'dominant' question and answer model, identifying problems with: comprehension (communication); recall or computation tasks; bias or sensitivity (judgement issues); and response categories. They may also include some elements of behaviour coding, such as identifying problems interviewers have in reading the question or recording the answer. In addition, they can cover what Willis et al (op cit) call 'logical issues' – that is items that cannot easily be conceptualised as due to problems in the cognitive processing chain (an 'other' category for problems that cannot be classified according to the question and answer model). Studies that have utilised such coding systems have tended to be those that have sought to evaluate the validity and reliability of cognitive interviewing methods (see for example, DeMaio & Landreth, 2004, Rothgeb et al 2001) rather than test survey questions per se. The coding scheme lends itself to the interview data being abstracted to a count or quasi-qualitative indication of the number of times a particular problem type was found is. This can be a time-consuming process, which may be a reason why such schemes are not in common usage among practitioners. Moreover the 'counting' of problems may not be that useful as an output to the practitioner who needs to understand why the problem occurred, in what circumstances, and therefore whether and how it can be rectified. As Miles and Huberman (1994) observe:

*"Just naming and classify what is out there is usually not enough. We need to understand the patterns, the recurrences, the **whys**. As Kaplan (1964) remarks, the bedrock of inquiry is the researcher's quest for 'repeatable regularities'."*

Many cognitive testing reports contain only a brief description of the analysis process (often just stating that the interview data were analysed but giving very little detail on how this was actually

done). This provides no real insight into how the findings were derived, which is problematic because the process is not replicable, transparent and open to scrutiny. Furthermore, to the researcher new to the world of cognitive interviewing there is little (with the exception to some extent of Willis, 2005) that is published that describes the analysis process in a way that could be replicated or followed. So how does one learn this skill? How can we be certain that when we compare cognitive interview findings for the same set of test questions across different organisations or research teams, that the data have been analysed in the same way?

Perhaps it is a reflection on the infancy of our discipline that little attention so far has been paid to how we analyse cognitive interview data. However, the credibility of our findings, and therefore our ability to convince funders of the value of cognitive testing, rests on the rigour of our methods. There is much we can learn from debates within the field of qualitative research in this regard.

2. Analysis as a component of quality

Over the past 20 years or so there has been a considerable debate in the qualitative research arena about the appropriateness and need for standards, guiding principles or evaluative criteria to help assess 'quality' in qualitative research (see for example, Spencer et al, 2003a, chapters 3,5 and 6 for a summary). The way in which the data are analysed and interpreted is seen as an important component of the research process, raising issues of:
validity (see Enderstvedt, 1989);
reliability in terms of consistency (LeCompte and Goetz, 1982) or '*auditability, dependability or reflexivity* '[18](see for example Lincoln and Guba, 1985); and
objectivity (see for example Patton, 2002).
Spencer et al (2003b) propose the following *'hallmarks'* of quality in any analysis process.

- Analysis remains grounded in the data (i.e. ideas and concepts emerge from the data rather than being imposed on it).
- The data reduction process is transparent (i.e. that there is a clear route back to the raw interview data).
- The process permits within and between case analysis. This facility will allow greater analytical power, assisting both thematic analysis and the identification of associations between phenomena.

We need to start to discuss these issues and develop guidelines for how analysis should be undertaken. In this paper I describe the methods used by NatCen's Question Design and Testing Hub to analyse cognitive interview data. These are routed in a qualitative approach to data analysis. First, I briefly discuss the principles of qualitative analysis and its value to survey methodologists concerned with question testing and development.

3. Qualitative approach to cognitive interview data analysis

As mentioned at the outset of this paper, it is my contention that cognitive interview data are intrinsically qualitative data: they are accounts of respondents' thought processes. As such it seems more appropriate to use qualitative analysis techniques than more quantitative ones, such

[18] This refers to whether the method is suitably well documented to allow others to replicate the process.

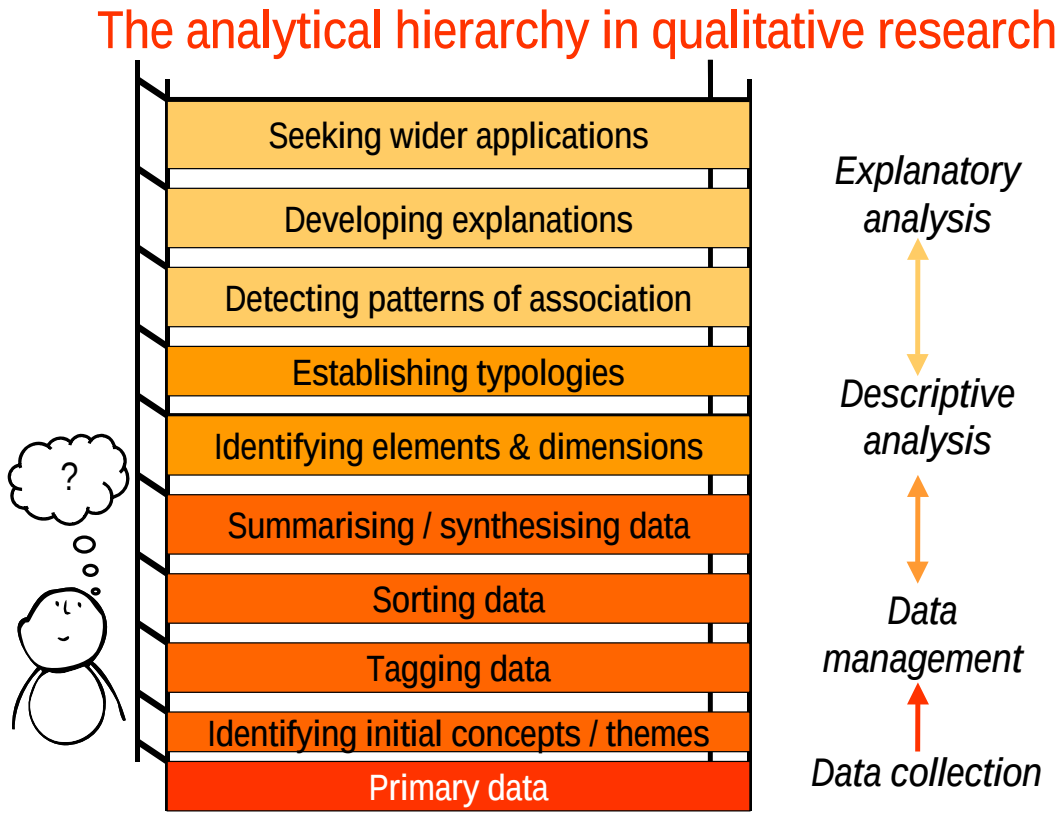
as standardised coding schemes. The question is, which one? There are many different analysis techniques that have been developed, reflecting the different epistemological assumptions made about the nature of qualitative investigation and the status of the researcher in it, and the aims and objectives of the traditions. Some of the main ones are summarised in Figure 1 below.

Figure 1 – Summary of main forms of qualitative data analysis

Traditional sociological/ethnographical approaches		
Ethnographic accounts	➢	detailed descriptions of cultures or organisations
Grounded theory	➢	generates analytical categories and the links between them through an iterative process of collecting and analysing data
Structural approaches		
Content analysis	➢	identifies content and context of documents, often involves counting (so not strictly qualitative)
Narrative analysis	➢	examines how a story is told and the intention of the teller
Conversation analysis	➢	examines the structure of (usually) naturally occurring conversations
Discourse analysis	➢	focuses on how knowledge is produced through the use of language
Hybrid approaches (concerned with meaning and context)		
Interpretative analysis	➢	attempts to present and re-present the world of those studied, by identifying and describing substantive themes, and searching for patterns between them

The process of analysis of qualitative data is inherently the same, whatever the technique used. The aim is to move from the individual case or raw data to higher levels of abstraction, that summarise, describe and explain. Figure 2 (below) represents this process. What is important is that this process is transparent, so each step in the abstraction process is visible and therefore open to scrutiny.

Figure 2 – Steps involved in the analysis of qualitative data



Based on Spencer, Ritchie & O'Connor (2003b).

At NatCen, our Qualitative Research Unit (QRU) developed a combined case and theme based, interpretative analysis approach, called Framework (see Ritchie & Spencer 1984). It is a matrix based analytical method that facilitates rigorous and transparent data management such that all the stages involved in the 'analytical hierarchy' can be systematically conducted. It also allows analysts to move back and forth between different levels of abstraction without loosing sight of the 'raw' primary data. In terms of where it fits within the data analysis approaches described in Figure 1, it is an hybrid, interpretative approach that is concerned with identifying substantive findings and addressing specific, often policy-related research objectives.

The name Framework comes from the thematic framework that is central to the method. The thematic framework is used to classify and organise data according to key themes, concepts and emergent categories. Main themes are identified and sub divided into a series of related sub-topics that evolve and are refined through familiarisation with the raw data and cross-sectional labelling. Once judged comprehensive, each main theme is charted in its own matrix, where every respondent is allocated a row and each column denotes a separate subtopic. Data from each case are then synthesised with the appropriate parts of the thematic framework. Currently this is done using Excel though in 2008 a standalone Computer Assisted Qualitative Data Analysis Software (CAQDAS) package will be available (see http://www.natcen.ac.uk/framework/index.htm for more information). Figure 3 summarises the main steps involved in carrying out this analysis.

Figure 3 Steps involved in carrying out a case and theme analysis

Step	Process
1) Familiarisation with data	Reading transcripts, listening to interview recordings
2) Identification of initial concepts/themes	Highlight, summarise, provisionally label
3) Categorisation leading to description	Iterative process of refinement, starting close to the data and becoming more abstract and interpretative, asking Qs of the data such as: • is this a different manifestation of that? • is this a subset of that? • is this of the same order as that?
4) Seeking explanations: informed by hunches and hypotheses, reflections during fieldwork and analysis and other research or theories	• Detailed within case analysis • Comparison between cases • Repeated interrogation of the data • Moving back and forth between cases searching for rival explanations

It should be noted that analysis is inherently a process of interpretation. We should not be afraid to ask questions of the data. These questions can be informed by theory or our own observations, hypotheses or hunches. If the analysis is rigorous and transparent then the data should be able to support or not support these. Steps 3 and 4 above therefore must be comprehensive: the data should be capable of supporting or refuting our ideas or theories, we should not (selectively) fit the data into the story we want to tell.

4. Tailoring Framework to cognitive interview data analysis

We were attracted to using Framework as an approach for analysing cognitive interview data because it provided a rigorous and transparent way of organising and interpreting our findings. In addition, on a practical note, we could utilise the in-house experience and expertise of the QRU and its training programmes. However we have had to adapt this approach to our needs, reflecting the different objectives of cognitive interviewing. This section is concerned with those adaptations.

It should be noted that other qualitative approaches might also lend themselves to the analysis of cognitive interview data. In particular a grounded theory approach or a conversation analysis approach. These may have useful application in particular settings. However, an attraction of the Framework, interpretative approach, is that it is focused on producing a substantive understanding of social phenomena and this translates well to cognitive interviewing, where we are interested in understanding how and why survey questions are failing to meet their measurement objectives.

The basic analytical process is the same as for qualitative data. Currently we tend to work from detailed notes made by the interviewer rather than verbatim transcripts. These notes, made after

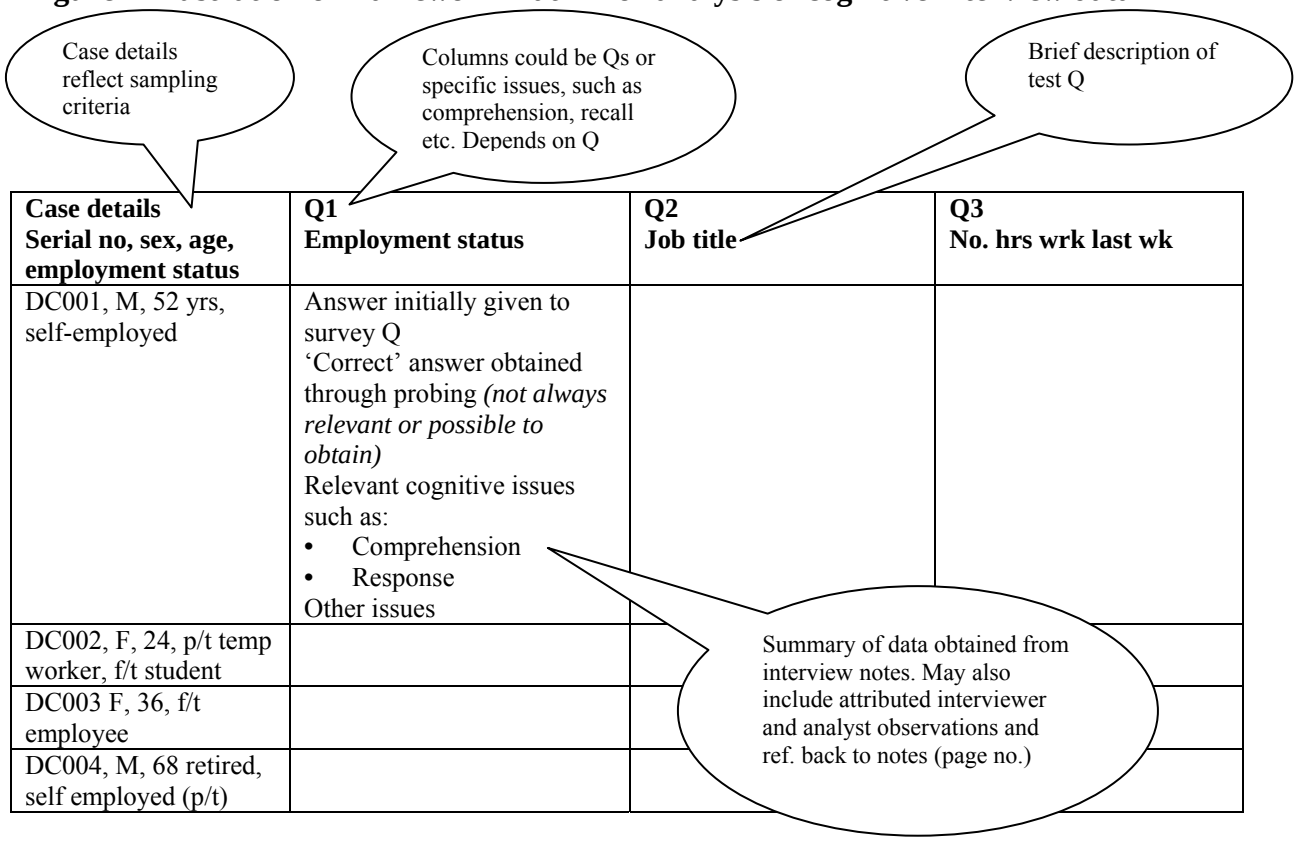
the interview by the interviewer reviewing the audio recording of the interview[19], are entered into a template, organised by test question and the key measurement issues to be explored (i.e. the aims of the test). The template is useful in providing a standard way for interviewers to report back on findings[20]. As well as containing details on respondents' understanding of the task or question, for example, interviewers can also record their own observations or explanations for what happened. These interviewer views are clearly identified (typed in uppercase or entitled interviewer note). Verbatim quotes are also recorded in the template (written in quotation marks, in italics). We tend to work from notes rather than transcripts for speed: the notes are produced electronically and emailed back to the research team in advance of the face-to-face interviewer debriefing. They are in a format that makes charting (completing the matrix) easier, as some sorting of the data into categories has already taken place in writing up the notes using the template. Whether this is approach is ideal is debatable, and an issue I will return to at the end of this paper.

Steps 2 and 3 in Figure 3 above - creating categories and classifying data within them - is a somewhat different process for cognitive interview data because the interview is much more focused or structured than a qualitative interview is. We are primarily interested in the cognitive processes respondents' use when attempting to answer a survey question or complete a questionnaire. In qualitative research, the categories would emerge from the data: in cognitive testing some of the analytical categories are implicitly built into the data collection process. The question and answer model informs our choice of scripted probes used in the interview, and these in turn form the main analysis categories. However, the framework is flexible and can be augmented to capture other issues that emerge from the data that do not fit within the cognitive framework. This is similar in some ways to the 'logical issues' categories developed by Willis et al (1999) in their standardised coding scheme, mentioned in section 2 above. Figure 4 below illustrates how the framework matrix for analysing cognitive interview data might be set up.

[19] All NatCen cognitive interviews are audio recorded with respondent consent.

[20] The template is often used on studies that have used probing and is based on the standard question and answer model. In addition, where think aloud is used we have found it useful to have these sections of the interview transcribed verbatim and then charted from the transcripts.

Figure 4 Illustration of framework matrix for analysis of cognitive interview data



Case details **Serial no, sex, age, employment status**	**Q1** **Employment status**	**Q2** **Job title**	**Q3** **No. hrs wrk last wk**
DC001, M, 52 yrs, self-employed	Answer initially given to survey Q 'Correct' answer obtained through probing *(not always relevant or possible to obtain)* Relevant cognitive issues such as: • Comprehension • Response Other issues		
DC002, F, 24, p/t temp worker, f/t student			
DC003 F, 36, f/t employee			
DC004, M, 68 retired, self employed (p/t)			

There are often several pages (worksheets within Excel) of the matrix. Each page may relate to a particular test question or to a set of questions (usually a series of questions that form a sequence). The structure of the matrix will reflect the complexity of the questions. For example, one survey question may involve respondents in a card sort exercise, looking through a series of possible reasons for why they are not currently engaged in paid work and being asked to put them into one of three piles: those that are a big factor, a smaller factor or not a factor. There are many cognitive stages involved in this task, and splitting these out into separate categories (columns) will be helpful in summarising and making sense of the data, rather than having a mass of detail in one cell.

Each cell should contain a summary of the key points, meaning it should contain enough information to clearly communicate these, referencing back to the original source material (in our case the interviewer notes). This linking back to the raw data is important in making transparent the data reduction process. The contents of the matrix may be refined, through an iterative review process, whereby additional information is added (for example reference to other cells within the matrix where the respondent demonstrates the same or different behaviour). The completed matrix would be reviewed by all members of the research team prior to analysis and reporting commencing. The team would then meet to discuss the findings, further analysis required and plan the report.

A key component of the Framework analytical approach, described in section 3, is the need for the analysis (steps 3 and 4) to be comprehensive. In analysing cognitive interview data we take this to mean we need to cover the diversity of strategies used and problems encountered by

respondents. In addition, as survey methodologists also need to comment on the significance of these problems, inferring whether such problems are a quirk (e.g. a logical problem, where the respondent just went off at a tangent) and thus could be seen as random error or are evidence of a more systematic error, stemming from the structural form and implementation of the question or questionnaire. The latter is something that we would want to address, for example by making recommendations for rewording the question.

In summary, we find the use of Framework as an analytical tool helpful in providing us with a systematic, transparent and rigorous way of exploring cognitive interview data. We can look at within and between case associations, which provides us with greater analytical power as the context in which respondents formulate their answers to survey questions can be captured. These features also assist us in being able to demonstrate the credibility of cognitive interviewing methods and the recommendations for changes in questionnaire design and wording that stem from them to our clients and sponsors.

5. Next steps

This paper has highlighted the lack of documentation and discussion about how cognitive interview data are analysed within the current literature. This threatens to undermine the credibility of cognitive interviewing methods as a tool for identifying problems with survey questions and questionnaires. Qualitative data analysis methods have much to offer us, and we should take note of the debate over quality in this field (see section 2) and start to consider how cognitive interview data should be analysed.
As a starting point I would suggest the following.

- There should be better documentation of how cognitive interview data are analysed in our reports and publications. Details such as how the raw data were captured, what was analysed (notes, transcripts etc) and how the data were summarised and interpreted need to be described (in more detail).
- We should consider whether the development of best practice guidelines for the analysis of cognitive interview data would be helpful, and if so what these should be. It is notable that those engaged in the enterprise of developing survey questions for comparative (cross national and cross cultural) research have begun to raise the lack of standardisation of methods as an problem in evaluating survey questions (see for example Willis et al, 2005; Miller, 2007).
- As part of this process we should consider what the source material should be for analysis; audio recordings, verbatim transcripts or detailed notes for example. If the latter then are there principles that should inform how these are made, for example, always making them after the interview, with reference to the recording of the interview?
- In addition, is there a case for integrating standardised code frames, such as those developed by Presser and Blair (1994) or Willis et al (1999) within a thematic framework approach? This might facilitate standardisation. However this would need to be done in such a way that the flexibility of the thematic approach is maintained, which ensures emergent issues can be captured, the analysis remains grounded in the data and the context within which respondents formulated their answers to survey question is retained.
- Is it appropriate to use statistics to describe cognitive interview data? I have argued in this paper that these data are inherently qualitative and as such it does not seem appropriate to use quantitative descriptors. Such practices may well serve to undermine the credibility of

our findings among (some) of our clients, who are often statisticians. Rather, we should utilise qualitative data analysis reporting techniques to better effect.

Ultimately the key issue is whether cognitive interview findings and the recommendations we make based on these are valid, reliable and objective. Improving the rigour and transparency of the analysis process will go some way to achieving this goal, but we also need to develop multi-faceted testing strategies that provide opportunities to triangulate different sources of data on the performance of survey questions, and that assess the success of any proposed change. These should include split ballot experiments to compare our recommended changes in wording or design with the original version, for example see the papers presented at QUEST by Consenza (2007), Fowler (2007) and McGee (2007).

References

Conrad F, Blair J, Tracy E (1999) *Verbal Reports are Data! A theoretical approach to cognitive interviews*. Federal Committee on Statistical Methodology Conference Proceedings. www.fcsm.gov/99papers/conrad1.pdf

Consenza C (2007) *How Much is Too Much? How adding information to a question may narrow, rather than expand, a respondent's understanding*. Proceedings of QUEST 2007, Ottawa. ADD WEB LINK

DeMaio TJ, Landreth A (2004) "Do Different Cognitive Interview Techniques Produce Different Results?" In S Presser, JM Rothgeb, MP Couper, JT Lessler, E Martin, J Martin and E Singer. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, New Jersey: John Wiley & Sons. 89-108.

Enerstvedt, R.G. (1989) 'The Problem of Validity in Social Science', in S. Kvale (ed) *Issues of Validity in Qualitative Research, Lund,* Sweden: Studentlitteratur.

Fowler J (2007) More Evidence on When Two (or More) Questions are Better than One. Proceedings of QUEST 2007, Ottawa. ADD WEB LINK

Kaplan A (1964) *The Conduct of Enquiry: Methodology for Behavioural Science*. San Francisco: Chandler.

Le Compte, M. and Goetz, J. (1982) 'Problems of Reliability and Validity in Ethnographic Research', *Review of Educational Research.* 52: 1, pp. 31–60.

Lewis J, Ritchie J (2003) "Generalising from Qualitative Research". In J Ritchie and J Lewis (eds) *Qualitative Research Practice*. London, Sage. 263-286.

Lincoln, Y. and Guba, E. (1985). *Naturalistic inquiry*. Beverly Hills: Sage.

McGee A (2007) *How do we assess whether we are improving instrument design? Using multiple methods to evaluate whether a re-designed travel record was better than the existing one*. Proceedings of QUEST 2007, Ottawa. ADD WEB LINK

Miles MB, Huberman AM (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. London: Sage.

Miller K (2007) Design and Analysis of Cognitive Interviews for Cross-National Testing. Paper presented at the European Survey Research Association Conference, Prague.

Patton, M.Q. (2002) *Qualitative Research and Evaluation Methods.* Thousand Oaks, CA: Sage.

Presser S, Blair J (1994) "Survey Pretesting: do different methods produce different results?" In P Marsden (ed) *Sociological Methodology*. San Francisco: Jossey- Bass. 73-104.

Ritchie J, Spencer L, O'Connor W (2003) "Carrying out Qualitative Analysis." In J Ritchie and J Lewis (eds) *Qualitative Research Practice*. London, Sage. 219-262.

Rothgeb J, Willis G, Forsyth B (2001) "Questionnaire Pretesting Methods: Do different techniques and different organisations produce similar results?" *Proceedings of the ASA Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Spencer L, Ritchie J, Lewis J, Dillon L (2003a) *Quality in Qualitative Evaluation: A framework for assessing research evidence*. London: Government Chief Social Researcher's Office. http://policyhub.gov.uk/docs/qqe_rep.pdf

Spencer L, Ritchie J, O'Connor W (2003b) "Analysis: Practices, principles and processes". In J Ritchie and J Lewis (eds) *Qualitative Research Practice*. London, Sage. 199-218.

Willis GB (2005) *Cognitive Interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis G, Lawrence D, Thompson F, Kudela M, Levin K, Miller K (2005) The Use of Cognitive Interviewing to Evaluate Translated Survey Questionnaires: Lessons Learned. http://www.fcsm.gov/05papers/Willis_Lawrence_etal_VIIIB.pdf

Willis GB, Schechter S, Whitaker K (1999) "A comparison of cognitive interviewing, expert review, and behaviour coding: What do they tell us?" *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Evaluating Filter Questions Used for the Participation and Activity Limitation Survey (PALS)

David Lawrence
Questionnaire Design Resource Centre, Statistics Canada

1. Background

In 1986, Statistics Canada conducted its first post-censal disability survey – the Health and Activity Limitation Survey (HALS) – following the 1986 Census. The survey was repeated after the 1991 Census. In both 1986 and 1991, the census long form included two general filter questions on activity limitations. Information collected from these questions was used to define the survey population and select respondents for the two post-censal surveys.

The early filter questions failed to perfectly pre-identify the desired target population for the two surveys on activity limitations. In 1998, an extensive research project was initiated to determine whether improved filter questions could be developed. It was hoped that better functioning questions would provide a more general, yet standardized indicator of disability that could be applied in other Statistics Canada social surveys.

A specific set of criteria was desired. The questions should be:

- applicable to entire household population: children, adults and seniors
- sufficiently succinct to be used in census as well as various social and health surveys
- portable across different modes of data collection
- broad enough to permit persons with all types and levels of disability to be included.

A new set of activity limitation filter questions was developed for the 2001 Census. These questions have been used to specify the target population for both the 2001 and 2006 post-censal surveys. The survey was re-named to the Participation and Activity Limitation Survey (PALS).

While both HALS and PALS provide detailed information about the demographic and socio-economic situation of persons with disabilities, the data from the two surveys are not comparable due to differences in the filter questions from 1991 and 2001, as well as differing questionnaire content and sampling plans used. Since 1999, several of Statistics Canada social and health surveys have used the modified filter questions in their data collection. These surveys include the *General Social Survey*, the *Survey of Labour Income Dynamics*, *Youth in Transition Survey*, the *Participation and Activity Limitation Survey*, the *Canadian Community Health Survey* and the *National Population Health Survey*.

2. Study Rationale and Objectives

Measuring disability is a broad, complex and frequently subjective challenge for survey organizations. Census filter questions offer the advantages of using detailed census information to improve the efficiency of the sample design and reducing response burden; however, operational and space restrictions often limit the number of filter questions allowed on a Census form. Hence, the identification of persons in the target population can often be less precise.

The PALS project team is interested in better understanding any anomalies between the Census filter questions and the post-censal survey. A small qualitative research study was initiated following data collection for the 2006 PALS.

- The primary intent of the study was to better understand why certain PALS respondents were identified differently between the Census and the survey.

- A second intent of the study was to better understand how PALS respondents interpret different questions on activity limitations. As noted above, the activity limitation filter questions are currently used by several other Statistics Canada social and health surveys. As part of an on-going initiative to improve the efficiency of social statistical data, Statistics Canada is attempting to harmonize various questions used in its social surveys.

This qualitative research project - the first of several phases - focused on two specific respondent-types:

- *False positives* occur when persons identified as having activity limitations on the Census questions are not similarly identified when administered the PALS questions.
- *Soft conditions*. There is a perception that the current filter questions may not firmly identify persons who have a participation or activity limitation due to an emotional, developmental, memory or learning condition.

The specific research objectives were:

- To obtain feedback from participants on their interpretation of the questions on the Census questionnaire and the PALS questionnaire and better understand the reasons why respondents are identified differently on the Census and PALS.

- To evaluate the performance of the Census filter questions by exploring differences among participants' responses and reactions to these questions as well as the filter questions used on the Canadian Community Health Survey (CCHS).

- To explore issues related to proper identification of respondents with so called "soft conditions" through the PALS questionnaire.

3. Methodology

A total of 58 one-on-one, in-depth interviews were conducted in the following four cities: Toronto (15 English interviews), Halifax (13 English interviews), Montreal (14 French interviews) and Ottawa (7 English and 9 French interviews). All in-depth interviews were carried out in a focus group facility equipped with a one-way mirror to allow for observation by members of the project team. The interviews lasted between 35 and 45 minutes. Each participant received a payment of $75 to cover any expenses they may have incurred as a result of participating in the study.

Participants were comprised of the following types of respondents:

- False positive - adults & children
- Adults & children having a "soft disability"
- Adult and child records with other reporting incongruities

(For children, the interviewed participant was the parent who responded to PALS by proxy)

4. Findings and Observations

During each cognitive interview, the test questions were administered orally to the participant. Concurrent probing was used to better understand participants' comprehension of each question and to explore how participants went about formulating a response.

Note: All test questions were read-aloud to participants during this study. While the Census questionnaire is self-completed, the Canadian Community Health Survey (CCHS) and the other social and health surveys using the activity limitation filter questions are interviewer-administered.

4.1 First Census Filter Question

Participants were first asked to answer the following Census filter question:

> [Do you / Does …] have any **difficulty** hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing similar activities?
>
> - Yes, sometimes
> - Yes, often
> - No
>
> Cette personne a-t-elle de la **difficulté** à entendre, à voir, à communiquer, à marcher, à monter un escalier, à se pencher, à apprendre ou à faire d'autres activités semblables?

While participants did not perceive the question to be particularly difficult to answer, their interpretations varied.

Question context

The activity limitation question is located at the beginning of the Census long form, immediately following household member profile questions such as name, sex, date of birth, and marital status. The question asks about 'any difficulty' without a reference to time, severity or chronicity.

- Several participants included short-term injuries and acute conditions such as broken limbs, sport-related injuries, and asthma when answering the question. Some participants reported wearing eyeglasses to correct a vision problem as a 'difficulty'.

- A few participants - reporting by proxy for elderly parents - indicated their parents had certain difficulties with mobility and hearing. These conditions were attributed solely to 'old age' and were not considered in any way as a 'disability' or 'limitation'.

- One respondent, who completed his Census form on behalf of his household included his 12 month old child. At the time of the Census, the child could not walk or speak. The respondent answered the filter question positively. It was only at the time of the post-censal survey that he realized he should not have included any difficulties that are attributed directly to age. He suggested during the interview that the question should not be applicable to infants and toddlers.

Question length

- For many participants, the question was long and vague. Several said they were unable to separate and process the list of difficulties. Some respondents either missed or misunderstood certain conditions that did affect them.

- A small number of participants felt the list was all inclusive; they had to experience difficulty with <u>all</u> the listed conditions in order to answer the question positively.

- In other situations, participants did not listen to the entire question. When they heard the condition that affected them, they immediately self-identified and stopped listening to the question at that point.

Question consistency

- The English question uses the word *'any'*. No similar word such as: *'quelconque'* is used in the French version. A slightly revised wording was tried with some French participants in Montreal and Ottawa: *«Avez-vous une difficulté quelconque à entendre... »*.

- Two effects were observed:
  - For some respondents having emotional or psychological difficulties, the use of *'any'* or *'quelconque'* helped them answer positively.
  - For some immigrant participants, the phrasing caused them to include any little problems that might occur daily. Others would be inclined to include the fact they wear glasses. A few Ottawa participants indicated that using the word *'quelconque'* made the question more inclusive.

Response categories

- For some respondents the response categories and the long list of conditions made the filter question difficult to answer. A young adult with a learning problem stated that she struggled when trying to answer the question saying for some conditions she experienced no difficulty, but for others it was 'often' or 'sometimes'.

- Other participants felt the offered response categories did not address their situation – the responses did not provide a natural response option.. They noted that their condition could not be simply summarized as a "Yes" or a "No". Others were confused when they heard

"yes, often or "yes, sometimes". They were uncertain if the question wanted a 'yes/no', or an often/sometimes' response.

- Frequently, participants answered *"it depends"* stating that factors external to the condition itself may influence how they would answer the question. Others noted that the answer would depend on one's perception. A mother reporting by proxy for her son noted that he had no problems communicating with his family or his teachers, but: *"… if he were to talk to you for the first time, then you would have problems understanding him."*

4.2 Canadian Community Health Survey (CCHS) Filter Question

The CCHS is a major national health survey conducted bi-annually by Statistics Canada. The survey questionnaire is administered as a computer-assisted interview. Typically, this question is asked mid-way through the CCHS interview, following other modules on health status and determinants of health.

Participants' interpretations and reactions to this version of the question were of particular interest to the study sponsors.

> The next few questions deal with any current limitations in your daily activities caused by a long-term health condition or problem. In these questions, a 'long-term condition' refers to a condition that is expected to last or has already lasted 6 months or more.
>
> Do you have any **difficulty** hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing similar activities?
> - Sometimes
> - Often
> - Never

Introduction and concepts

The CCHS version frames the filter question using concepts such as "current limitations", "daily activities", "long-term health condition or problem" and "expected to last or has already lasted 6 months or more".

- For several participants, the introduction, although viewed as being more precise, did not change how they answered the question.

- Others answered the CCHS question differently than the census question. When probed, respondents indicated that the terms caused them to think differently about the question. Some explained how they processed the questions as: the CCHS question was about '*lifelong'* problems, whereas the Census question was *'asking about the present'*.

- Several participants found the CCHS introduction to be long – particularly for a telephone interview. Some participants who suffer from episodic conditions such as migraines or bouts of arthritis were confused by the introduction. Although their conditions are chronic and very limiting, they do not experience the difficulties every day.

- In a few situations, participants changed their answer on the second reading of the question simply because they did not completely hear or understand the complete list of conditions when read the first time. (This again speaks to the vagueness and question length of the filter question in an interview setting).

Question context

- The concepts did not effectively clarify the question for all participants. For example, some respondents with 'soft conditions' such as learning difficulties, explained that their difficulty had lasted for more than 6 months; however, it did not really limit their daily activities. As a result they did not know how to best answer the question.

- A mother reporting for her son indicated 'sometimes' for Census and 'never' for the CCHS. Her 9-year old son has a communication problem. In her mind, she answered the census question thinking about his speech problem when he first started school. For the CCHS, she was thinking more in terms of his current situation and improvements that he has made.

Long-term condition

- Most participants viewed 'long-term' as something that has been present since birth, or something that might last for one's entire life. Others suggested it may be a condition that is the result of an accident or illness. Some respondents indicated it is something that is permanent: *"not something that will heal"*. Invariably, 'long-term' was interpreted to mean a period longer than 6-months. Including both statements in the question was confusing for many respondents.

4.3 Census Filter – Activity Reduction Questions

The second Census filter question relates to activity reduction. This question is also asked on the CCHS survey.

> Does a physical condition **or** mental condition **or** health problem **reduce the amount or the kind of activity** you can do:
> … at home?
>
> O Yes, sometimes
> O Yes, often
> O No
>
> … at work or school?
>
> O Yes, sometimes
> O Yes, often
> O No
>
> … in other activities, for example transportation or leisure?
>
> O Yes, sometimes
> O Yes, often
> O No

Physical condition

- Most participants associated a *physical condition* with any limitation or restriction of mobility. Some respondents also considered injuries affecting one's mobility as a physical condition.

Mental condition

- When asked about *mental conditions*, participants had two distinct perspectives on the term: Many participants having some form of learning problem (or reporting by proxy for someone with a learning problem) associated a mental condition with any activity related to concentrating, memory, reading, writing or math. However, these respondents did not like using the term 'mental condition' to describe their situation. They felt it had a negative connotation. Some said that they could not easily categorize their condition as physical, mental or a health condition.

- Participants not suffering from a learning problem more frequently associated *mental condition* with conditions such as depression or schizophrenia. Several respondents said that it would be a condition that was diagnosed by a doctor. At least two participants wondered if emotional conditions should be included here, or if mental or emotional conditions were separate issues.

Health problem

- Many young participants associated health problems with conditions such as asthma or even a cold, flu, or fever. Older participants tended to identify health problems as other chronic conditions such as heart disease, arthritis and emphysema.

- Several participants suggested that the question would be clearer if asked separately for 'physical condition', 'mental condition' and 'health problem'.

Activity

- Participants found the term 'activity' to be quite vague. It was suggested that the question should include examples to clarify the scope and intent. For example, one participant with a learning problem, wondered if personal finances and banking should be considered as "at home" activity. If this activity was in scope, then he would respond "often" to the question, otherwise his answer would be "no".

- Another respondent answering on behalf of her elderly mother indicated that her response would depend on what was considered an activity. Accounting for her age, the respondent felt her mother was not at all limited in her day-today activities, but remarked: *"…I wouldn't ask her to go hiking or something…"*.

Summary of Findings and Observations

The intent of this preliminary study was to explore participants' interpretations of certain activity limitation questions in order to better understand the reasons why certain respondents identify differently between the Census and the PALS post-censal survey.

This paper has focused primarily on the findings and observations from a series of in-depth, one-on-one interviews conducted with respondents who recorded as *false positives*, or who were identified as having a *soft condition* on the 2006 Census and 2006 Participation and Activity Limitation Survey.

Issues of question context, question location, terminology used, mode of data collection and proxy/non-proxy responses were key factors that influenced reporting differences for the test participants.

Extensive qualitative and quantitative research was carried out in the late 1990's to develop an improved set of Census filter questions to accurately pre-identify the target population for the post-censal activity limitation survey. These questions seem to work very well at screening in persons with severe or moderate disabilities and generally work very well with the intended methodology and sample design.

It is unclear how well these questions pre-identify persons with less severe limitations, or those with *soft conditions*. More research is required to further assess how to ask questions that might consistently identify persons with these attributes.

What next?
More research is proposed in this on-going effort to better understand this complex issue.

- A second qualitative study with 'false negative' respondents to the 2006 PALS is scheduled for May-June 2007.

- Further evaluation of 2006 PALS data

- Further research exploring possible revisions to the Census filter questions may be proposed.

- Address on-going issues pertaining to the harmonization of activity limitation questions at Statistic Canada.

Who Is More Likely To Attend? A Study Of "No-Shows" In Qualitative Research

Benoit Allard
Questionnaire Design Resource Centre, Statistics Canada

Abstract

Since early 2005, Statistics Canada's Questionnaire Design Resource Centre has been compiling basic information on people who were recruited to participate in cognitive interviews and focus groups. In addition to the variable of interest (whether the recruit actually attended the interview or focus group), information is being gathered on two broad categories of variables: 1) information about the testing activity (such as the topic under study, the testing methodology, language, location and time of testing, amount offered to participants, etc), and 2) basic demographic information about the recruited person.

The resulting data are now being explored to look for correlations between the recruits' characteristics and their likelihood of actually showing up for the interview or focus group. Through this study of "no-shows", we hope to find answers to questions such as: Are there factors (or combinations thereof) that influence a recruit's likelihood of participating in the study? Can we identify these factors and adjust recruiting specifications to account for them (for example, by over-recruiting when the risk of having "no-shows" is high)? Which companies are better at finding recruits who actually participate? How much does the amount offered to recruits influence their turnout? This paper presents the first results from this exploratory analysis.

Background

Cognitive interviews and focus groups are the main methods for questionnaire testing by Statistics Canada's Questionnaire Design Resource Centre (QDRC). Whether for questionnaire testing or for other purposes, qualitative research such as interviews and focus groups is subject to "no-shows" – that is, persons who accept the invitation to participate but do not participate for some reason. With this study, we examined the characteristics of "no-shows" in comparison to those who actually attended, to try to better understand the factors that influence attendance at qualitative research activities.

Recruitment for the QDRC's research is done mostly by private firms, although sometimes QDRC staff will recruit the participants (for example, when using lists from confidential survey data). Participants are usually recruited to obtain a good mix along such variables as age, sex, education and income, although tighter criteria are occasionally applied.

For research involving individuals (as opposed to businesses which are not covered by the study), "no-shows" make up roughly 15% of all recruited persons. While this averages out to 1 or 2 missing people per focus group, there is some variability in attendance – for example, groups have (fortunately, rarely) been held with only 5 or 6 participants out of 11 or 12 recruited. The "no-show" study was motivated by questions which recurred periodically at the QDRC, such as:

- Which recruiters do the best job?
- Does the amount offered to participants influence their turnout?
- What factors influence attendance – which groups are more susceptible to "no-shows"?
- Do certain times (e.g. time of day, day of the week) present a higher risk of "no-shows"?

By answering some of these questions, the QDRC might be able to improve turnout for its testing activities by adapting the recruiting specifications (for example, by limiting the number of recruits with some high-risk characteristic), with a better choice of location and schedule, and/or by over-recruiting in situations that might present a high "no-show" rate.

Methodology

The "no-show" study is an ongoing, quantitative study of individuals who have been recruited to participate in the QDRC's qualitative research activities such as focus groups and cognitive interviews. Since early 2005, the final lists of recruits submitted by recruiters are being captured into the "no-show" database. These are the individuals who accepted the invitation, after last-minute replacements have been incorporated. No identifiers are captured.

For each recruit, the following variables are captured:

- Variables relating to the testing activity:
  - Topic of discussion
  - Type of testing (focus group or 1-on-1)
  - Location (facility or respondent's home)
  - Province
  - Date and time of testing
  - Payment amount
  - Recruiter
- Variables about the recruited person (these are gathered during recruiting via a screening questionnaire):
  - Age
  - Sex
  - Presence of a spouse/partner
  - Presence of children in the household
  - Work status
  - Income
  - Education
  - Attendance (i.e. did the person actually attend the activity?) – the variable of interest.

When enough records were accumulated, the first exploratory analysis was made in late 2006 (approximately 1,300 individuals). The rate of "no-shows" was examined for various combinations of factors – log-linear models were used for significance tests.

There are a few limitations to this kind of study, the first being that it is blind to the recruiting process – we have information on recruits who have agreed to participate in qualitative research, therefore any difficulty in getting them to accept (e.g. too low a payment, or being a single parent) has been overcome already. The "no-show" data do not reflect the level of difficulty in recruiting participants. The effect of the amount being offered to participants, for one, does not seem to influence "no-shows"; recruiters tell us that it influences the difficulty they have in recruiting participants (indeed, the more we offer participants, the less recruiters will charge us for the work) but this is of course not reflected in the "no-show" data.

Another limitation is that (in this particular case at least), the data was being exclusively generated by the QDRC's production work, with very little control over the experimental design. However, the variables collected were standardized, i.e. the various recruiting companies were asked to use consistent response categories for age and income.

Results

While the gathering of data is ongoing and further analysis is pending, here are the first results to come out from the "no-show" database.

Recruiters

Which recruiting firm(s) to hire is a decision made on a project-by-project basis. The QDRC has standing offers with 8 recruiting firms scattered across Canada. The choice of recruiter(s) for a particular project is usually based on their capacity to recruit in certain regions of the country. As some projects involve testing activities in several regions of Canada, sometimes a single recruiter is hired to recruit at all locations, with perhaps a separate company recruiting in French when there is also testing of the French version of a questionnaire.

The data has revealed variation among recruiters as far a "no-show" rates go, ranging from about 5% to 20%. An interesting find is that, within a region, the companies that show the best "no-show" rates tend to be those based in this region. Companies recruiting outside their "home" region, in general, achieve lower turnouts. One can speculate the following explanation: being less familiar with the area, a recruiter might give less accurate directions to the focus group facility than if they were recruiting people within their own region or even their own facility.

To minimize "no-shows", therefore, it is preferable for the QDRC to hire a separate recruiter for each region (West, Ontario, Quebec, and East) rather than use a single recruiter across the country. This entails a little extra paperwork but should help reduce "no-shows".

As the standing offers between the QDRC and the recruiting companies is subject to renewal every year, and expires after five years, results about recruiters' "no-show" rates will help to better assess their performance when evaluating proposals and considering renewal.

Location of one-on-one interviews

Whether to conduct cognitive interviews at a central facility or at the respondent's home is also decided on a project-by-project basis. Both locations have their advantages and drawbacks: a facility can accommodate many observers behind a one-way mirror, and provides a more professional setting; a home interview may allow the participant to feel more at ease (in their "home court", so to speak). As far as costs go, both options are more or less equivalent – the extra time and money spent going door-to-door is offset by the cost of renting a facility (this would be different for an agency that had its own testing facility).

The data reveals twice as many "no-shows" for interviews in a facility, as in the participant's place of residence (15% vs. 7%). Intuitively, this makes sense – in order to avoid an interview at the last minute, one needs only stay home if the interview is taking place somewhere else; not so if the interviewer is coming to one's home.

There are two ways to take advantage of this result: one can favour interviewing at the participant's home when the conditions allow it (only one observer per interview, longer testing schedule); or one can compensate by over-recruiting when the interviews have to be done in a facility.

Age, Sex, Family

Young adults (18-25 years), with a "no-show" rate of 24% overall, are the least likely to attend after having been recruited for some qualitative testing activity. When a particular project specifically targets this group (for example if the questionnaire for a survey of graduates were being tested), therefore, the QDRC has become aware of the need to over-recruit this population – as many as 13 recruits(as opposed to the usual 11) may be required to ensure an actual group size of 9 or 10. At the other end of the age spectrum, seniors (65 +) and retirees had "no-show" rates under 10%.

Other demographic variables, like sex or family situation such as the presence of a spouse/partner or children in the household, were not correlated with attendance. Whatever differences there are have been accounted for in the recruiting process – for example, a single parent may be more difficult to recruit but once they have accepted to participate they are as likely to attend as the population as a whole.

Time variables

From the captured date and time of the testing activity, we were able to derive variables like period of the day (morning, lunch, afternoon, late afternoon, evening), day of the week (Monday, Tuesday etc.) and season (spring, summer, fall, winter) to see if these factors influenced turnout. Few differences arose except for the following.

The QDRC does significantly less field work on Mondays (three times more field work is done on both Tuesdays and Wednesdays), for several reasons: we often use Mondays to travel to a test site; some of the staff are on a part-time schedule and don't work on Mondays; others prefer to schedule testing later in the week in order to spend the Monday for preparation. As it turns out,

Monday "no-show" rates are somewhat better (9%) than those in the rest of the work week (Tuesdays to Thursdays are average, and Fridays are the worst at 17%). The Monday result is close to being significant (more incoming data could confirm it), so it appears that Mondays should be favoured when possible.

Time of day is also an interesting factor: Interviews done in the morning have a very low "no-show" rate (7%) while the late afternoon rate (defined as beginning between 4:00 and 6:00 pm) is 20%. Of course, one cannot schedule all interviews in the morning if 5 or 6 have to be done during the day. But this information can help improve scheduling, for example by placing the dinner break so as to avoid interviews starting in the late afternoon.

Conclusion & Further Research

While far from revolutionizing recruiting practices, a "no-show" database is allowing the QDRC to pinpoint some of the factors that influence attendance and take appropriate action to minimize the impact of "no-shows".

Further research on the "no-show" database will include:

- Investigating other variables and more complex interactions
- Case studies, in particular past events with low turnout: if only 5 or 6 persons showed up out of 11 recruits for a focus group, was it bad luck or a combination of unfavourable factors that could have been avoided?
- As data keeps being gathered, more results will become significant. However, we must also verify that the effects observed are constant over time.
- The data will also serve to validate the effectiveness of any corrective action taken by the QDRC to reduce "no-shows" as data gathered after these steps have been taken comes in.

Harmonization of the Minimum Health Module in the European Health Interview Survey (EHIS)

Gunilla Davidsson
Statistics Sweden

Introduction

This paper just want to give you an idea of the struggle Eurostat and the Member States have to deal with when trying to harmonize surveys to be used in many countries with many languages and also when having the same questions in different surveys introduced at different times for different purposes.

Background

EU and the member states have agreed on conducting a number of surveys in different areas in a harmonized way. To make this come true imperative regulations concerning each survey have been taken by EU Parliament and Council. In the health area Eurostat has spent several years to organize the development of a series of modules concerning health status, life style, care consumption, care costs, disability, etc. The intention with having a number of modules is to give the countries possibility to choose themselves how to implement the modules – all in one survey of its own or embedded in an already existing national health survey or as extra modules added to one or several other national surveys. In the following I talk about the modules as a health survey since the first version is launched as a survey of its own.

In order to get a health survey of harmonized questions in all 27 EU member states and another five European countries (Iceland, Norway, Switzerland, Turkey and Croatia), special proceedings have been developed. At first a reference questionnaire in English has been produced by using old well-known, internationally often used questions as starting-point. To assure a harmonized translation, which takes cultural differences into consideration, conception cards with background and rationale have been produced about each question. Unfortunately there is no rationale to be find for some of the questions, i.e. the mental health score of SF 36. A special, rather far-reaching, time-consuming translation procedure has also been decided upon.

A draft of the reference questionnaire was cognitively tested by ONS in UK. The result of the test was a number of suggestions to changes. But the problem was that another Eurostat survey (called EU-SILC), already running, is using some of the health questions with exactly the same, slight or more considerable differences in the wording. It is a very time-consuming procedure to make any change in an already regulated survey. The EU-regulation of SILC, which is an annual survey on income and living conditions, is unfortunately very detailed. Another problem was that some countries were well on their way to start data collection of their national health surveys. They were using the first version of the by Eurostat recommended questions and had now neither time nor means to start the translation process all over again.

State of the art at the moment: The regulation of the health survey is on its way, but yet not taken. The first round of data collection in the 32 countries is proceeding with not fully harmonized surveys. The possibility to improve the final health survey still remains, but not for a

very long time. This also means that the problems in harmonization with the SILC-survey has to be solved very soon.

Some early noticed problems were connected with using questions developed for use in the US (for example SF 36). Most of these sets of questions have probably never been properly tested in a cognitive way in Europe before, they have only been translated and used as they are. To find out more about these problems Eurostat offered the member states to make pilots and cognitive testing of their national versions of the reference questionnaire. Some of these results are presented and discussed in this paper.

In Sweden a second purpose with the cognitive study we performed was to find out if immigrants living in the country since many years would understand the questions in the same way as people with Swedish as their mother tongue do. This is important to us since about 10-15 percent of the Swedish population do not have Swedish as their mother tongue and Statistics Sweden very seldom translate questionnaires to other languages.

Among the other European countries Finland, Norway, Estonia, Czech Republic, Austria, Cyprus, Poland, Romania, Italy, Slovenia, Slovakia, Spain and the Netherlands had pilots or cognitive testing. The cognitive testing were performed in different ways, but the results are still interesting.

Some Results of the Cognitive Testing of the Minimum Health Module

In the following is presented some results from cognitive testing both by ONS and by some of the other countries. To keep this paper short the discussion is only concerning the set of three questions called the Minimum Health Module – self-perceived health, chronic or longstanding conditions and limitations due to health problems. This set of questions is already used in EU-SILC and will be used in EHIS (and other future Eurostat surveys) as well.

Question 1

The **first version** *of question 1 was the "old traditional" question wording recommended by WHO a long time ago.*

How is your health in general? Is it very good, good, fair, bad or very bad?

During the translation work many countries had problems finding proper words in their own languages to the middle alternative 'fair'. Since the English word 'fair' can have both a positive and a negative meaning, the problem is which one to chose if you have two different words in our own language. This resulted in a new version by Eurostat with a new middle alternative 'neither good nor bad'. This ***second version*** was used in the cognitive testing by ONS and the countries.

How is your health in general? Is it very good, good, neither good nor bad, bad, very bad?

ONS recommendation was to go keep to the original middle alternative 'fair'.

Comments from other countries:

The understanding of the term 'in general' is somewhat differing. Some respondents take into account their social and mental well-being as well as their physical well-being, while others do not. Some respondents define good health as a state where illnesses and diseases are kept in check with medication or other treatment, while others define this state as 'neither good nor bad'.

It is obvious that people of different age have varying frames of reference when talking about 'own health in general'. In the eldest group health is often related to not being dependent of care. Somewhat younger but still senior people relate health primarily to physical conditions and have health of people of same age as frame of reference. The middle-aged refer to mastery of everyday activities and many of them use themselves somewhat younger as frame of reference. Younger people think a lot about being both physically and mentally fit and compare themselves to people of the same age.
Several countries describe problems with the middle alternative. A common view is that 'neither good nor bad' has no real content. One respondent said: "I think 'neither nor' is a bit wrong expression because then my health is nothing, non-existing - am I really alive? or am I a zombie?". Other respondents pointed out that the middle alternative ought to be 'good' since that is the natural standard to measure against when you talk about health. Some thought that 'neither nor' also could be interpreted as 'sometimes better/less pain/etc and sometimes worse/more pain/etc'. Unfortunately the Swedish equivalence to 'fair' is not widely known by immigrants regardless of how many years they have been living in Sweden. Another solely Swedish problem is that the meaning of Swedish equivalence to 'fair' is changing from being used to describe something as slightly negative among elderly people to somewhat slightly positive among younger people.
The outcome of the cognitive testing and also the fact that middle alternative in EU-SILC is 'fair', has made Eurostat to go back to the wording in the first version. The ***third prevailing version*** which is now officially decided upon:
How is your health in general? Is it very good, good, fair, bad or very bad?

In the additional background and rationale is stated:
The reference is to health **in general** rather than the present state of health, as the question is not intended to measure temporary health problems. All dimensions of health, i.e. physical, social and emotional function and biomedical signs and symptoms, should be included. It omits any reference to an age as respondents are not specifically asked to compare their health with others of the same age or with their own previous or future health state. It is not time limited.
- Fair: this intermediate category should be translated into an appropriately neutral term, as far as possible.

Question 2

The **first version** *of question 2 was the same as the one already being used in EU-SILC.*
Do you have any longstanding illness or health problem?

ONS recommended after cognitive testing the following wording:
Do you have any longstanding illness or longstanding health problem? By longstanding I mean anything that has troubled you over a period of time or is likely to affect you over a period of time.

Comments from other countries :
In all countries several respondents have difficulties understanding what is meant by the term 'longstanding'. How long has the illness or health problem to be present to be considered as longstanding? Then there is the problem with medication, some respondents answer No even if they have serious health problem as long as the problems are kept in check by medication (i.e. one respondent with diabetes dependent on insulin who did not consider diabetes to be neither an illness nor a health problem). Other respondents with similar conditions answer Yes.

Longstanding health problem is by many seen as a problem that can not be cured, while 'temporary health problem' is curable. This can be either physical or mental health problems. Some point out that states that are passing – i.e. a broken leg – is a temporary health problem even if it is complicated and last for more than 6 months. Longstanding health problems express a state of lasting problems without the view of getting well. Many respondents point out that an illness has to be really serious to be counted as longstanding.

Some examples from the Swedish testing:
All respondents ought to have answered 'Yes' since all of them had been selected because of having some longstanding illness or longstanding health problem. But they did not! The respondents were asked to describe what longstanding illness or health problem meant to them. They were also asked of the difference between longstanding illness and longstanding health problem if they thought there was any difference. Some descriptions related to mental health were barely mentioned at all. One respondent said: "Can be anything – problems with the bowels, the heart, the kidneys and so on or with the ability to concentrate – can be anything that affects like the social life, economy, relations to family and friends, love – everything!" and another said: "Health problem means you don't feel well physically or mentally and longstanding illness is when you have got a diagnosis". Physical conditions were the only conditions people formed associations with when asked.

Time was, of course, mentioned as a criteria: "You have to be ill since many, many years" or "longstanding means several years, not only for example two months" or "something lasting more than a year, something being diagnosed" or "longstanding, then I think of my mother having Crohn's disease, but also asthma and allergy, something more or less chronic, going on for a long time. But of course if you have something making you really ill for a month or more then it's rather longstanding as well".

Another criterion mentioned was seriousness. "You have to be really, really ill and suffer a lot"; a "longstanding illness is an illness you will have for ever, getting worse and worse like Parkinson's disease, compared to health problem which is something you never get rid of, but

might not cause you a lot of suffering, for example diabetes"; "pollen allergy is something longstanding, but being allergic to some food or to horses is not longstanding, because you can avoid that – it's your own choice"; "longstanding means something you will have for ever and die of, something that will affect the rest of your life, but health problem is not that serious – you can have health problems from time to time during your life"; "longstanding illness means you need treatment for ever, it's chronic, but health problem is more like having pain and a need to take medicine for a long time, but not for ever"; "longstanding is like being paralysed, disabled, really tired, not active, but health problem is more like you have to withdraw from social life for example if you have asthma, yes asthma can be a health problem"; or "for example diabetes, angina pectoris, vascular spasm – you might be able to do something to it when you feel very bad, but you still have it – that's longstanding illness and health problem".

In all of the countries it was obvious that longstanding illnesses or health problems were under-estimated. The minimum health module was followed by a list of a number of diseases. According to the answers of these questions the prevalence of longstanding conditions are much higher than according to question 2.

The ***second prevailing version*** which is now (almost) officially decided upon:
Do you have any longstanding illness or longstanding health problem? [By longstanding I mean illnesses or health problems which have lasted or are expected to last for 6 months or more].

The sentence in brackets diverge from the same question in EU-SILC, which means that it is uncertain if it will stay. Now it is necessary for the two working groups at Eurostat to find a solution that is accepted by both groups and also able to adapt to the existing SILC-regulation.

In the background and rationale states that the general concept is self-reported longstanding illnesses and longstanding health problems. It is then up to the countries to find the best understood words in their own languages which describe the (by Eurostat) given definitions of longstanding (or chronic) and illness or health problem (or condition). The intention is to ask if people 'have' a chronic condition, not if they suffer from it.

Unfortunately it will in several countries be necessary to have a whole set of questions to assure the general concept is measured according to the definitions by Eurostat.

Question 3

The **first version** *of question 3 was the same as the one already being used in SILC.*
For at least the past 6 months, to what extent have you been limited because of a health problem in activities people usually do?

ONS recommended after cognitive testing the following wording:
Now thinking about all your health problems, not just those who are longstanding: For at least the past 6 months, to what extent have you been limited because of a health problem in activities people usually do? Would you say you have been severely limited, limited but not severely, or not limited at all?

Comments from other countries:
Lots of confusions especially about the time period of six months occurred – is there any sense in excluding longstanding or chronic conditions that will perhaps be lifelong only because they did start less than six months ago? What about conditions that might have been going on more than six months but ended (temporarily or permanently) a months ago? What about illnesses that had not spanned the entire six months period? What is supposed to have last at least the past six months, the health problem or the limitation or both?

The complexity of this question is also reflected in other difficulties respondents have when trying to answer the question. Several respondents forgot the six month reference time given in the question, as there is too much other information given at the same time – limitations + to what extent + caused by health problems + comparison in activities people usually do. Many countries also suggested to split the question into two or three parts and use follow-up questions for duration and for severity.

The term 'activities people usually do' is also troublesome for the respondents. Most of them referred to activities they themselves normally do, but those who tried to imagine what are normal activities that the general public usually do, got confused and could think of anything all people do except eat and sleep. Many related only to work activities, others only to leisure activities, and so on. Most people only thought of activities people of their own age usually do. How to measure the same way in all ages and in all countries? This is the major problem! Example: Today a young person, having had a car accident, is much more limited compared to six months ago in certain activities. He usually is an active young person (downhill skiing, mountain biking, etc.). But compared to what people usually do, he is not that very limited because people usually do not do those activities. It is rather obvious that people in a situation like that tend to answer they are limited. Another example: If you have been severely limited for a long time and your ability to do things have improved a lot, but still be severely limited compared to other people, the chance of getting an 'objectively correct' answer is probably small. The respondent will be happy to tell that he is not that limited as he used to be.

Another type of problem occurred in Finland and Estonia, countries belonging to the same language group. The question became very long and got a complicated sentence structure, which made it difficult for the interviewers to read the question aloud. In Estonian all of the interviewers mentioned that the question was too difficult to read aloud. Many countries had translation problems, especially with the term 'activities people usually do'. Norway, for example, found it hard to find acceptable expressions in Norwegian capturing the meaning of question. The

difficulty was to find a translation that secured a flowing language in Norwegian. The solution was to split the question into sub-questions.

The ***second prevailing version*** which is now officially decided upon:
For at least the past 6 months, to what extent have you been limited because of a health problem in activities people usually do?

The background and rationale to this question is very long and gives many definitions and clarifications, still not covering all difficulties.

Finally
This paper has only discussed three of more than 250 questions in the comprehensive European survey on health. We are already aware of a lot of problems, which obviously are not easy to solve. A complicating factor is the EU-regulations. Many countries need to get a detailed regulation in order to get appropriations from their governments. Other countries only want to have a regulated framework which gives the countries more "freedom" in how to conduct the survey. A compromise will probably end up in regulating almost everything except the mode of data collection, which at least in Sweden for economic reasons might result in using CATI as main mode, while most other countries will use CAPI or PAPI.

Is there a fair chance for Eurostat to succeed in launching a harmonized survey in 32 countries? Will it be possible to clarify to all respondents all around Europe how to understand the questions in order to measure all questions unambiguous despite language and culture differences?

Towards a More User-friendly Reporting System for KOSTRA

Trine Dale, Tore Notnaes and Bente Hole
Statistics Norway

Introduction

KOSTRA was established in 1995, as a web-based offline reporting solution, where municipalities could report data to the central government in Norway. All municipalities have been required to report their data through this system. There are more than 40 questionnaires on different topics and with different respondents within the municipalities in the system. In Statistics Norway there has been a myth that KOSTRA was a well functioning electronic reporting system, in spite of reports of quality problems with the data and a need for extensive revision in some of the surveys in KOSTRA. However, in 2005 evaluation and testing of one of the questionnaires uncovered severe problems with the KOSTRA solution (Dale et al, 2006; Dale and Hole, 2005). What was once a rather innovative way of data reporting, had now become old fashioned and bothersome to the respondents. In order to do their reporting tasks, the municipalities had to do a lot of extra administrative work. They had to make paper copies of the questionnaires and instructions, distribute them to the right respondents, collect the completed (parts of the) questionnaires, key the data correctly into the electronic questionnaires and return them by e-mail. Also, for some of the questionnaires a lot of people had to be involved in the response process, each responding only to a few questions. In appendix 1 we show a flow chart of the most complex response process we came across in our studies. This is how the reporting is organised in one of the larger municipalities. In addition to this, design problems led to reporting errors that a more modern electronic reporting system should be able to control for.

As a result of the rather depressing information that came out of the evaluation, Statistics Norway decided that it was time for a total modernisation of the KOSTRA system. Since some of the problems in the old KOSTRA were caused by the fact that it was an off-line system that had to be uploaded in the municipalities and required a lot of administration, it was decided that the new reporting system should be online. The change of platform made it necessary to change the design and layout of the portal. This also led to some layout changes for the content of the reporting solution, even if this was not the main objective at this stage of the revision process. Since the 2005-evaluation uncovered problems in the question- and questionnaire design, as well as in the information flow and structure (Dale and Hole, 2005), improvements are required in these areas as well.

The structure of this paper will be as follows: first we will go through the main parts of the online reporting solution for KOSTRA, presenting and describing some of the main screens, then we will present and comment on some of the most important test results. In this first step of the redesign-process, focus has mainly been on getting online, as well as on the design of the portal and the layout of the web pages – not on the content and design of the questionnaires. We are, however, aware that visual design of questions and questionnaires are important factors that influence on nonresponse and measurement quality (Dillman, 2007). In this matter we are fully in line with other researchers who have emphasised the importance of error prevention through improved instruments rather than error correction (Willimack, Lyberg, Martin, Japec, and Whitridge, 2004). Design issues have long been ignored in business surveys and it is not always easy to get acceptance for the necessity of changing established routines and practices. Since the

tasks we require the respondents to solve in business surveys often are demanding and challenging, it is however even more important to use available tools to help them solve these tasks (Cox and Chinnappa, 1995; Willimack and Nichols, 2002).

The Redesign Process

Since some of the problems uncovered in the 2005-evaluation were rather severe, it was important to get a new system up and running as soon as possible. Time was limited, so a two-step process was chosen. The main objective in the first step would be to get online before the 2006 reporting, which started in January 2007. This to, hopefully, ease the administration process and perhaps also reduce the number of people that have to be involved in the reporting. In order to do this, it would be necessary to settle for only a minimum of functionality and focus on getting the layout of a more permanent solution in place. The main infrastructure should be developed at this stage, but it had to be possible to make future improvements. More functionality and redesign of questionnaires and instruction s were postponed until a later stage. The plan was to have a more permanent solution in place for the 2007-reporting.

IT-developers and survey methodologists teamed up and started working on the design. As time was limited, all requests and wishes could not be met. However, it turned out that it was possible to develop more functionality than what was expected at the start of the process. The most important feature was a forwarding function for questionnaires. This function can be used to administer the reporting when multiple respondents are involved in one questionnaire or when there are many institutions reporting the same data in separate questionnaires.
Before changing the platform or introducing a new mode for data collection in the field, testing is always recommended (Dillmann, 2007). Unfortunately, there was no time to test the new online solution on actual respondents before it went into the field. This testing was therefore conducted after the reporting was completed in January/February of 2007 with three municipalities with different population size[21]. Since the TPs had already used the online version, the setting became a combination of testing and debriefing. Even so, for simplicity reasons, we will refer to the encounters as tests. The TPs were invited to our lab, and one particular questionnaire was used as a starting point or case. This was necessary in order to select test persons (TP) and for them to be able to actually use the solution. The questionnaire was about day care facilities, and multiple day care centres in each municipality had to report on this questionnaire. Each test lasted two and a half to three hours, and the tests concentrated on the organisation of the reporting and the TP's experiences with the online solution. One moderator and two observers were present at each test, in addition to the TP. The tests differed from "normal" usability testing, as the moderator and the TP talked about the solution while looking at it and trying out different possible actions to see what happened. There were no tasks or assignments to solve as is normal in these settings. The debriefing nature of the testing led to focussing more on the TP's experiences while reporting – what had they done, which functionality had they used, what did they liked and disliked about the solution etc.

The Online Solution

The main objectives for designing a new, online reporting platform for KOSTRA, was to improve data quality and reduce the response burden for the municipalities by simplifying the reporting process. Quality problems and actual and perceived response burden are often

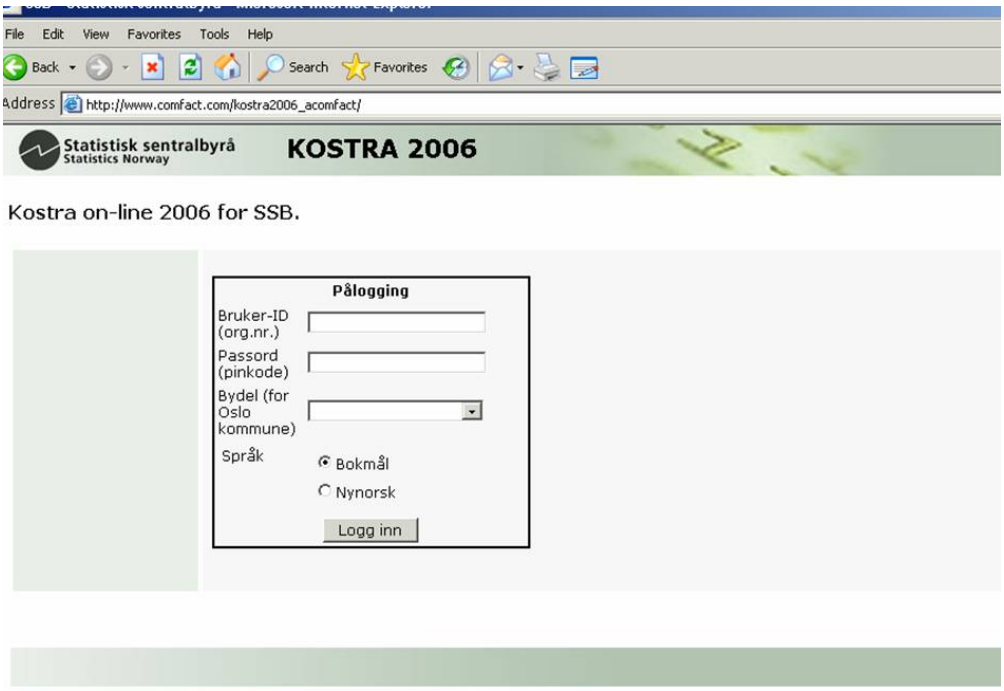
[21] We would have liked to do more tests, but unfortunately we lacked both time and money to do so.

correlated in the way that poor design and bad questionnaires lead to a higher perception of burden by the respondents and vice versa - a high perception of burden may lead to lower quality in the reported data (Dale et al, 2007; Haraldsen and Jones, 2007; Hedlin et al, 2005; Jones et al, 2005). As we have already mentioned, some municipalities used to have very complex administration procedures in the "old" KOSTRA (see appendix 1). With the new online platform the need for administration should be reduced radically for these municipalities. Some administration will still be required, but our testing and knowledge of the online solution suggests that three levels or less should now be sufficient: KOSTRA-coordinator, questionnaire coordinator and respondent (see flowchart in appendix 2).

Logon page

The logon page (figure 1) of the online solution is kept clean and simple. The respondents are asked to log on, using the user-ID, which is the organisation number of the municipality, as well as the password or pin code they received in the introduction letter. There is an extra response field for Oslo, which is organised in a special way that differs from other municipalities. Norway has two official written languages and the respondents are able to choose which one they want to use – bokmaal or nynorsk. In the offline solution, you had to scroll down on a page crammed with information in order to be able to install the questionnaires on your computer (appendix 3), something that caused several problems (Dale et al, 2005; Dale and Hole, 2006). Also, only a few people in the municipalities had access to the electronic version and opportunity to upload it, something that caused problems if these persons were not available. The simple access in this new system is therefore a major improvement.

Figure 1: Logon page



The logon page of the online solution was perceived as intuitive and easy to understand and use by the TPs. However, the extra response field for Oslo was commented upon by the TPs, and

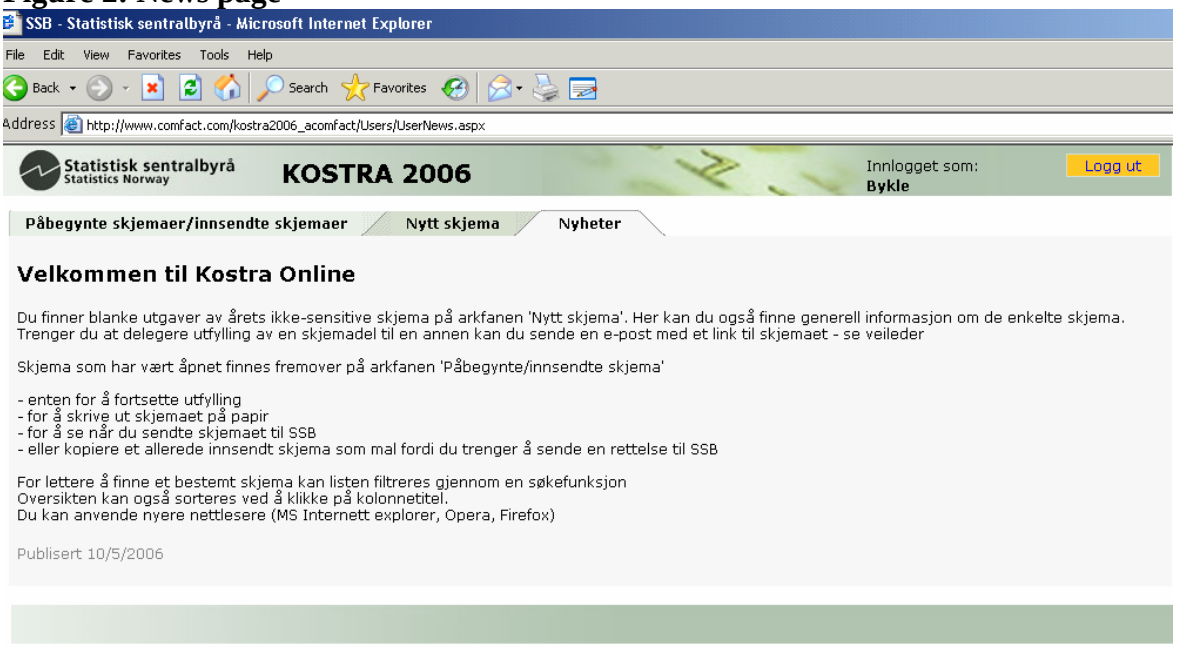
perceived as noice by at least one TP. Even if it was a source of irritation, it did not seem to make the task harder for them.

The layout of the screens can be described as follows: the pages and headings are coloured in shades of pastel green, while the answering boxes are white and stand out against the background. This is according to Dillman's design principles (Dillman 2007), and is meant to help the respondent to navigate in the questionnaire and to respond to the questions. Once you are logged on to the solution, the heading of each page will always show which municipality you are logged on for. In the examples used in this paper, the names of the municipalities are Bykle or Oppegaard (top, right). Once you get into the different questionnaires, the number and name of the questionnaire will also be listed on top of each page (see figure 4). The name of the municipality assures the respondent that she/he is reporting for the correct municipality. The name and number of the questionnaire is important because some respondents report on behalf of several institutions and on several different questionnaires within the KOSTRA solution. This information may help prevent erroneous reporting. It is not possible to logon for another municipality unless you manage to get their password. There are also other features to help the respondents complete their tasks, but these will be described later.

News page

The first time you log on you go directly to a page labelled "News" (Nyheter) (fan 2, figure 2). On this page you get a welcoming message, in addition to some general information. News will also be posted here. The next time you log on you go directly to the page labelled "Started questionnaires/Sent questionnaires" (Påbegynte skjemaer/innsendte skjemaer – fan 1). Once a questionnaire has been opened it will be placed here. The third fan is called "New questionnaire" (Nytt skjema). On this page you will find all new questionnaires that have not yet been opened or worked on.

Figure 2: News page

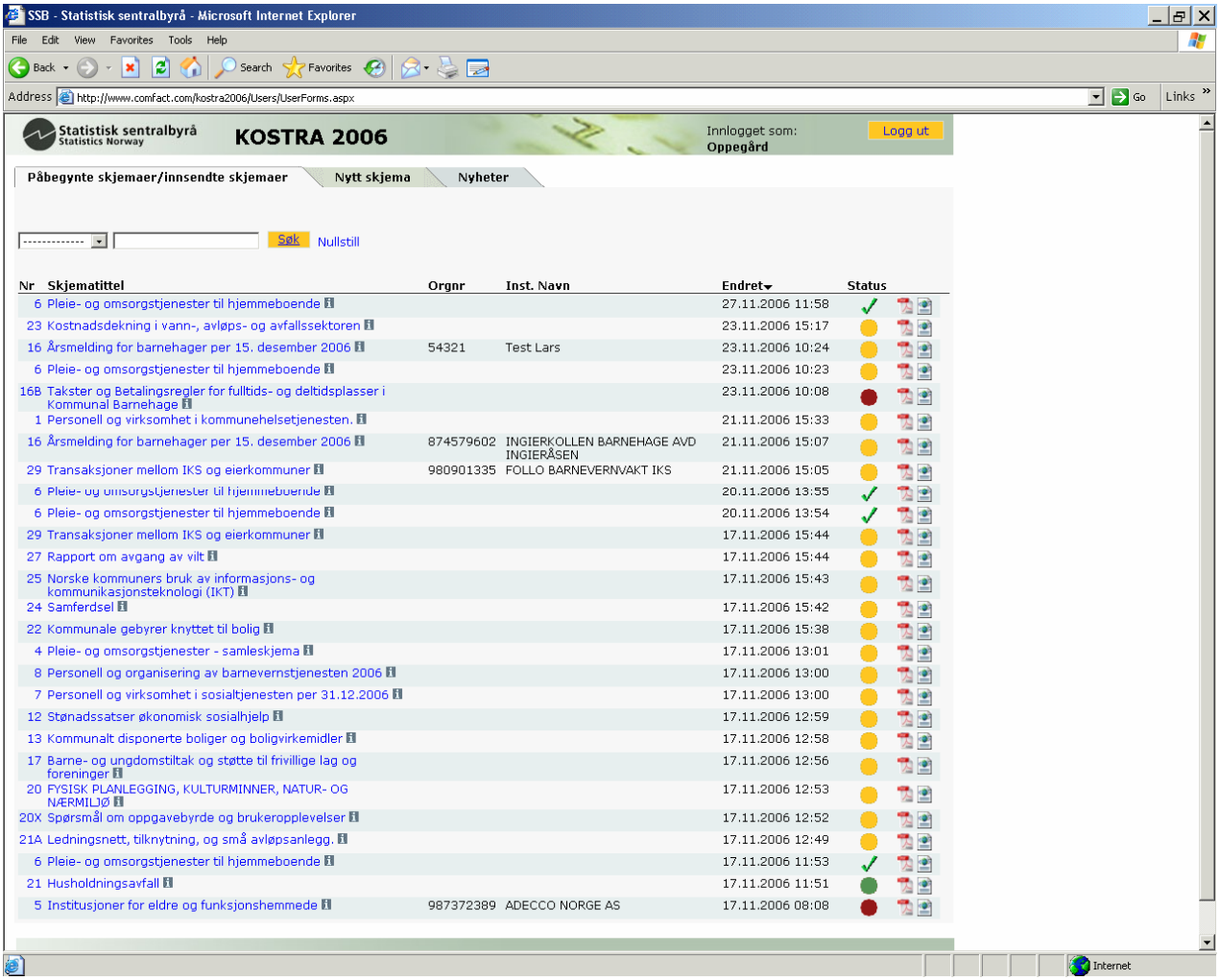


The TPs read the content of this page, and the alternation between the fans worked well. They seemed to understand what was behind the different fans.

Status page

The status page for questionnaires you have opened, started working on or already sent in is quite complex and contains a lot of information (figure 3). On the top, right, there is a search function – a drop-list where you are required to filter your search and a search field where you specify the search item. In the main part of the page there are eight columns. The column labels have a built in sorting function. The filter-function in the search field is based on these column labels.

Figure 3: Status page



On the far left the numbers of the questionnaires are listed. In KOSTRA this is important information to the respondents, as they often refer to the questionnaires by number instead of title. In the next column are the titles of the questionnaires, with an information icon following directly after the text. Column three lists the organisation number of the municipality or the institution in question. In column five the name of the institution the reporting should be done for

is listed. Date and time for the last change/action in the questionnaire is given in column six, and in the last column with at heading (column seven), traffic light symbols show the status of the questionnaires. These were developed and included to make it easier for the KOSTRA- and questionnaire coordinators to keep track and stay on top of the reporting process. Yellow means that questionnaire has not been controlled ant that it contains possible errors, red means that the questionnaire has been controlled and contains errors, green means that the questionnaire has been controlled and is without errors,
while the green tick means that the questionnaire has been controlled and returned faultless. An exclamation mark (not shown in this example) means that the questionnaire has been returned but that it contains errors.

To the far right of the screen are icons that allow the respondents to get pdf and xml versions of the questionnaires. The pdf is included to make it easier to print the questionnaires, as we know from earlier tests that many municipalities use prints of the electronic questionnaires when collecting data (Dale and Hole, 2005). The xml version contains, among other factors, a summing function for questionnaires that have to be filled in by many institutions in the municipalities (like day-care centres). This function will provide aggregated data for the municipality.

We will not go into detail on all the findings on this page. However, the tests showed that it will be necessary to do some adjustments on the status page. For instance, the TPs had not understood that the column labels had built-in sort functions. Even so they would like an additional function to sort search results after using the search function. None of the TPs had used the search function - one of them had not seen it at all. Another TP had problems using it, not understanding that the search had to be filtered. To filter a search you need to specify a search area in drop list (number or title of questionnaire, organisation number, name of institution etc) and then narrow the search by specifying for instance the name of an institution, the number of a questionnaire etc. This function needs to be made more visible and its use more intuitive.

When clicking the information icon, you are transferred to the complete instruction manual for KOSTRA. The information found was too detailed at this stage. The TPs would like only general information, otherwise there would be too much to read. The status symbols did not work that well, as they were not intuitive enough. The main reason for this was that the colours were not used in a logical way. For instance, the colour for questionnaires that had been returned containing errors was green with an exclamation mark. One TP thought it should have been orange, since the questionnaire was not accepted. Two TPs suggested smiley-symbols instead of the symbols used. The pdf and xml icons were not intuitive either, and as we have commented on before, the municipalities do like to have a printed version of the questionnaires. They found the pdf-printout of the questionnaire to be very messy and complex. None of the TPs had found the summing function for day-care centres, which would have provided aggregated key numbers from all day-care centres in the municipalities and thus have been very helpful for them. So, even if this page is an improvement from the off-line version, there is still room for vast improvements. It is important that the design and functions are intuitive and easily understood by respondents.

Questionnaire

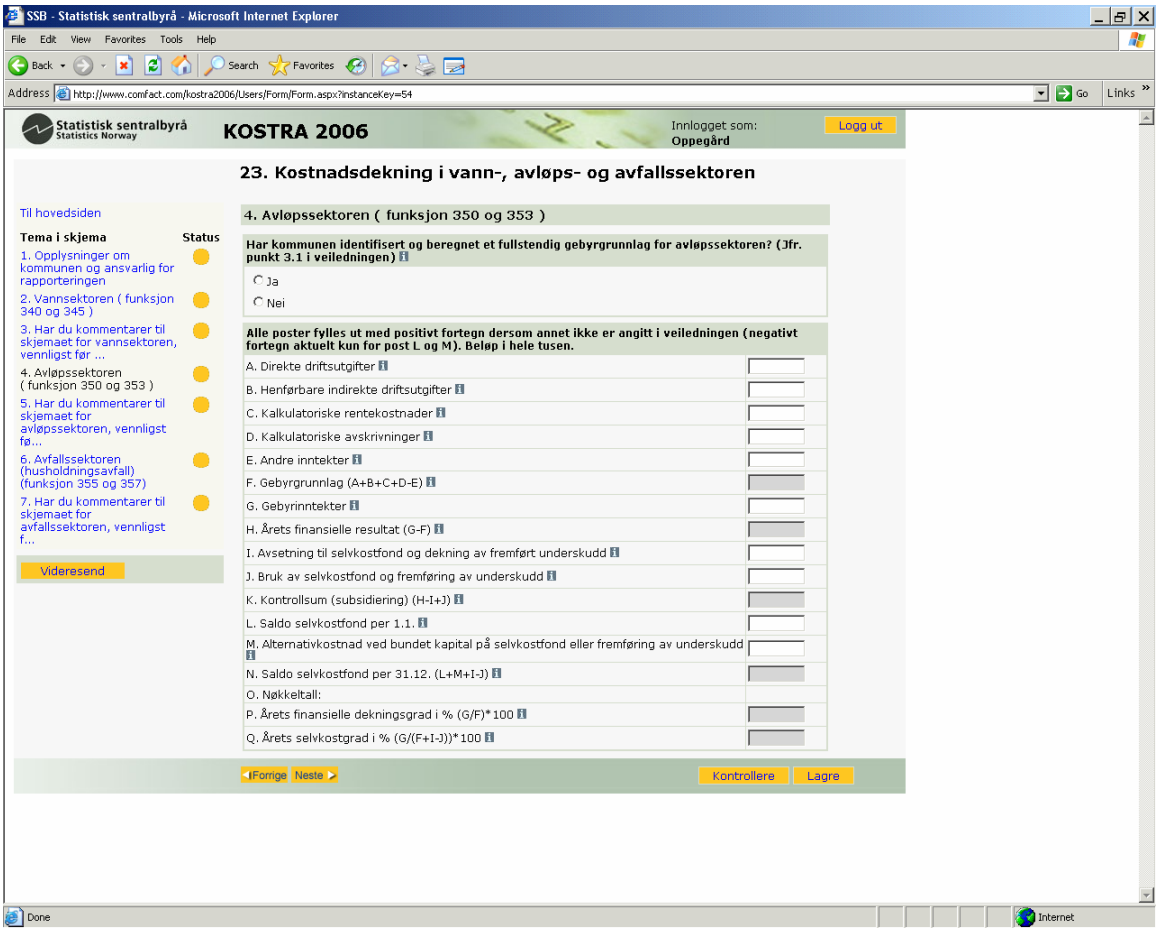
A typical questionnaire page in KOSTRA can be seen in figure 4. Even if we have not worked on the content and layout of the questionnaire directly, some changes have been made in some

questionnaires due to the evaluation report from 2005. For instance, the numbering policy has been improved and made more intuitive. In the old version, the numbering of the first question in a new section all started on 1. In this example, only the themes have numbers. Sub-questions are numbered by using letters. The numbering practise is still not coherent across questionnaires and normally questions are not numbered continuously. Dillman (2007) recommends consecutively and simple numbering from beginning to end in a questionnaire. As this is the most logical way of numbering questions it will also be the format respondents can most easily relate to.

Some changes were necessary do to the transition to an online platform – both in layout and in functionality. Whereas the offline questionnaires consisted of one, long scrollable page, the new online questionnaires are divided into pages according to different themes. The pages are divided into three different fields: on the left a status field, the main part of the screen consists of the questionnaire itself, and on the bottom a navigation field. On the bottom left are the previous and next buttons, and on the right a control button that allows the respondent to check for errors and a save button. On the far left, at the bottom of the status field, there is a forward button, which function will be described in more detail later. In the status field, the same traffic-light symbols that were described on the status page are used. The meaning of the symbols is, however, a bit different once you are working within a questionnaire. Yellow means that the questions have not been controlled (possible errors), red means that the questions are controlled and contain errors, and green means that the questions have been controlled and are without errors.

As has been mentioned before, the number and title of the questionnaire is on the top of each page. In many questionnaires the title has an information icon next to it (not this example).

Figure 4: Questionnaire



When you click on the icon you now go to the main instructions for the questionnaire. In the future the plan is to have more tailored instructions, at least on theme level. This was not doable in this first version.

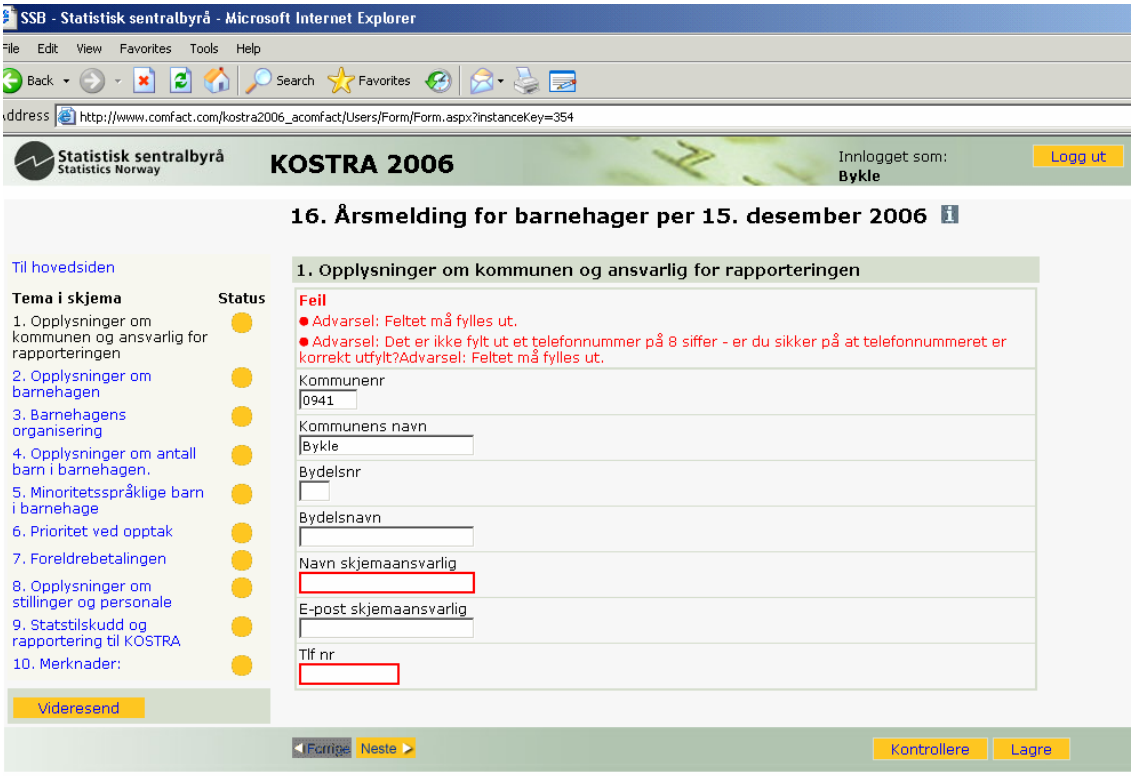
The tests showed that the TPs would like tailored instructions, preferable on a question level. The information icon was not intuitive enough, as this is not the symbol normally used for instructions in surveys at Statistics Norway. This information button should be named "veiledning" (instruction). One TP said that he expected go get instructions to the questions on this page, not general instructions to the KOSTRA system. That information should be on the main page or on the KOSTRA pages on ssb.no. Two of the TPs pointed out that the status field on the left should be visible at all times. It now disappears when you scroll down the page. Design-wise and with regard to response quality, the best would probably be to avoid scrolling altogether. The forward button was not intuitive enough, and the TPs did not understand what it was for. They thought that they would return the questionnaire to Statistics Norway if they clicked this button. As we will see later, this is not the case.

Error messages

When the control button is used, error messages are displayed in red writing. The response fields containing errors get red frames (figure 5). This is done to help the respondents find and correct

the errors. If the control uncovered errors in one or more parts of the questionnaire, the status symbols in the left column should turn red. If they are green, everything is ok. The fact that they are yellow in this example should indicate that the questionnaire has not been checked. This screen was taken from a test-version of the program and didn't come out quite right, as can be seen.

Figure 5: Error messages



The TPs liked the fact that the response fields containing errors were easy to find. However, they often had problems understanding the error messages, as they did not contain enough information to identify and correct errors. Some of these problems were due to programming errors. Research shows that there are three important principles to take into account when designing error messages. It is important to identify where the error is, what the error is as well as how to correct it (Haraldsen, 2004). These principles are not met in KOSTRA.

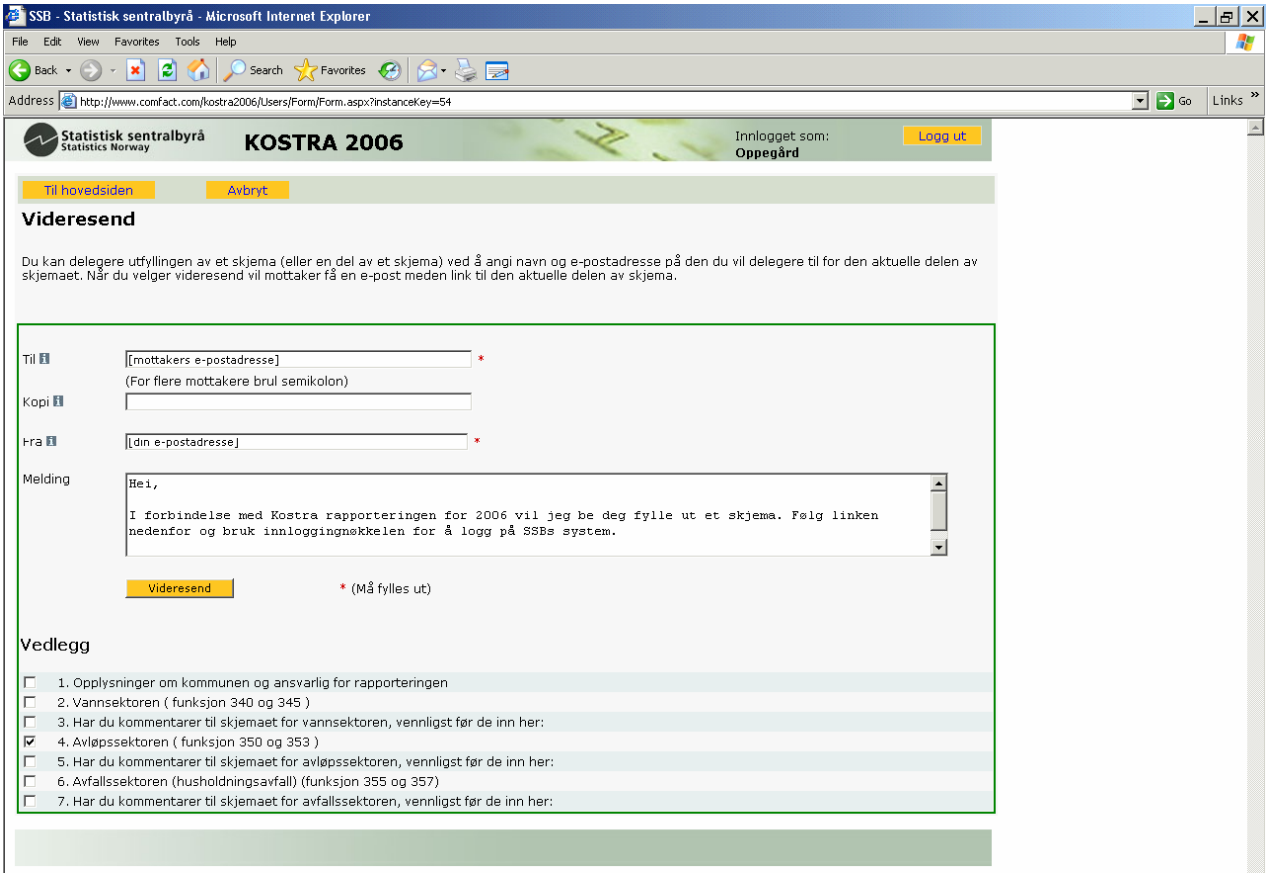
Forwarding function

The forwarding function is a new function, designed especially for KOSTRA. As we have mentioned previously, many KOSTRA questionnaires have multiple respondents, and in some surveys questionnaires must be distributed and collected from many respondents. This was causing a lot of administrative work in the offline solution of KOSTRA, and is likely to have contributed to a perception of higher response burden than what would otherwise be the case. It has not been not possible for most municipalities to do their reporting electronically because of these facts and most of them therefore collected the data from different respondents on paper copies (often bad) of the questionnaires. At Statistics Norway we wanted the online solution to

facilitate electronic reporting at all levels, from all respondents, as this would ease the administration part of the reporting and mean fewer steps and fewer people involved in each step of the reporting process. This again, would hopefully reduce the risk of errors occurring. To deal with this, the forwarding function was developed.

The forwarding function allows the KOSTRA or questionnaire coordinator to delegate questionnaires or parts of questionnaires to other people who are the actual data providers. This is done by e-mailing the link to the questionnaires or questions to the people in question,

Figure 6: Forwarding page



along with a password that allows them to log on. The forwarding page is displayed in figure 6. As in regular e-mails, it is possible to write a message, telling the recipient what you wish her/him to do. The recipient(s) of this e-mail may then log on and fill in the necessary information. They will only get access to the questionnaires/questions/themes the coordinator has specified in the forwarding function. The rest of the questionnaire or other questionnaires remain inaccessible. After the recipient has filled in the necessary information, she/he is supposed to press the send in button and thus return the data to the coordinator. The data is automatically included in the "host" questionnaire, and the information can be seen by the coordinator in charge.

The TPs liked this function when they understood what it can do. They had not all understood what it was for when they did their reporting. Only one TP had actually used it, forwarding questionnaires to day-care centres. However, the tests uncovered that there is much to gain on improving the design of the page. This function and page was not subject to any influence from or testing by experts in questionnaire design or users, as it was developed just before KOSTRA was sent into the field. This is evident both in the layout and in the choice of terms and labels on the functions. These are not consistent with the rest of the solution.

The green frame around the content should be removed, as it is not used anywhere else in the solution. This frame divides the questionnaire part of the page from the instruction part on top, and from a design-point of view this is unfortunate. The button labelled "videresend" (forward) should be moved to the bottom of the page and the send in button should be visible at all times.

The name of the page itself, the name on the button that got you there from the questionnaire and on the button that actually allows you to forward (parts of) the questionnaire are identical. The TPs found this confusing. Also, they did not like the term used – "videresend" which means forward. The TPs thought it meant that the questionnaire would be returned to Statistics Norway. They suggested that the page and function should be labelled "delegere" (delegate) instead. One of the TPs also did not like that the term "vedlegg" (attachment) was used on this page, as he understood attachments to be something other than or outside the questionnaires. He therefore suggested that they should mark the *themes* or *sections* they wanted to delegate, not the *attachments*. The TPs would like to be able to use a local address book for this function, as that would save them a lot of typing. Sometimes the questionnaires have to be sent to many respondents, access to local address books would therefore make the delegating job a lot easier.

Delegated questionnaires

The respondents get access to their delegated questionnaires/questions by using the link and a password given to them by e-mail from the coordinator. These e-mails are labelled with the name of the software producer, which is not known to the respondents. Some of the recipients thought the e-mails were spam and did not open them. Changing the e-mail heading is therefore imperative to avoid such misunderstandings.

Many respondents receiving their questionnaires by e-mail did not understand that the send in button would send the completed questionnaires/questions back to the coordinator. They believed that the questionnaire would be sent directly to Statistics Norway if they clicked this button and consequently did not dare to use it. This button therefore needs to be more

Figur 7: Logon for delegated questions

intuitive, making it clear that the questionnaire will go back to the questionnaire coordinator. This can easily be done by naming it "Return to KOSTRA coordinator" or something similar.

Figure 7 shows the logon-page for respondents that have received delegated tasks. It is simpler than the regular logon-page, and only asks for a password. The respondent is allowed to choose between the two official written languages in Norway and some links to help-functions are also provided. In figure 8 part of a delegated questionnaire is shown. In the status-field on the left we can see that two themes in this questionnaire have been delegated to this respondent – number 4 and 6. As we have already explained, the respondent can only access these to themes or sections. The rest of the questionnaire cannot be accessed or seen.

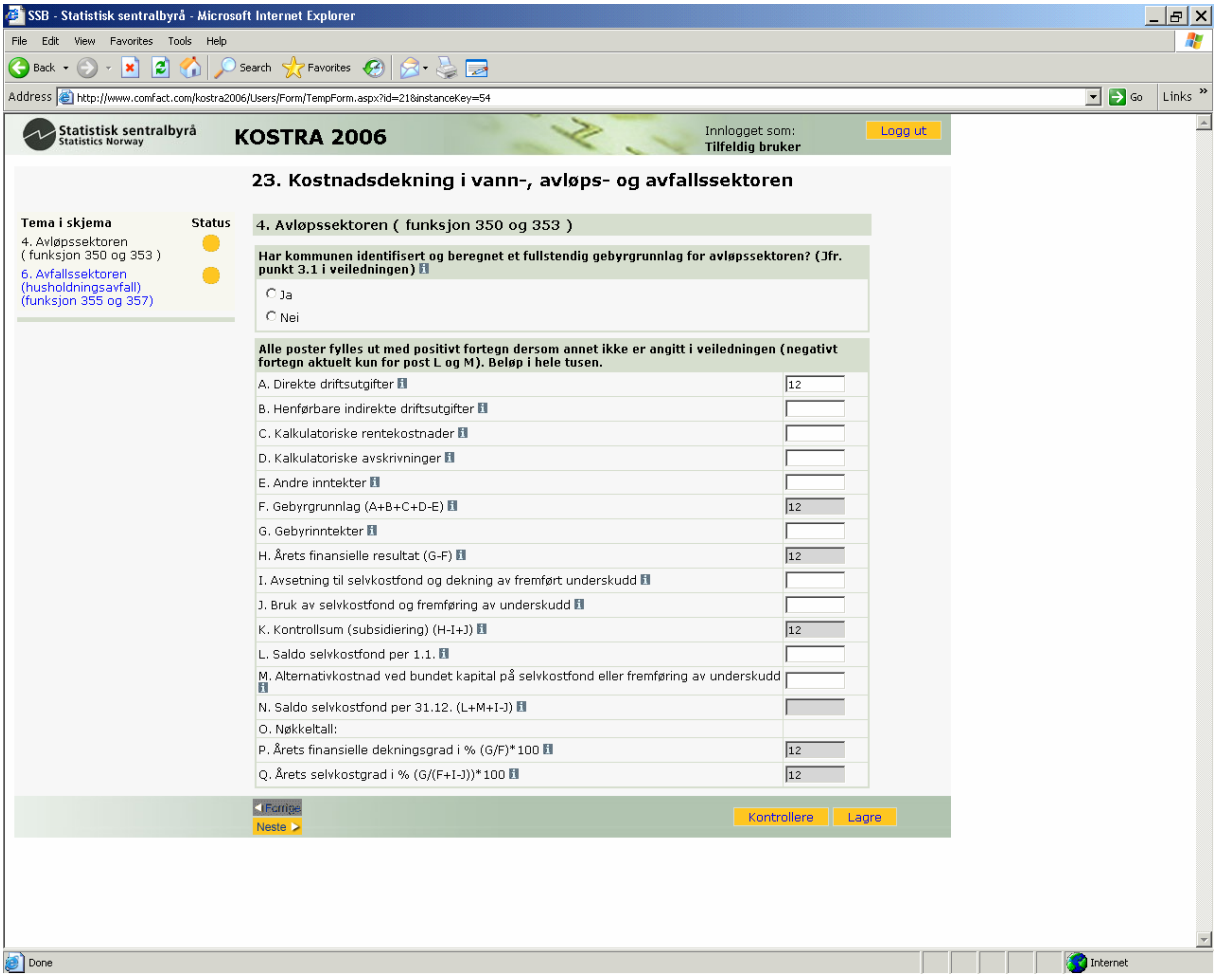
If the forwarding function is to be a success in KOSTRA as well as in other types of business surveys, it must meet the needs of both respondents and coordinators. It must be intuitive and usable on both ends of this process – for the persons delegating tasks (coordinators) as well as for the persons on the recipient end (respondents). The preliminary testing that has been conducted here suggests that most of the challenges are now on the respondent side, even if accessibility and design also has to be improved for the coordinators.

Conclusion

It is important to notice that even if the online reporting solution was available and functioning in the 2006-reporting, the offline solution was still the recommended option. There was no promotion of the online-solution. It was introduced as an option, but nothing more. Even so, about one third of the municipalities chose the online solution. This is a strong indication that there was a need for a new reporting solution. The feedback we got supports this, as the users were generally very positive. Another factor that might be interpreted as support for this assumption was that even though a new system was introduced, there were fewer calls for assistance than there had normally been.

Throughout the debriefing and testing three problem areas have become evident. First, there is what we may call the information problem. Second, there is the administration problem, and third there is the design problem.

Figur 8: Questionnaire for respondent with delegated questions



The information problem

Both the 2005 evaluation and our debriefing/testing in 2007 uncovered problems with the way the information material in KOSTRA has been organised and the information flow. The main problems seem to be that there is no paramount information policy that considers the needs of both the informer and the receiver, that the information is not tailored to different actors in the response process, and that instructions are not tailored to the contents (questions) of the questionnaires. For instance, we have uncovered problems both in the content and the location of instructions - which should be provided where they are needed and be as to the point as possible. The error messages are not specific enough to guide the respondents in correcting the errors, and as we have pointed out earlier the questions where, what and how should be answered. In 2005 we also found that there was confusion about who provided which information as there were three official suppliers of information (Dale and Hole, 2005). The information content has not yet been adjusted and redesigned to fit the new online solution, and this is probably the most pressing

task now. For instance, question specific instructions should be tailored on the question level or at least on theme level, thus making them more accessible to the respondents.

The administration problem

As we have shown, administration of the reporting process will still be a challenge in the online solution. The delegation or forwarding function can, when improved and corrected, ease the administration process. It will, however, not make the need for administration disappear. That would mean a total redesign of the content of the KOSTRA reporting system, splitting up questionnaires according to how the municipalities are organised. This would not be easy, as the organisation varies quite a lot between municipalities (Dale and Hole, 200). The TPs in this project seemed to appreciate the idea of a forwarding function, even if some of them were insecure of how to use it before it was explained to them. They also believed that this functionality would help introduce actual online reporting. Access to a good paper version of the questionnaires will, however, still be essential – as the TPs expressed quite clearly.

The design problem

The focus has now been on designing the portal, both when it comes to layout and functionality. Even so, our findings have shown that there is still a need for adjustments and improvements – both when it comes to layout and functionality. On the layout side, more testing and experimentation is required to find good graphic symbols and to use colour symbols in an effective and informative way. When it comes to functionality and interactivity, much work remains. We still haven't started working on the content and layout of the questions and questionnaires, and a redesign and development process is badly needed. Most of the questionnaires are still forms consisting of statements and key-words instead of questions. Dillman (2007) recommends using a question format in business surveys, just like in social surveys, as this makes the response process easier for the respondents. In his principles for good questionnaire design, he also recommends a clean and simple layout following consistent "rules" and using tools and effects in consistent ways. That is not the case in KOSTRA yet. For instance, the vast use of frames and visible lines on the questionnaire screens is unfortunate and should be avoided. There is also a need to find better ways to handle large matrixes and thus minimize or avoid scrolling (at least horizontal).

Even if the online solution is not optimal and adjustments and improvements are required on design, functionality and content, it is a big step forward in the right direction. When improvements are incorporated, the online solution will provide a much better tool for both administrators and respondents at different levels. Once the online solution has been tried and used by most municipalities, we also hope for actual online reporting. This will contribute to reducing the likelihood of errors occurring, as fewer people will be involved in each step of the reporting process. Roles will still be important in KOSTRA, as there will still be a need for coordinators at different levels. With the new online solution, however, we hope their job will be less complicated and easier to administer.

References

Cox, B.G. and Chinnappa, B.N. (1995). *Unique Features of Business Surveys*. In Business Survey Methods, B. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds). New York: Wiley, 1 - 17.

Dale, T. (ed), Erikson, J., Fosen, J., Haraldsen, G. (ed), Jones, J., and Kleven, Ø. (2007) *Handbook for Monitoring and Evaluationg Business Survey Response Burdens*. Eurostat

Dale, T. and Hole, B. (2005) *Evaluation of electronic questionnaires in KOSTRA. Case: Questionnaire no. 20 – Land use planning, cultural heritage, nature and local environment*. Statistics Norway. In Norwegian

Dale, T., Hole, B. and Høie, H. (2006) *It Takes More than a Computer to Respond to Electronic Questionnaires. A case study*. Paper presented at Q2006, Cardiff.

Dillman, D. (2007). *Mail and Internet Surveys. The Tailored Design Method. Second Edition. 2007 update with new Internet, Visual and Mixed-mode guide.* Wiley.

Haraldsen, G. and Jones, J. (2007). *Web and Paper Questionnaires seen from the Business Respondent's Perspective.* Paper presented at the Third International Conferens for Establishments Surveys, Montréal 2007

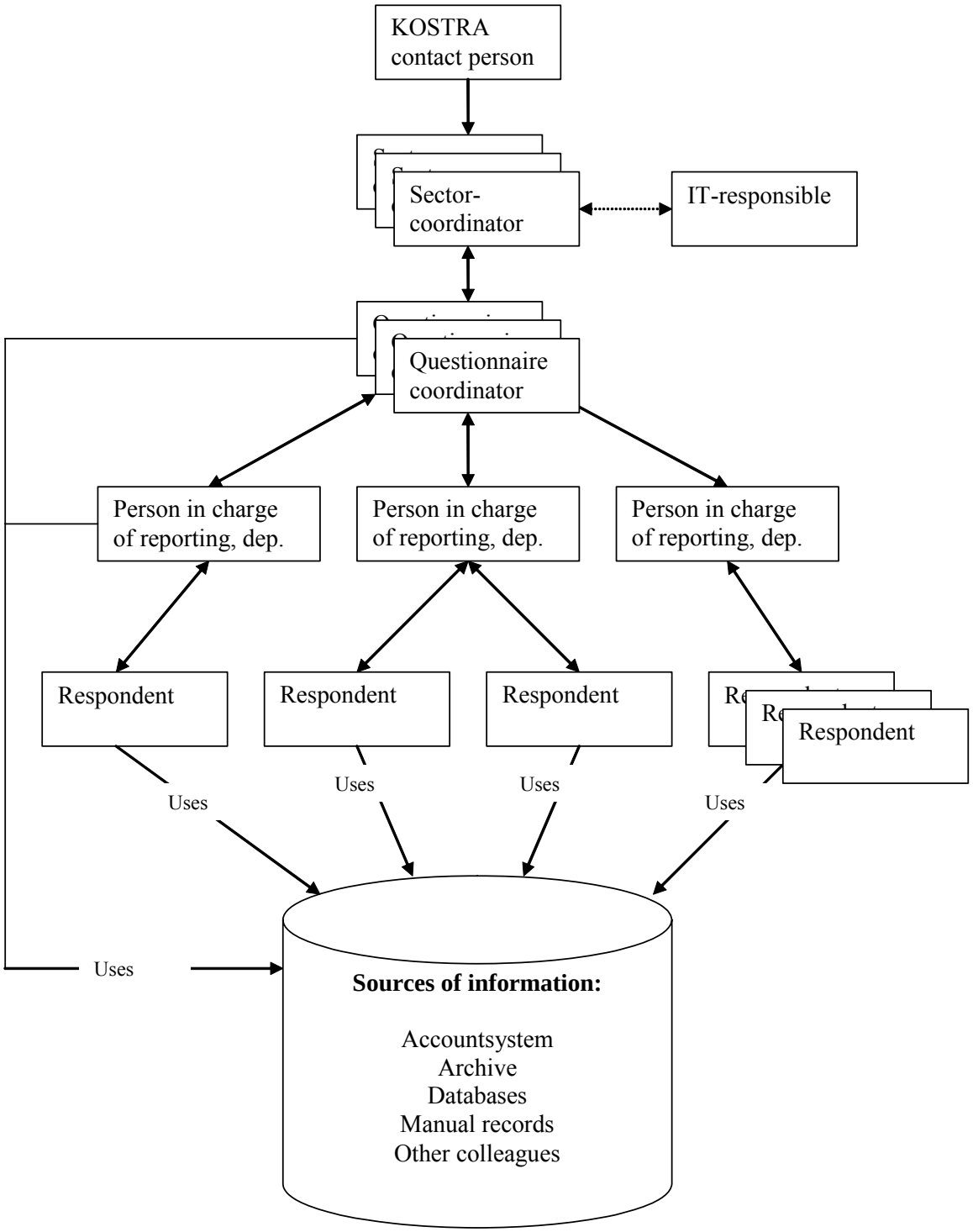
Haraldsen, G. (2004). *Guidelines for developing and designing web-questionnaires.* Håndbøker (Handbooks) 2004/81, Statistics Norway. In Norwegian.

Hedlin, D., Dale, T., Haraldsen, G. and Jones, J. (2005). *Methods for Assessing Perceived Response Burden.* February 2005

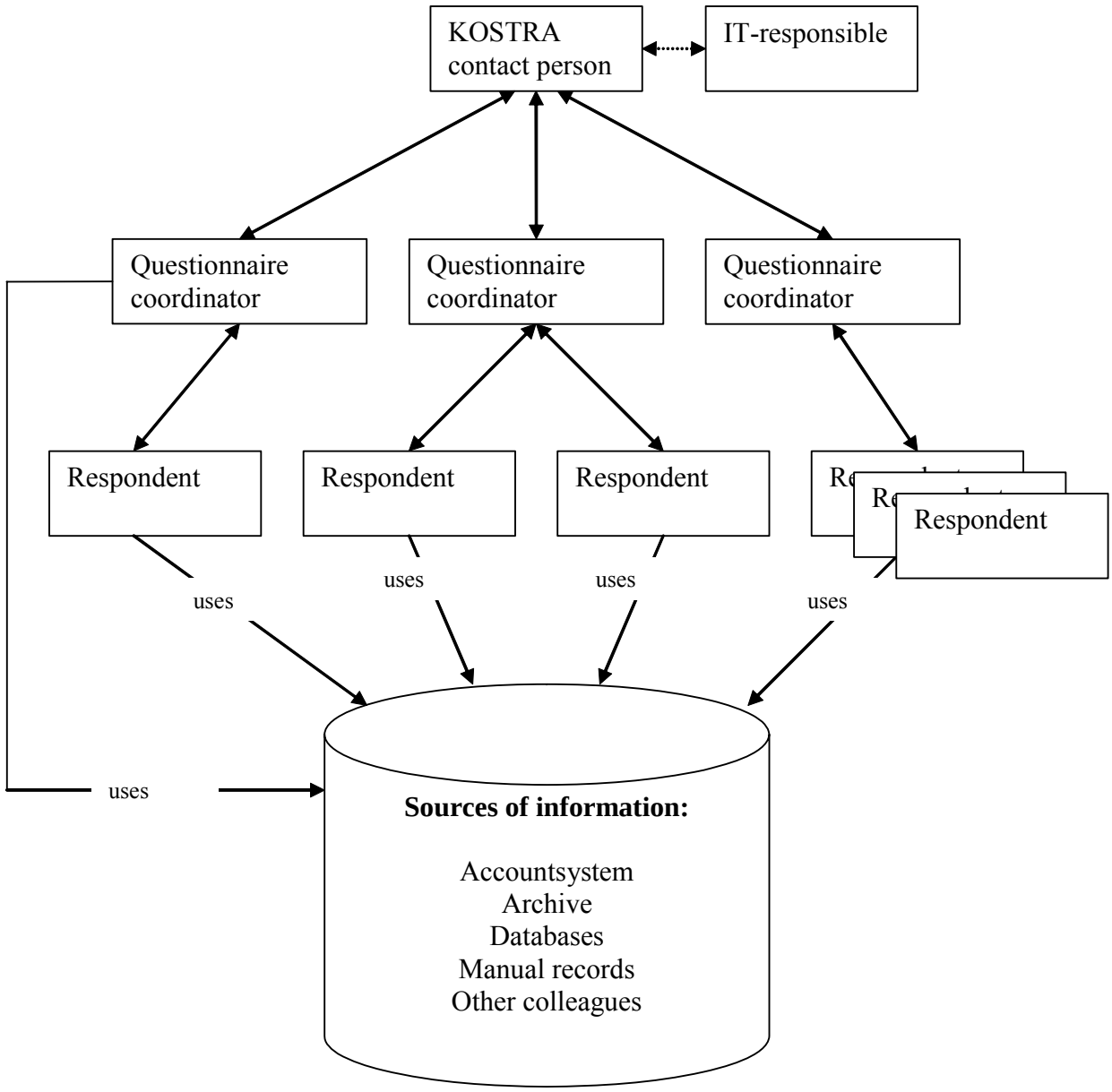
Jones et al. (2005) Conceptualising Total Business Survey Burden, Survey Methodology Bulletin, UK Office for National Statistics, No. 55 pp. 1 – 10.

Willimack, D.K., Lyberg, L.E., Martin, J., Japec, L., and Whitridge, P. (2004). Evolution and Adaptation of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys, Chapter 19. In Methods for Testing and Evaluating Survey Questionnaires, S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer (eds). New York: John Wiley and Sons, Inc.

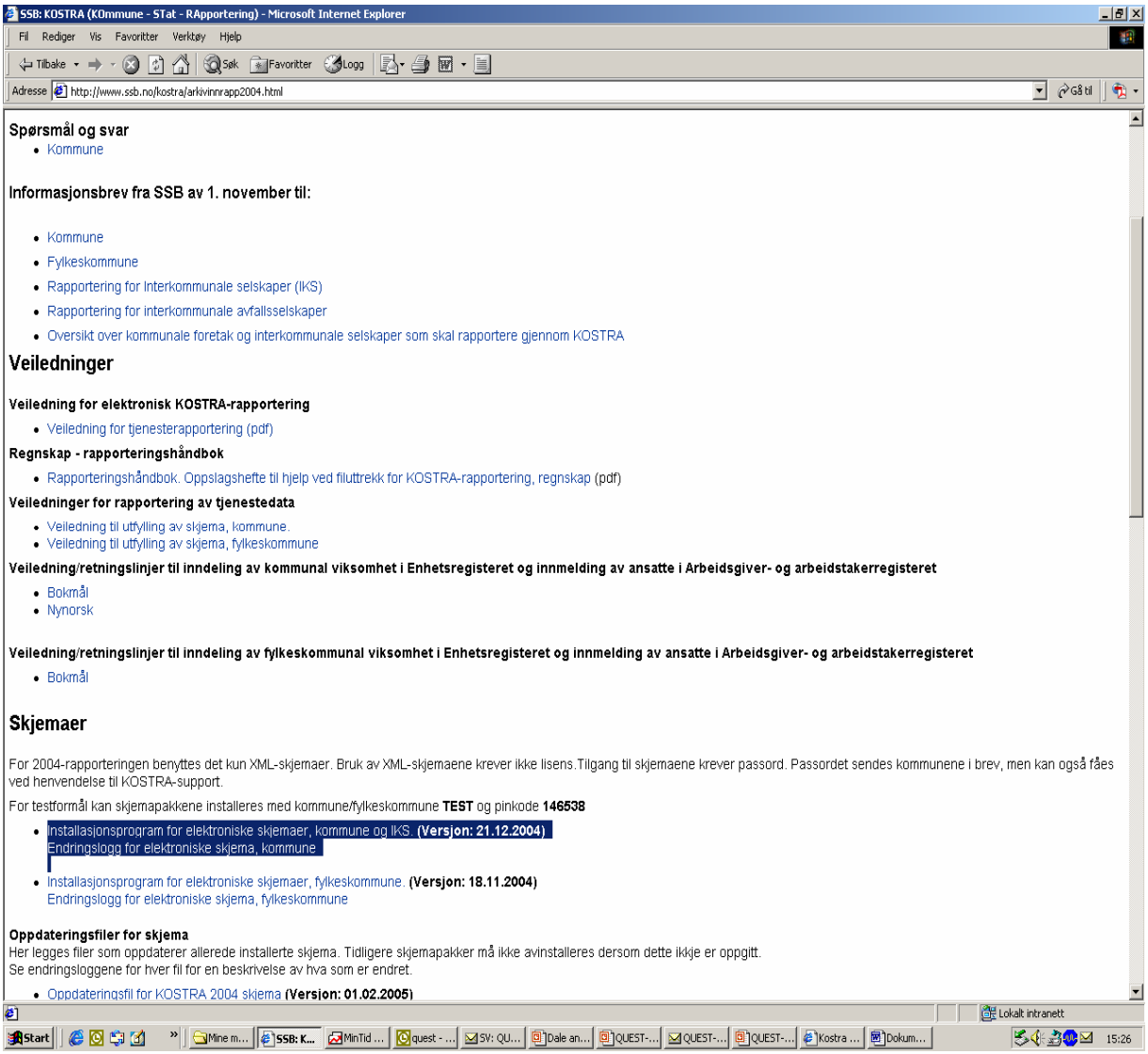
Appendix 1: Flow chart over response process in large municipality



Appendix 2: Flow chart over response process in small municipality



Appendix 3: Introduction page in KOSTRA – offline



The Usability of a Website Evaluated by Survey Methodologists

Bente Hole, Tore Notnaes and Gustav Haraldsen
Statistics Norway

Introduction

As survey methodologists we are normally engaged in evaluating and testing questions and questionnaires, on paper and on web. During the last year or so, we have had the opportunity to convey the knowledge we have on questionnaire testing and development to a slightly different, but related area, namely web design. Through several projects run by Statistics Norway's Department of Communication we have been involved in evaluating and testing some web pages and web applications, among them Statistics Norway's home page ssb.no. This was a quite new and interesting experience for us. After the testing was finalised we tried to figure out what we, as survey methodologists, can contribute with when it comes to web design and development that differs from the skills and insight that are offered by people with more traditional IT and usability competence. The following is an attempt to sum up some of what we found.

Tools used for question and questionnaire evaluation

When we do expert reviews on questions and questionnaires, we typically use a simplified version of Forsyth's O-Questionnaire Review Coding System combined with Snijker's Expert Questionnaire Appraisal Coding System (figure 1 below). This checklist is mainly used for text and content evaluation and focuses on three main areas; comprehension problems, problems concerning the task itself and problems with response categories. Especially the part concerning comprehension can be useful, not only when reviewing questionnaires, but also if one assesses a web page or a web application. Large amounts of text and use of complex syntax can be fatal when designing a home page, as well as a questionnaire. Too few instructions and definitions, on the other hand, can turn out just as bad. It's all about finding the right balance. In general, one should keep it short, but make sure the user gets as much information as is needed in order to make use of the product in question.

When it comes to visual design in questionnaires we use our own checklist inspired by Don Dillman (figure 2). This list consists of four parts. The first section concerns the first or overall impression of the questionnaire, typically the front page. The second deals with question order. The last two parts cover navigational and visual issues, such as numbering and referencing, and use of colour and symbols, respectively. We find that especially the first and fourth of these sections can be used more or less directly when it comes to evaluating and developing web pages. Also the two remaining sections are indeed relevant and useful, though in a more figurative way. From a web design perspective, they can be said to cover topics such as grouping of information, information flow and navigation. Dillman focuses on how important it is to lead a respondent through a questionnaire. The same is equally important when it comes to web pages. If the user doesn't know where to start or how to get to the information he or she seeks, it's a lost case.

Figure 1: Simplified version of Forsyth's O-Questionnaire Review Coding System combined with Snijker's Expert Questionnaire Appraisal Coding System (text and content evaluation)

1. Problems with comprehension		2. Problems related to the task		3. Problems with response categories	
1 Insufficient text or instructions		1 Unclear task		1 Missing response labels	
2 Unclear terms		2 Abstract task		2 Unclear labels	
3 Complex syntax		3 Extensive information retrieval		3 Missing response categories	
4 Insufficient navigational instructions		4 Long or problematic retrieval period		4 Wrong units of measure	
5 Several questions in one		5 Difficult considerations		5 Too detailed or general response categories	
6 Mismatch between question and response categories		6 Difficult adaptation		6 Overlapping response categories	
7 Ambiguous references to time and space		7 Sensitive questions		7 Multidimensional response categories	
8 Other problems		8 Other problems		8 Other problems	
COMMENTS:					

Figure 2: Checklist, questionnaire and question design

A. First/overall impression of the questionnaire	1	The questionnaire is too long
	2	The questionnaire is too compact
	3	The layout is messy
	4	Too many effects used
	5	Effects that doesn't have any obvious meaning or relevance
	6	The questionnaire is without contrast
	7	Font size and/or font type is hard to read
	8	Other problems with readability
B. Question order	1	Misleading main or sub title
	2	Mismatch between title and first question
	3	The first questions do not apply to all respondents
	4	Background questions before questions on the relevant subject
	5	Unclear or disorderly topic division
	6	Unfortunate placing of sensitive questions
	7	Other problems with question order
C. Navigation help	1	Question numbering or other indication of question order missing
	2	Complicated question numbering
	3	Changes of topic not indicated clearly enough
	4	Questions not placed according to normal reading direction
	5	Inaccurate references
	6	Reference arrows and texts placed too far away from relevant response alternatives
	7	Complicated references
	8	Reference to former submitted responses
	9	Breaks from the main structure not indicated clearly enough
	10	Other problems when it comes to navigation

D. Visual message	1	Colour and contrast combinations that make important things hard to read
	2	Not consistent use of font types and sizes on different text elements
	3	Instruction texts not placed where needed
	4	Complicated symbols
	5	Symbols that are similar, but have different meaning
	6	Use of different symbols to communicate the same meaning
	7	Symbols that don't mean anything
	8	Questions, instructions and response labels that don't appear as units
	9	Response fields that are easy to overlook
	10	Lacking accordance between questions and the design of response fields
	11	Design or placing of response fields that can seem leading
	12	Other problems when it comes to visual message

What can survey methodologists contribute with?
Apart from the tools mentioned, what kind of survey methodology knowledge may come in handy as one engage in web design and development?

First and foremost we find that insight in cognitive psychology and training in how to explore the cognitive processes a user goes through as he uses a certain product (in our case, normally a paper or electronic questionnaire) is useful. We have certain rules of thumb that we follow in order to make a product such as a questionnaire work: for instance one should expose the user to one task at the time, put things in a – to the user – logical order and group information in a sensible and comprehensible manner. The user should be led through the questionnaire in an easy, clear and consistent way. Measures such as these help by reducing the burden on the user's working memory and will ideally contribute to a good user experience by adjusting the product to the way a typical user thinks and reasons. These rules also apply to web pages.

As survey methodologists we tend to have a strong focus on text, the use of concepts and on the interaction between textual and graphical elements. With respect to usability it is important that text, concepts and graphics are used with care and in a simple and unambiguous manner.

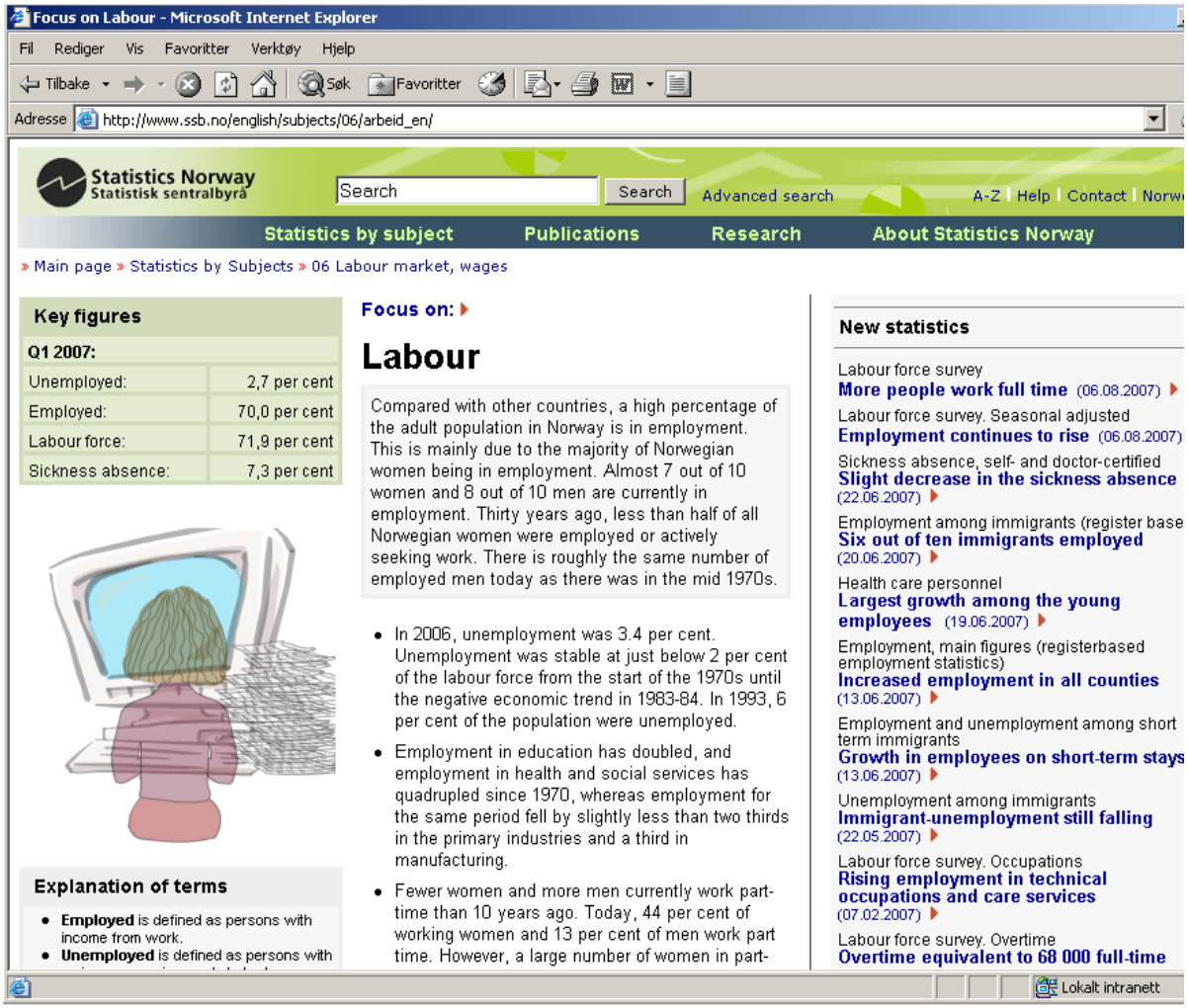
In the following we present a few examples that illustrate just how important it is to consider aspects like information flow, the use of text and concepts and the interplay between text and visual design in order to develop usable web pages.

Information flow

Example 1 shows one of the Statistics by Subject-pages on the ssb.no-site. Test results indicated that this kind of page can be confusing, especially to new users. The page conveys little information about where to start. It is difficult to tell what is most prominent and important on this page and where to look if you are searching for specific information. Most of our test persons found this type of page a little overwhelming; they perceived it as dense and rather messy. There is simply too much going on at the same time. We also found that the hyperlinks in the right-hand part of the screen seemed to take a lot of focus. The blue bold text here attracts the user's eye.

Some of the test persons felt that it would have been easier to cope with a more hierarchical structure with several layers and thus less information and text on each page. More experienced users, on the other hand, seem to have grown accustomed to the design; they are familiar with what type of information is placed where and tend to like the non-hierarchical, comprehensive design.

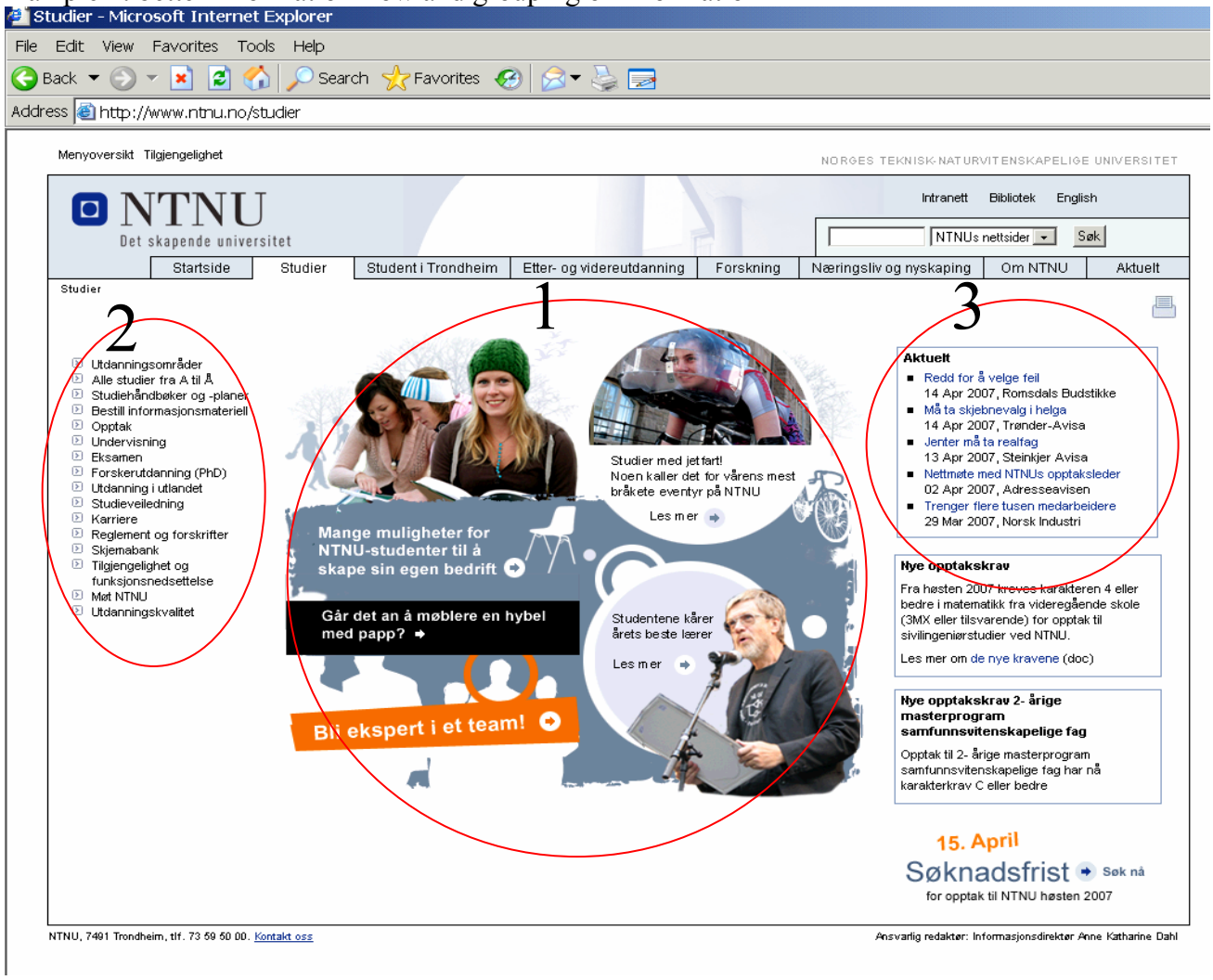
Example 1: poor information flow



Example 2 shows a university web page which tells about what kind of studies the university offers. We have not tested this page. Still, we think that this page illustrates how a page that covers a given theme or area could be constructed. Compared to the ssb.no-page in example 1 the visual design here is simpler and clearer and thus quite easy to follow, even for new users. The amount of text is kept to a minimum. Instead there is a menu to the left which helps the user navigate through the different related pages. Like in the previous page, links to related articles are placed in a separate field to the right, but on this page they are less pronounced and don't steel as much of the user's attention.

The middle part of the page (marked with "1" in the screenshot below) catches the eye as one enters the page and serves as an appetiser. It gives the page a young and fresh look and depicts the dynamism and energy associated with a university environment. The next thing one sees is typically the menu to the left (marked with "2" in the screenshot below). This is consistent with how one (in our part of the world, at least) normally reads; i.e. from top to bottom and from left to right. This is where the key information related to studies is found.

Example 2: better information flow and grouping of information



The links placed in the right-hand part of the screen (marked with "3" in the screenshot) lead to more peripheral, but related information connected to studies, typically news articles featuring themes relevant to students.

Interaction between text and visual design

We find that Dillman's principles for visual design are universal and as crucial in connection with ordinary web pages as they are in connection with questionnaires. Dillman considers visual effects a language of its own. He looks upon visual design as a non-verbal language that can be divided into three subcategories: a graphic language, a symbolic language and a numeric language (Dillman, D. 2000). The graphic language consists of graphical effects like size, distance and contrast. The symbolic language comes into play as soon as one makes use of culturally interdependent symbols like arrows, check-boxes and response fields; whereas the numeric language consists of numbers. The graphic language is superior to the two other languages, since both symbols and numbers are graphical. This is also the most powerful language of the three; research shows that especially size and distance have a great impact.

As we did the assessment of ssb.no, we registered a few things that differed from Dillman's guidelines. For instance, we found that the combination of text and visual design in the "Statistics A-Z"-field in the top left margin could be improved. As can be seen in example 3, the "Statistics A-Z.." text is separated from the letters (the hyperlinks) themselves, both by space and a line. This violates what is called Law of Proximity within gestalt and perception psychology. This law says that elements that are placed close to one another belong together, i.e. they are perceived as a unit, and vice versa.

Example 3: text and visual design that doesn't play together well (old version of ssb.no)

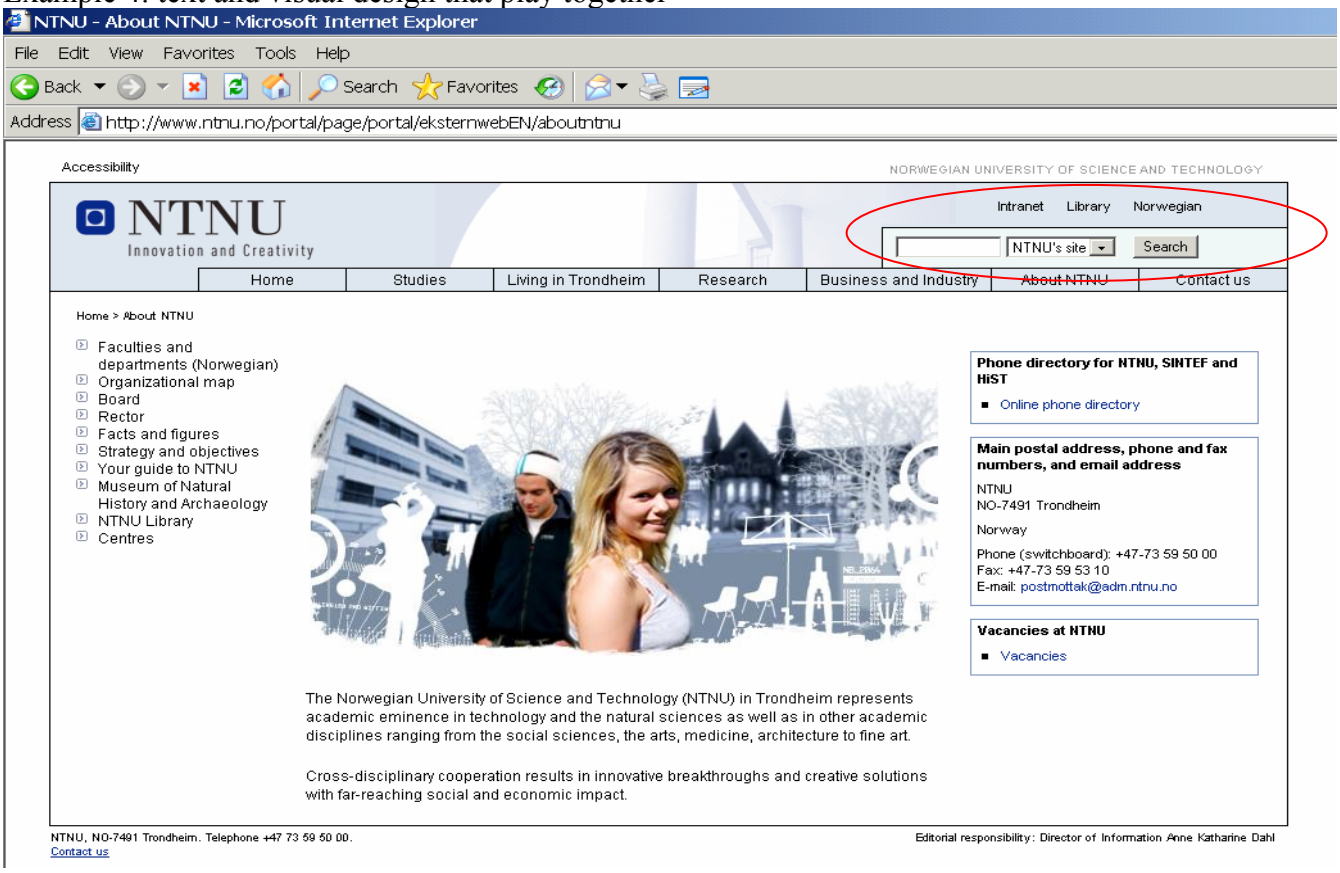
In this case, the distance between the “Statistics A-Z”-text and the letters is as big as the distance between the letters and the language links below. To make it visually clearer what goes together with what, the letters should be placed a little higher up and the topmost line removed.

The opposite is the case when it comes to the search field and the menu below it in the top right corner of the page. Since the search field and the menu elements below it are placed so close together, they may be perceived as a single unit. The menu does have a slightly different background colour than the search field does, but the difference is too subtle. One of our test persons actually thought that one could use the menu alternatives to limit the search, i.e. search only through the A-Z pages, the Help pages, etc.

The news articles in the middle of the page typically have a short ingress in small fonts above the heading, which is written in bigger, bold letters. The fact that the titles are emphasized like this makes them stand out and, consequently, they mark the point where people start reading if they read this part of the screen. This is probably the designer’s intention and all good and well, but what then, is the point in placing the short ingress before the title? Testing indicated that many users simply don’t see this text.

The university site constitutes a better example of how a search field can be designed and of how different elements that belong together can be grouped together visually. As example 4 shows, there are a few links placed rather close to the search field here too, but the use of colour, space and the frame around the search elements prevent confusion and clearly show what’s what. In fact, both the search function on ssb.no and on these pages enables the user to limit her search, but only the university page actually indicates *visually* that this is possible.

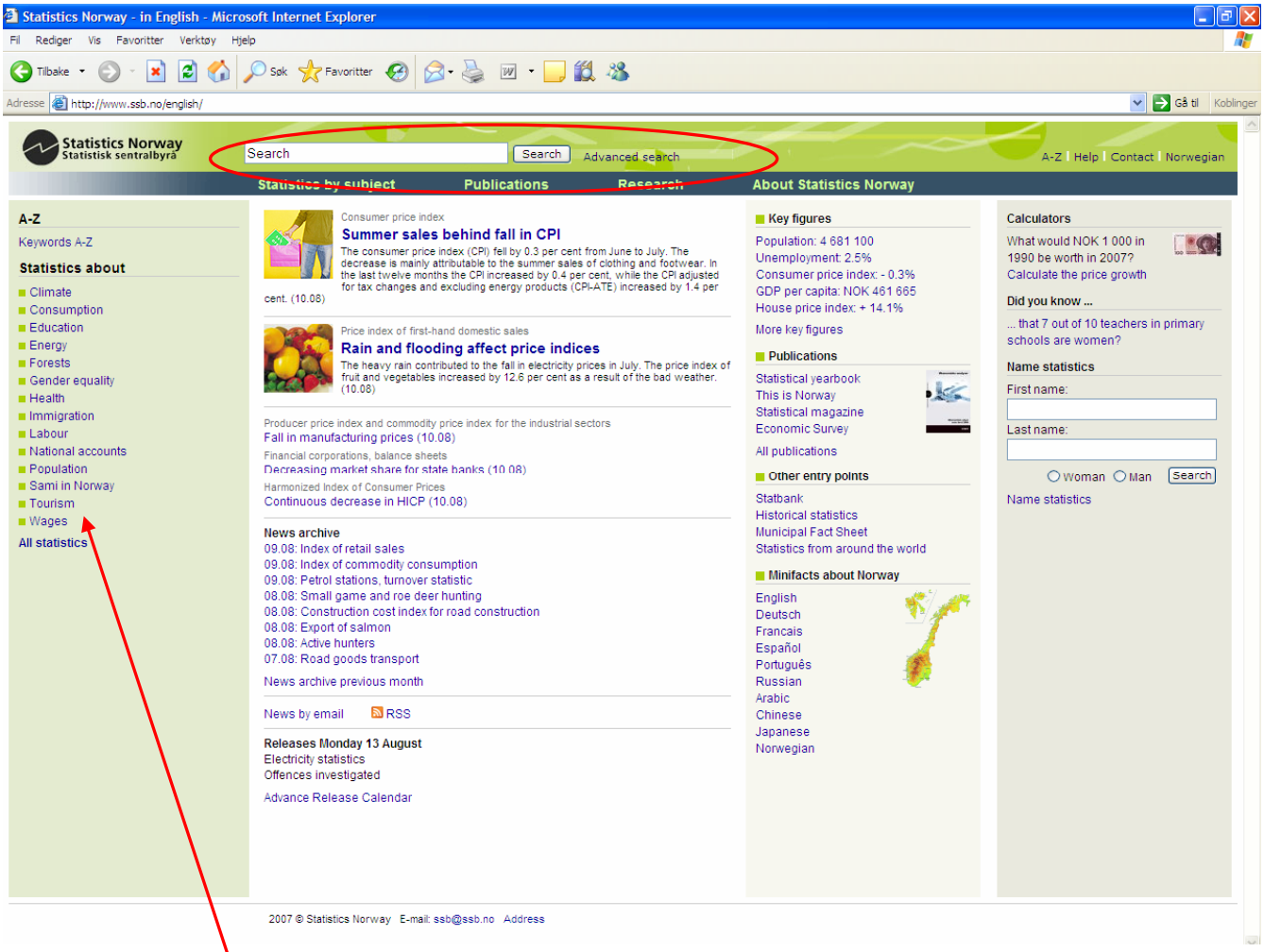
Example 4: text and visual design that play together



The new version of the ssb.no home page has also been tested and has proved to be more user friendly than the old version. The new page is, as one can see in example 5, divided into several columns, each with a different background colour, in order to group and separate the many different elements better. We feared that four columns might be too much, and still think that there is too much information crammed into one page, but testing indicated that users are able to cope. Most test persons answered that they preferred the new version to the old one.

The search field has been moved and has thus been separated from the small menu to the right. Also, the link “Advanced search” has been added just next to the field so that the user is made aware that there are more sophisticated search options available.

Example 5: Text and visual design that play together (new version of the ssb.no home page)



Use of text

In the left margin on the new version, the so-called theme pages are now listed under the heading "Statistics about". On the old page one would have to click on a link in the middle part of the page called "Focus on…" in order to get to the same overview of theme pages.

"Focus on…" is obviously not the most intuitive text; it doesn't tell the user much about what it leads to. During testing we found that users who knew about the "Focus on…"-link and had used the theme pages before, tended to use these pages a lot in their search for information on different subjects, whereas new users and users who weren't aware of the theme pages, didn't necessarily try out the link/find these pages at all. In fact, the whole collection of links which "Focus on…"

is a part of is a bit of an odd mix of different types of links; there is no clear relation or thematic similarity between them and thus no reason why they should be grouped together like this.

When it comes to unclear text, there are also a few examples to be found on the old version of one of the pages subordinate to the main ssb.no-page, the *Subjects*-page. In addition to the poor use of text, the visual design on this page is simply bad. As example 6 shows, the combination of short, often none-informative link titles and a messy layout contributes to create a rather overwhelming and chaotic impression. The use of arrows and the colour combination with blue and red doesn't help either. During testing, we found that quite a few test persons gave up and hit the back-button immediately as they got to this page, without even trying to make sense of it or have a look around to see if what they were after could indeed be found here; they simply assumed there would have to be an easier way to find the wanted information.

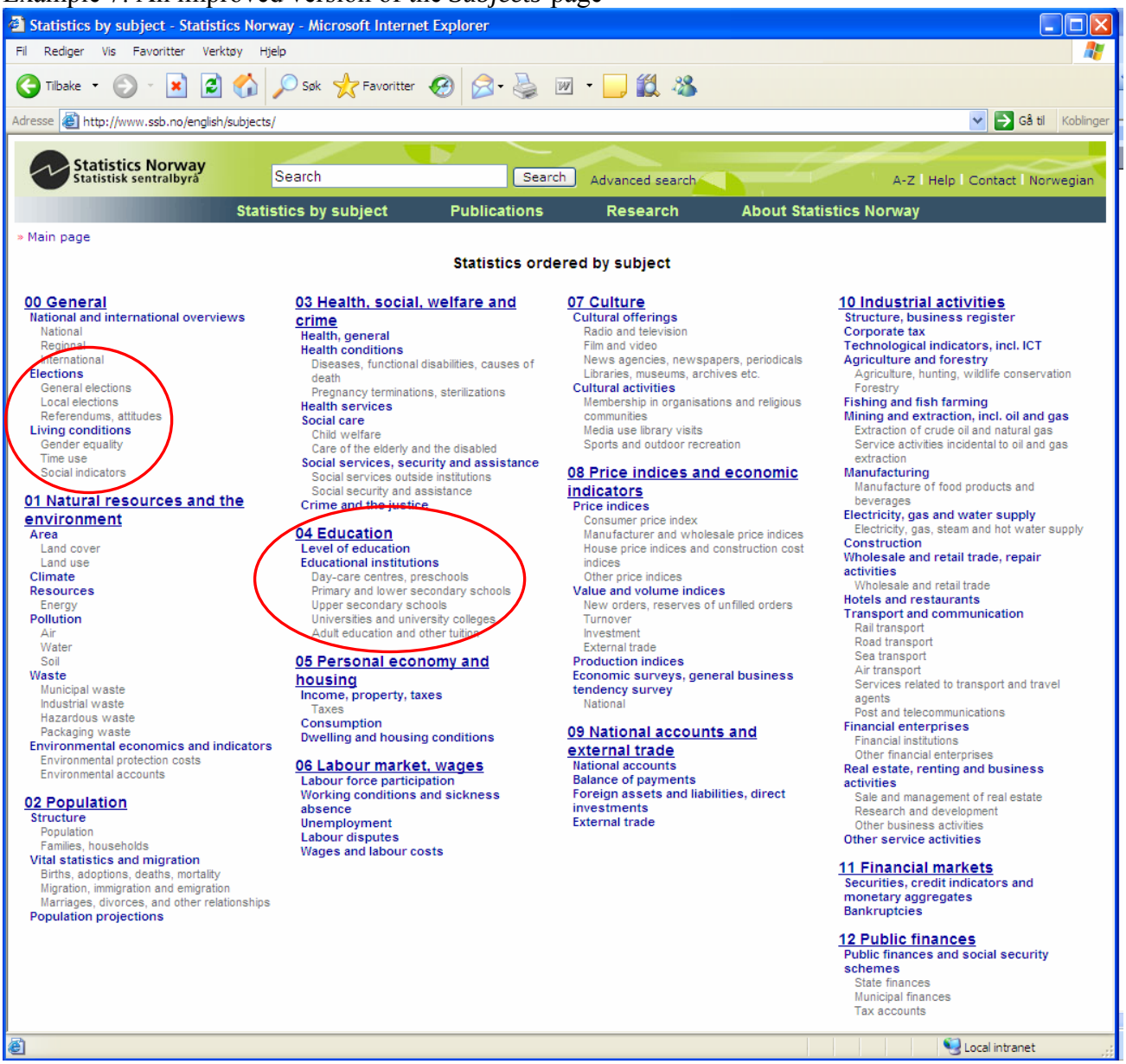
Example 6: Old version of *Subjects*-page. Unclear text, poor visual design and an discouraging overall impression

Words and terms like *General*, *Local*, *Gender*, *Time* and *Social* in the upper part of the page, and *Primary*, *Upper* and *Adult* a bit further down aren't terribly informative on their own. Even seen in connection with their respective heading, i.e. *Elections*, *Living conditions* and *Educational*

institutions, they don't really convey a whole lot about what the user can expect to find if she chooses to click on them.

Example 7 shows the new version of the *Subjects*-page. Just the new design in itself – the way the different subjects have been grouped together and structured – makes the page a whole lot better and less discouraging. The use of different font sizes, underlining and indentation conveys a clearer and tidier overview of the extensive collection of statistical categories.

Example 7: An improved version of the *Subjects*-page



When it comes to the use of text and concepts, the new page differs from the old one in that the text elements, the hyperlinks, are more specific and supplementary. Together with the stronger

hierarchical visualisation, the longer and thus more comprehensive links make it easier for the user to cope.

Wrapping up: what can survey methodologists learn from traditional web-design?
Of course there are evaluation criteria that are unique to web applications and that must be considered in addition to those that we as survey methodologists are trained to focus on. Technical preconditions, confidentiality and data security measures, efficiency, the capacity of server and PC, screen size, audio and video effects are just a few.

One of the problems we struggle with at Statistics Norway is that we are not good enough at taking advantage of the possibilities the web offers. We should look at web design and learn how to make better use of the web as a medium, including the use of navigational tools, user interaction functionality, the use of menus, buttons, links, audio and video effects, etc. By combining this kind of web design knowledge and the knowledge that we have about perception and cognitive processes, we should be well equipped for creating both better electronic questionnaires and better web pages.

Business Surveys – Testing Strategies in German Official Statistics

Karen Blanke
Federal Statistical Office (FSO), Germany

1. Background

German official statistics are collected from more than 170 surveys based on questionnaires. The majority of surveys collect information on businesses, farms, institutions, or public administrations. Only a minority is related to social statistics. Until now there are no systematic, standardised procedures for testing questionnaires. This is changing due to recent strategic decisions in the German statistical system and requirements of the European Code of Practice (CoP)[22]. Both strongly express the need for regular and systematic questionnaire testing. As a consequence, the FSO Germany has started to implement the recommendations formulated in the "Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System", published by Eurostat[23].

During the last months, a pretest on trade statistics[24] has been carried out. The survey in retail and wholesale trade is conducted annually. There are about 55000 retail and wholesale outlets participating in the compulsory survey (see chart 1). The businesses are selected randomly. Generally companies are in the sample for 10 years.

The survey aims to produce information about the structure of businesses, especially concerning efficiency and productivity. The data collection mode is either self-completion by paper and pencil or by an electronic form. Currently, more and more businesses deliver their data online, although the majority still use the paper and pencil version. The instrument covers the following topics (chart 2):

Chart 1: Survey design

- Compulsory, annual survey
- Random sampling (two stage)
- Sample size: 55000 retail and wholesale outlets
- Data collection mode: self-administered
  (paper or pencil or electronic form)

[22] Eurostat Code of Practice
http://www.epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47141301/VERSIONE_TEDESCO_WEB.PDF

[23] Eurostat (2006): Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System
http://epp.eurostat.ec.europa.eu/portal/page?_pageid=2273,1,2273_47143267&_dad=portal&_schema=PORTAL#METH

[24] Trade statistics, based on EU regulation Nr. 58/97 20.12.1996 (EG, Euratom)

Chart 2: Topics of the trade statistics

- Number and structure of employees
- Stock-related issues
- Expenditure
- Investment
- Turnover
- Subsidies
- Trade via the internet

The variety of different topics is presented on 8 pages. The questionnaire is considered by the companies to be too long. However, not every business needs to fill in the whole questionnaire. At least medium-sized and large businesses usually need two different divisions to provide the information: staff management and accounting. Consequently, the survey is considered as very burdensome and, from the businesses' perspective, has no relevance for their own work. Moreover, companies complain that the survey neither reflects the reality of politics nor that of companies. However, the survey provides structural data on the economy in Germany and is used by a wide range of users: government and ministries, trade unions and associations of enterprises, economic institutes and universities.
But also the users are not fully satisfied with the data presented. Two main aspects are criticised: late publication of the results (1½ years after the reporting year) and the inconsistency with other surveys covering comparable data. Therefore, it is necessary to test this instrument.

2. Concept of testing

Since the FSO is rather inexperienced in the field of testing business surveys, we basically used three sources to develop a testing concept:

1. The Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System,
2. the cognitive model on the answer process for businesses and[25]
3. the testing procedures employed by CBS Netherlands when testing the same instrument.

According to the Handbook it is recommended to implement several methods for pretesting, as each method has its own quality. With regard to the cognitive model on the answer process in businesses, sensitive aspects have to be checked such as record formation, the selection of the "right" respondent, the priorities of companies to fill in forms and to fit the information into the right category, and the authority from the management level to publish the data. Furthermore, we

[25] Willimack, D.K. and Nichols, E. (2001). Building an Alternative Response Process Model for Business Surveys. Proceedings of the Annual Meeting of the American Statistical Association.

found that CBS Netherlands has already conducted a pretest on business statistics based on the same EU regulation. The description of their proceedings helped us to develop our concept. However, as the pretesting team at the FSO is very small, a realistic scheme based on limited resources had to be launched. Consequently, our concept was based on three principles: (a) to implement simple but different methods, (b) to use what is available (e.g. data sets of previous surveys), (c) to communicate with all users of the instrument (respondents, experts, data entry staff and customers).

Following these considerations, a comprehensive, five-step concept was implemented (see chart 3):

Chart 3: Pretesting steps

Step 1:	Analysis of data (post evaluation)
Step 2:	Examination of terms and definitions (desk evaluation)
Step 3:	Hotline experiences
Step 4:	Expert groups
Step 5:	Cognitive interviews (company-site visits)

Analysis of the data sets
In a first step, we compared raw data from an earlier survey with the final data sets. The aim of the analysis was the evaluation of corrections and inconsistencies in the data sets, which might be related to shortcomings of the instrument. This step of evaluation was found to be more of a quantitative nature, as the frequency, e.g. of corrections, was one issue.

Examination of terms and definitions (desk evaluation)
This analysis was aimed at comparing current terms and definitions by examining the concepts of official statistics and those applied in the accounting systems of businesses. The examination shall help to uncover possible inconsistencies with other data sources, which result in (apparently) different figures on the same topic.

Hotline experiences
The collaborators responsible for trade statistics offer the businesses the opportunity to contact the FSO or the Länder offices, if they have problems to fill in the forms. This hotline is used rather often (approx. 5000 calls). Conversely, the companies are contacted if the data entry shows problems resulting from plausibility checks. Even though some items and problems are well known, the calls were systematically listed for the first time and gave an opportunity to identify possible systematic errors.

Expert groups

Subject matter experts on trade statistics - who are responsible for the organisation of data collection and partly of data processing - were consulted to describe their experience with the questionnaire. The discussion was led by a member of the pretesting team and was based on a compendium for the discussion. A second expert group discussion was conducted with members of the data entry team. Both groups were considered as users of the forms and to be involved in the data processing at different stages: either through organisational issues and data collection or by data entry.

Cognitive interviews
The most important method of this pretest was to carry out company-site visits in order to identify problems that the businesses encountered with the instrument when filling in the forms. For the first time, the producers (of the questionnaire) and the respondents of trade statistics had the opportunity to discuss drawbacks of the survey. Previously, either the hotline or remarks on the forms had been the only possible feedback companies could give. By visiting the companies, we wanted to find the appropriate person in the business to answer our questionnaire, and to know whether the requested data were available, whether the instructions and definitions were user-friendly, and whether terms like e.g. E-commerce and Leasing are understood in the way wanted.. Fourteen businesses participated (out of a planned 15), which was an adequate number, although retail trade companies were underrepresented. The interview was partly standardised, conducted by a member of the pretesting team, lasted 30-45 min and was tape-recorded. There was no special incentive involved. Once a company agreed to participate, they were usually interested, open-minded and had useful comments. Some used the opportunity to complain about the increasing load of surveys (market research and official statistics), which is in fact a problem in Germany.

3. Findings

The aim of the following presentation is to point out the methods and their usefulness for specific purposes, not to show and discuss the content and direct interpretation of terms, wording etc.

Analysis of data (post evaluation)
The comparison of raw data with final data sets was carried out by the subject matter experts of trade statistics. The evaluation revealed several problematic categories and some incorrect data delivered by the businesses. Relying on the businesses' intention to provide valid data, inconsistencies should either be related to unintelligible definitions and confusing terms or to the non-availability of data. Most corrections were due to interrelated values which did not fit. About five to eight items turned out to be problematic because of the high frequency of corrections; other errors were diverse and not systematic. Within the context of the corrections, it was possible to show where overestimation and underestimation posed a problem and, consequently, to see where respondents used a wider or more limited definition of what should have been reported. But the drawback of this method was that it did not reveal the reasons for errors and thus gave no indication of how to improve the questionnaire. As a consequence, items with high rates of corrections were listed for discussion during company-site visits. In addition, high rates of corrections did not always correspond with the problems discussed via the hotline. Obvious problems revealed by the evaluation of data sets are different from those, of which the respondents are aware and which they need to clarify when calling the hotline.

Examination of terms and definitions (desk evaluation)

The examination of the terms and definitions used by official statistics in comparison with those applied in the accounting systems was not very extensive, due to limited manpower. Apart from that, the subject matter experts were often convinced that the terms in use had been checked in detail while implementing the regulation and should be comparable. Basically, the terms in use have been predetermined by international agreements on the definitions to be used in accounting systems. However, some representatives of trade associations doubted that the respondents in businesses knew the definitions and applied them correctly (see also company-site visits). Besides, other terms not originally defined for the accounting systems were detected and the definitions scrutinized. Finally, a list of terms which might be problematic was compiled and served as the basis for our company-site visits.

Hotline experiences

The systematic documentation of calls and requests was very useful, as it produced a list of problematic items. Some items have been well known among the colleagues directly involved, but have not been systematically transferred to the group of experts who are responsible for the survey. In addition we found that a separate list of the problems discussed following initial requests from businesses or initial call-backs by experts was helpful. The background of calls initially made by the companies has been different from that of requests made by the office: Requests by the companies deal with problems, of which they are aware, but which are seldom associated with the terms in use and their definitions, because people are not as sensitive to problems concerning well-applied definitions as statisticians would want them to be (see company-site visits). Their requests were mainly linked to the basic motivation to fill in the form or contained specific questions connected with the special conditions of the company. Initial call-backs from experts in the office were motivated by problems that arose when entering data and when plausibility checks showed inconsistencies. Thus, more interrelated errors could be detected and solved case by case. Consequently, requests by the companies are linked to their awareness of problems, whereas hotline call-backs are more associated with errors that (maybe) are not encountered by businesses.

Expert groups

Two kinds of expert group meetings were held: one with the subject matter experts of the Länder who are responsible for the data collection, and one with the data entry staff. Previously, basic debriefings had been carried out with the interviewers after the data collection in order to get a feedback on the instrument. The problems that were expressed by the staff responsible for the fieldwork showed a more administratively oriented perspective. This helped to understand organisational problems, which are always neglected. The main discussion was focused on allocation time and constraints, applied terms and definitions. Sometimes the experts had a good feel for problematic terms; they were often astonished that these issues could be difficult for the respondents. Hence, some common terms of trade statistics, like retail trade, need an exact definition, even though one may expect people to know the subject.

Additionally, the data entry staff were invited to talk about their experience. The discussion was very lively, frank and open. Most remarks were linked to layout effects and the order of questions. For example, the position of item x was bad in comparison with previous forms, or

bold letters were to heavy, etc. The data entry staff have also been responsible for call-backs, and they found out that several pages were overloaded. Most call-backs were linked to one basic question of the survey, which they thought was placed too close to the end, where people get tired completing the questionnaire. Thus, very useful comments concerning the survey form can be expected when conducting discussions with these colleagues.

Company-site visits

The recruitment of businesses to participate in the pretest was very burdensome and time-consuming, even though the chamber of industry and trade served as a reference to promote the participation. An invitation letter informed the businesses on the planned pretest. Some days after sending the letter, the businesses were contacted by phone. Sometimes more than five calls were necessary, only to get in contact with the person in charge to decide if the business would participate. Of those who agreed to participate, only half filled in the form before the interview, others just scanned it before the interview. The 14 participating companies were very diverse. The contact persons were either managers or book-keepers. It is not possible to determine a "perfect" person for answering the questionnaire, as this depends totally on the size and structure of a company. Where businesses employed tax consultants to do their tax declarations it would have been best to interview them. However, in Germany it is quite unrealistic to interview a tax consultant due to the fact that they get paid by the companies, but not for pretesting. The most helpful respondents were the book-keepers of the companies: They know their subject and argue about the terms, reference periods etc.

Another interesting aspect when conducting company-site visits has been the treatment of instructions and definitions by the respondents. Most people use their own definitions. This especially applies to terms that are in daily use, in newspapers and on television, and so to say are "common" vocabulary. Good examples are terms such as part-time work, turnover and social contributions. The respondents use their own understanding of the terms and initially do not examine whether their definition is the same as in the questionnaire. Consequently, currently used terms of daily life are seldom checked against the instructions. This even applies to technical terms like E-commerce and Leasing. The only way to get some people to read definitions is to place them directly into/below the questions. On the other hand, some respondents do expect instructions and become irritated if they cannot find them. In these cases, it is advisable to provide instructions so that definitions and instructions can be be checked by demand. These results could only be achieved by conducting company-site visits.

4. Conclusion

This contribution outlines the conceptual frame for testing trade statistics at the FSO in Germany. The FSO implemented several methods: post evaluation of data sets, expert group discussions, checks on terms and definitions, discussion with data entry staff and company-site visits. The purpose of this short paper is to give a feedback on the different methods implemented and their potential use. In general, the implementation of several methods has proven to be useful, as the different methods showed different potentials to uncover shortcomings of the instrument: Weaknesses in the layout were especially well observed by data entry staff and the respondents. Problems concerning data sources and definitions were best identified by conducting company-site visits, whereas fieldwork officers provided better advice on the organisation of the survey,

and subject matter experts on either organisational issues or the conceptual frame. Interrelated problems due to overlapping categories were detected by checking raw data and final data as well as via the hotline. As regards the further planning of pretests, three aspects have proven to be worthwhile for further investigation: First, it is necessary to improve the strategy of recruiting businesses. Secondly, a profound and systematic evaluation of cognitive interviews (either of companies or private respondents) needs to be developed, and thirdly, a more extensive post evaluation of the data is required.

Standards for Questionnaire Design and Layout of Business Surveys

Birgit Henningsson

1 Background

Statistics Sweden has just started a comprehensive work to standardize our surveys as much as possible, particularly business surveys. We are annually sending out about one hundred different surveys to businesses. They all vary substantially concerning, for example, variables and layout. In addition the Swedish government has commissioned us to reduce the response burden for businesses with 25 % within 4 years. At the same time we are supposed to increase the production of statistics in the economic field. Moreover we need to reduce the amount of editing. All together, this means that improved questionnaires are necessary. We believe that standardization of the surveys will increase the effectiveness and also reduce the response burden.

2 Our project

Our project "Standards for Questionnaire Design and Layout in Business Surveys" went on for three months. We scrutinized nine of our surveys which need a lot of manual editing before data entry. We had four debriefings with the working groups doing the data editing. The problems in the questionnaire are often very complex. Therefore the staff in these groups needs to have many contacts with the respondents.

We have guidelines for our web surveys, as we discussed at an earlier QUEST meeting. Today we also have special software for web design; however, paper is still the dominant data collection mode. Therefore, one of the goals of the project was to create better guidelines for the paper versions. However, since the project only lasted for 3 months, the output represents merely a start.

The goals for the project were

Standardized questionnaire layout

- Makes the task as easy as possible for respondents
- Designers should concentrate on contents
- Higher efficiency

Questionnaire development

- Evaluate that questionnaires are adapted to data provision capabilities
- Systematic review and improvement
- Additional tools for evaluation

Most business surveys are surveys that are repeated monthly, quarterly or yearly. Therefore we want a cyclic model in which the questionnaire can be improved continuously in multiple rounds. That means:

Cyclic Model

- Define the Survey
- Questionnaire Design
- Testing
- Data Collection
- Evaluation
- A new round

3 Enterprise characteristics

There are several differences between household and enterprise surveys.

First, the response process is quite different. One needs to find the right person to target and, very often, it is not only one but several respondents involved.

Second, the statistical units differ – they are not homogeneous – some are big whereas others are small. The population change all the time.

Third, regulations from the European Union and measurement problems due to globalization complicate the situation further.

Fourth, you also need to coordinate between different surveys. As such, terms and definitions are very important and not always easy to decide and apply.

4 Questionnaire Design

4.1 Guidelines

Statistics Sweden has guidelines for household surveys but since enterprises are different compared to households, the guidelines cannot be generalized to business surveys. So, we did a literature review and made guidelines with the following headlines:

- Types of questions
- Response alternatives
- Language
- Definitions and instructions
- Order of questions
- General advice and guidelines for writing questions

Our ambition is that the guidelines should be compulsory.

4.2 Checklist for expert review

Today the Measurement Lab consists of about ten people. With that many people involved, we need some policies or checklists to follow. We used the guidelines mentioned above to develop such a checklist for expert reviews of business surveys. The headlines for the checklist were:

- General things
- Covering letter and instructions
- Questionnaire

- Structure of the questionnaire
- Formal issues
- Visual layout
- Tables
- Questions
- Response categories

5 Visual layout

We have guidelines for web surveys, but we do not have similar guidelines for the paper versions. Such guidelines are necessary since paper is still dominant and will remain important in the foreseeable future. Again, we reviewed the literature, focusing on how existing principles are applied in practice.

It is well established that the layout has a large effect on the data process. However, there are many things to consider when designing a questionnaire. The questionnaire should, of course, be compatible with the scanning procedure. It is also important that there is a clear flow through the questionnaire. We also discussed and decided on where to place information concerning confidentiality, the name of the survey and the last day for returning the questionnaire. Furthermore, the boxes for "username" and "log in" should look the same and put in the same place in different cover letters and questionnaires.

Questionnaires for business surveys vary considerably. It is not possible to make *one* standard and apply it in every case. Still, we created some visual examples of standard questions in the nine surveys. I will give you one example from "Rental charge for new apartments".

On the next pages you can see the old version and the new one. First we changed the first table into two tables. "Square meter" was put in one table together with "how many apartments". The cost in SEK came in the next table. This makes it easier for the respondent to recognize what we want them to enter. The navigation is easier. The spaces for the numbers also help +in showing whether we want figures about one apartment or the total amount. We also made some other improvements such as:
- Comments in the last question
- Special apartments got clear instructions
- Definitions in the table
- Ready for using a scanner

Old version:

Lägenhetstyp *OBS! Ändra bara på lägenheterna ovan. Är rum räknas som ett rum*	a) Hur många lägenheter av varje typ finns i hushållet?	b) Hur fördelas enbart kvadratmeter på respektive lägenhetstyp?	c) Hur stor är summan av alla årshyror/avgifter för respektive lägenhetstyp? **	d) Hur stor är summan av insatser för bostadsrätter/ kooperativa hyresrätter?	e) Hur många lägenheter var outhyrda/ osålda den 31 december 2005?
	Ange antal	*Ange hela kvadratmeter*	*Svara i kronor*	*Svara i kronor*	*Ange antal*
1 rum och kokvrå.	011	012	013	014	015
1 rum och kök	021	022	023	024	025
2 rum och kök	031	032	033	034	035
3 rum och kök	041	042	043	044	045
4 rum och kök	051	052	053	054	055
5 rum och kök.	061	062	063	064	065
6 eller fler rum och kök	071	072	073	074	075
Övriga lägenhetstyper (ex. kokvrå med fler än 1 rum)	081	082	083	084	085
Summa samtliga lägenheter	091	092	093	094	095

New version:

1. Fyll i följande uppgifter för bostadslägenheter färdigställda 2006 som är tillgängliga för alla.
Lägenheter inom samma lägenhetstyp redovisas i en grupp oavsett eventuella skillnader i ytor eller hyror.
Räkna även med outhyrda lägenheter.

Lägenhetstyp	**a) Hur många lägenheter av varje typ finns i huset/husen?** *Ange antal*	**b) Vilken är den <u>sammanlagda</u> bostadsarean för varje lägenhetstyp?** *Ange hela kvadratmeter*	**c) Hur många av lägenheterna var outhyrda/osålda den <u>31 december 2006?</u>** *Ange antal*
1 rum och kokvrå	011	012 **m²**	015
1 rum och kök	021	022 **m²**	025
2 rum och kök	031	032 **m²**	035
3 rum och kök	041	042 **m²**	045
4 rum och kök	051	052 **m²**	055
5 rum och kök	061	062 **m²**	065
6 eller fler rum och kök	071	072 **m²**	075
Övriga lägenhetstyper	081	082 **m²**	085
Summa för samtliga lägenhetstyper	091	092 **m²**	095

6 Editing Staff Debriefings

6.1 A new method

We tested a qualitative method to take advantage of the experiences from the production process. Our goal was to have specific information about the unique survey. We also wanted to see how a debriefing with the editing staff would work. The editing staff is doing a lot of data checking. During that job they have many contacts with the respondents since the problems are often complicated and need to be discussed. Maybe they could help us with more information? We made four debriefings within the project. Our four editing staff groups had long experience from the survey. We held a meeting with the person responsible for the survey. Then, we went on with the editing staff debriefing. To make sure we were consistent, we developed and followed a topic guide for the debriefings.

6.2 Topic Guide

Here are the basic contents of the guide:
We started with an overview of the survey. Then, we go on to the details, which mean the variables and the instructions.

- Start with the aims of the debriefing
- Ask the responsible person to describe the editing process from the point where they get the questionnaire back from the respondent to when the editing is finished
    - Ask the rest of the group to fill in
    - How is your documentation?
    - How is the documentation used and how often?

- In what ways may the respondent answer the questions (web, paper or a file)?

- What reactions do you have from the respondents?
  - Positive?
  - Negative?

- Are there any problems with the variables in the survey?
  Time for more specific question in the survey
  Questions from the survey manager
  Changes this year?
  What effects?

- What are the reasons for the problems?
  Technical ones?
  Other: for example definitions, regulations etc.

- Do you think the respondent use the instructions?
  If Yes: Still problems?
  If No: Why not?

- What kind of businesses has problems?

And why? Bad quality of the register, different …, different attitudes from the respondent.

- Do you need further improvements in your system?

6.3 Results and recommendations

This was another tool for us to use when we want to evaluate a questionnaire. The debriefings gave us a lot of good information about the error sources in the survey. The debriefing can pinpoint where the problems are. Very often we also got information about the reason why there was a problem. The experiences and opinions from the debriefings represent many different respondents. A debriefing with the editing staff gave us the opportunity to separate the hard errors which happen very often and other errors that occur more seldom. As such, we know what to focus on.

Cognitive testing with "real" respondents is of course the best way to evaluate the measurement instrument. Still, our experience is that you can use debriefings as substitute or "proxy" when you don't have time or money to visit the firms and test the questionnaire cognitively. It can also be used as a complement to cognitive testing or quantitative methods such as process data.

Our suggestion is to combine a debriefing with a cognitive test. The debriefing can give you input how to formulate probes in your topic guide.

7 What's left to do after the project?

The group who ordered the project was very satisfied with the very practical results. Now, we need to go through all our questionnaires for business surveys and apply our new guidelines. We also need to:

- Make cognitive tests with "real" respondents
- Finish the job with guidelines and templates. This is not easy. The variables are so different and we have a lot of regulations from the European Union that we must follow.
- Use the results from this and other projects to finalize a complete cyclic model

Using Cognitive Interviews to Test Business Surveys

Marcel Levesque
Questionnaire Design Resource Centre, Statistics Canada

Background

The Unified Enterprise Survey (UES) began in 1997 with a pilot of a small number of industries and has continued to the present time where it now encompasses the majority of businesses surveys conducted by Statistics Canada.

The UES is basically a single questionnaire with a variant for each industry sector. There is a set of core questions asked of every industry and industry specific questions for each different sector.

Core questions cover such topics as the reference period (financial year of the establishment) covered by the survey, revenues and expenses, personnel, sales according to types of clients and their location and, if appropriate, questions on international transactions.

Industry specific questions refer to the main commercial activity of the establishment and industry characteristics, namely sales activity specific to the sector.

The UES attempts to cover all provinces and territories, all industries and all types of businesses, large or small, private or public, employer and non-employer and incorporated or unincorporated.

The purpose behind the UES is to develop a common approach with concepts harmonized across all industries to facilitate data collection and data capture, reduce and better manage response burden and increase response and data quality.

Determining priorities and establishing issues to be tested

Prior to proceeding with the testing of the UES questionnaires, the Questionnaire Design Resource Centre (QDRC), in consultation with the Enterprise Statistics Division at Statistics Canada responsible for the UES survey, determines the issues to be tested.

Issues usually explored with typical survey respondents include the following:

- The appropriateness of questions, concepts and terminology
- Respondents` willingness and ability to respond
- Respondents' understanding of the questions and what information to report
- Respondents' use of external sources of information such as financial or administrative records and their need to consult other individuals to provide the information requested
- The compatibility of questions and response categories with respondents` record keeping practices
- Difficulties respondents may encounter in retrieving information and completing the questionnaire

- Language translations (usually English to French)
- Respondents` suggestions on how to improve the questionnaire

Choosing and recruiting respondents for questionnaire testing

Respondents for cognitive interviews are usually recruited from Statistics Canada's Business Register or from another frame of establishments provided by the industry-specific program areas. They are identified based on the project team's specifications, designed to ensure a variety of business establishments for each industry sector to be tested.

Appointments are made with the person within the business establishment who usually completes questionnaires for Statistics Canada. The ideal respondent is the person in the business who is most knowledgeable about the data requested, who has access to this data and who also has the authority to release it.

Respondents are provided with an explanation of the purpose of the cognitive interview with emphasis on the fact that Statistics Canada is consulting with them to obtain their feedback and opinions. The importance of their participation is emphasized and they are given assurances of confidentiality.

Respondents are also told they will receive a copy of a Statistics Canada publication, usually the Canada Year Book, in appreciation of their participation in the testing of the questionnaire. This helps ensure cooperation and is a positive respondent relations gesture that brings an appropriate closure to the interview.

When testing business survey questionnaires, the QDRC usually recommends that approximately six to ten cognitive interviews be conducted in a minimum of two different Canadian cities. The more diverse the survey population, the higher the number of cognitive interviews that should be included in the questionnaire testing.

Whenever possible, an initial draft of the questionnaire is reviewed by the QDRC prior to proceeding with the testing. Then a revised version is tested with selected respondents. This approach has proven effective in maximizing the usefulness of the testing results and in providing an opportunity to make improvements to the questionnaire.

Using cognitive interviews as a testing method

According to Tourangeau (1984), a respondent, in answering survey questions, will walk through the following four steps of the response process: [26]

[26] Tourangeau, R. (1984), "Cognitive Sciences and Survey Methods" in T. Jabine, et al. Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, Washington, D.C.: NationalAcademy Press

- **Comprehension**: understanding the meaning of the question.
- **Retrieval**: gathering the requested information
- **Judgment**: assessing the relevance of the retrieved information to the data requested
- **Communication**: reporting the response to the question

Cognitive interviewing is used to identify difficulties or potential measurement error in each of these four stages of the response process and suggest possible modifications to reduce errors. It relies on an interviewer guide designed to explore respondents' understanding of questions, their strategies to retrieve the information requested by the survey, their judgment of the adequacy of the retrieved information and their ability to report the required data.

The following are examples of difficulties that can be encountered in each stage of the response process taken from a paper presented to the Federal Economic Statistics Advisory Council in December 2000. [27]

Comprehension

In understanding a question, a respondent may encounter difficulties with the following:

- Instruction content - inaccurate or conflicting instructions, complicated content, complex syntax, instructions separate from the question or provided too late, unclear examples, unclear layout.

- Question wording – critical definitions missing, ambiguous terms, multiple definitions for a single concept, industry specific language

- Question structure – implicit assumptions, several questions in one, Question /answer mismatch

- Question content – complex topic, topic carried over to another section, question too short and not sure of the meaning

- Reference period – may be undefined, period too long, abrupt change in reference period from one section to another

- Navigational instructions – inaccurate instructions (move to the wrong place), confusing flow, complex information

Retrieval

In gathering the requested information, a respondent may encounter difficulties with the following:

[27] *Using cognitive methods to improve questionnaires for establishment surveys, Cognitive applications for establishment surveys working group, presented to the Federal Economic Statistics Advisory Council Dec. 2000)*

- Multiple sources – information may be distributed in multiple sections /departments in the organization

- Record retrieval – records not available, records in multiple sources, record access issues, authority issues

- Memory retrieval – shortage of memory cues, reference period too long, recall problems and telescoping errors.

Judgment

In assessing the relevance of the retrieved information to the data requested, a respondent may encounter difficulties with the following:

- Data incompatible with existing records

- Coordination and collaboration with other respondents necessary

- Need to evaluate and synthesize multiple sources of information

- Burden may lead to guessing or estimation

- Potentially sensitive information, strategic or proprietary issues may prevent revealing of certain information.

Communication

In reporting the response to the question, a respondent may encounter difficulties with the following:

- Response terminology – critical definitions may be missing or vague terms may be used

- Mismatch with technical language

- Industry-specific terminology may be used

- Missing response categories.

Questionnaire testing for the UES uses cognitive interviewing to focus on this response process, with interviews carried out at the respondent's place of business. Because of the financial nature of the data collected in these questionnaires, emphasis is placed on the process respondents go through to retrieve the information and their need to access information through external sources such as financial and /or administrative records.

Wherever possible, the QDRC tests the whole questionnaire package including accompanying introductory letters and reporting guides. Testing the whole questionnaire package is important to see if respondent relations material such as the introductory letter affects respondents' willingness to complete the questionnaire.

Problems encountered in the UES 2005-2006 UES testing

The UES 2005 questionnaire testing focused on the service industry. Both core questions and industry specific questions were tested for questionnaires on automotive repairs, accounting and bookkeeping, periodical publishers, film and video distribution, film and video production, employment services, rental, leasing and property management and real estate brokers

The 2006 questionnaire testing also focused on the service industry but in this case on six different industry sectors, namely architecture, engineering, advertising, theatre arts, sports and variety shows, entertainment and leisure.

This section presents some of the problems encountered in each of the four phases of the response process and solutions recommended to reduce the potential for error.

Comprehension

Question structure

One of the difficulties encountered in the comprehension phase of the process was in relation to the structure of the questions, more specifically to a mismatch between the question and some of the answer categories.

For example, in the 2005 survey on automotive repairs, respondents were asked to record sales for different types of goods and services related to automotive repairs and maintenance and were provided with different answer categories to do so. Although the section referred to sales of goods and services, the first answer category provided related to labour costs for automotive repairs and maintenance. Consequently, respondents were confused because this answer category referred to expenses while the section was focusing on sales.

In discussing this issue with those who designed the questionnaire, it was found that the intention behind this question was to obtain the sales value of labour costs billed to the client but respondents failed to understand the question in this way. Although the difficulty here was one of question wording, it was also a mismatch between question and answer, which led to confusion.

Once respondents understood the intention behind this question, they recommended that the expression 'labour costs' be replaced by 'sales resulting from labour costs billed to clients'.

Question wording

Another difficulty encountered in the comprehension phase of the process was in relation to the question wording in the French version of the questionnaire, more specifically resulting from the translation from English to French.

One question in the 2005 questionnaire on automotive repairs asked respondents to report expenses for each of a series of items, one of which was costs for employment agencies. The French translation used was *'coûts des bureaux de placements et des services de location de personnel'*. The expression 'personnel agencies' translated into '*services de location de personnel'* did not make sense to the French respondents since the word '*location'* in French means rental or leasing. Since personnel is not rented or leased, respondents did not understand the meaning of the expression '*services de location de personnel'*. As a result, many respondents did not know what costs to report in this column.

Since '*bureaux de placement'* was a more common expression, it was suggested that only this category be included in the question, leaving out the reference to '*services de location de personnel'*.

Retrieval

Multiple sources and record retrieval

One of the difficulties encountered in the retrieval phase of the process was in relation to the lack of fit between the question structure and the record keeping practices of the respondent.

For example, some respondents in the 2006 survey on entertainment and leisure were involved in the health fitness industry and had difficulty reporting some of the expense items listed in the question on expenses. The respondents offered a variety of services such as a spa, massage, tennis courts, pool and exercise rooms and expenses were recorded in different accounts. In order for them to answer questions on expenses, they would need to consult the accounting books for each different service department. Furthermore, some of the expense items listed did not correspond to their accounting books. As a result, the respondents clearly indicated they would provide estimates for many of the items and would not know where to report some of the others.

Examples

- Item one in the questionnaire asked for salaries in Canadian dollars and item five asked for the dollar value attributed to contract workers. Some employees are hired as replacement staff and others for specific contracts. Although staff are hired on contract, the costs are listed under salaries in item one. In this case, the respondents would not have reported any dollar value in item five for contract workers and there was no way for them to indicate that these expenses were all under item 1 on salaries. Consequently the data quality for both items one and five could be compromised.

- Item fifteen in the questionnaire asked for telephone and telecommunications expenses. In some establishments, these were grouped with other expenses and recorded by department. Consequently, respondents would not have tried to uncover these expenses by department nor would they have tried to extract them from the expense groups in which they were recorded. As a result, respondents would have reported an estimate of these costs or would have left the question unanswered.

Since expenses are part of the core questions applicable to all industry sectors, it is quite difficult to have answer categories corresponding to all industry sectors and all establishments. It was, therefore, important to provide a means to the respondents that would allow them to explain why they were unable to. In order to increase the possibilities of getting a good estimate or getting some indication of how the respondent had answered these questions, it was recommended that a space for 'comments' be added under the section on expenses to allow respondents the possibility of explaining the items they are unable to provide as listed i.e. telephone included in other expenses by department). It was also suggested that an instruction be provided mentioning that, in such cases, best estimates were acceptable.

Judgment

Data incompatible with existing records and the need to evaluate and synthesize multiple sources of information as in the examples given above will often have an impact on the respondent's decision on what data to provide. If records are largely incompatible with the structure of the questionnaire, the resulting burden may lead to guessing or estimation or the respondent may simply choose to ignore the question and not provide any data. In a business environment where respondents' main preoccupation is on the success of their business, time is limited and few will consider spending time to research information if it is not readily available.

As mentioned above, in order to increase the possibility of getting some indication of why it is difficult for the respondent to answer these questions, it is recommended that a space for 'comments' be added in such cases to allow respondents to explain the items they are unable to provide as requested. Consequently, once the questionnaires are received and such missing data is identified along with accompanying explanations, it would be possible for data analysts to contact the respondents and ask them to provide corresponding figures from their ledgers. In such cases, analysts could search for the appropriate figures and record them in the questionnaire. Such a suggestion was given by actual respondents who participated in the testing of these questionnaires, be logistically difficult and expensive to do.

Potentially sensitive information, strategic or proprietary issues may also prevent respondents from revealing certain information. In the 2005 survey on service industries, some respondents involved in "real estate, leasing and property management" said they would not provide information on revenues because their company had a policy on confidentiality which would not allow the release of such information. They did, however, mention that they would be willing to provide estimates.

In such cases, providing revenue categories may encourage respondents to provide information on revenues they would not report if exact figures were requested.

Communication

One of the difficulties encountered in the communication phase of the process for the 2006 service industry surveys was related to response categories.

One section asked for a breakdown of sales for different types of products related to architecture. Some respondents who designed airports could not find any category related to sales for projects in airports. Although there was one last category for 'other sales' in the questionnaire, they would not have reported this information because they could not find any specific relevant category applicable to their sales.

Other respondents involved in architectural landscaping asked where they would include sales related to such landscaping for condominiums. The answer categories provided referred to architectural landscape products for residential and non-residential projects, projects for open spaces and consultation. Did residential projects include condominiums?

In another case still related to architectural landscaping products, there was an answer category for urban planning projects (*projets d'urbanisme*), a term which seemed industry specific but was not familiar to the architects interviewed. One asked if this referred to projects with cities and universities.

In all these cases, respondents were left not knowing where to include sales which could have been significant. They would not have reported their figures or would have included them in inappropriate categories resulting in poor data quality.

In all these cases it was recommended to add significant answer categories and define those which were not clear (i.e. *projets d'urbanisme*.)

Observer notes

Cognitive interviews are conducted by a QDRC consultant, based on an interview guide designed with the client, the client usually being those responsible for conducting the survey and designing the questionnaire.

Since interviews are carried out at the respondent's place of business and considering the level of detail of the questionnaires, an observer from the client division is usually asked to participate. The observer's role, in this case, is to take notes during the interview and to report them to the QDRC interviewer once the interviews are completed. Although the interviewer will also take notes during the interview, it is difficult for him to capture all relevant information concerning difficulties encountered with the questionnaire. Also, the observer is usually someone knowledgeable about the specific industry being surveyed and can therefore respond to comments or questions the test respondent may have, if appropriate.

Although recording of the interviews is sometimes done, with the respondent's authorization and with the assurance that information provided is confidential, observer notes are always important since they provide another perspective on the issues observed during the interview. Since

observers are generally involved in the design of the survey questionnaire, their presence at the interview will give them a first hand view of the respondent's reactions to the questionnaire and provide a greater understanding of the difficulties encountered in answering the questions and the response burden that may result from the questions.

Since response burden is a critical issue for the UES and may have a significant impact on the results of any survey, it is important that observers be aware of the level and the nature of this burden.

Report writing and debriefing

A comprehensive report is usually prepared and presented to the client with findings from the interviews and recommendations for improving the questionnaire.

Such a report presents the findings for each section of the questionnaire tested and is based on the issues identified at the onset, as included in the interview guide.

A description of the methodology used is also included along with a profile of the participants.

A short introductory text informs the client that there are limitations to any study using qualitative research methods such as cognitive interviews. Since the research is qualitative, findings and conclusions are not necessarily representative of the whole population. The client is therefore advised not to view the results as statistical findings but rather as a method to provide important insights into the participants' opinions on the issues discussed and on their response process.

Conclusion

In a business environment where a respondent's main preoccupation is the success of his business, time is limited. Few respondents will consider spending time to research information if it is not readily available or spending time trying to understand questions that may not be clear.

As indicated in this paper, many issues may arise in the designing of a questionnaire.

Some, such as question wording, question structure, instruction content, navigational instructions are related to comprehension. Others, such as multiple sources of information and record retrieval are related to the retrieval of information. Still others, such as the need to coordinate findings with other persons and the presence of potentially sensitive information may make it difficult for a respondent to assess the relevance of the retrieved information to the data requested. Finally, other issues related to the communication of information, such as missing response categories or technical language may make it difficult for a respondent to report information.

All of these may contribute to response burden and consequently have a negative impact on the respondent's participation and on the results of a survey.

Because of the potential impact of such issues on a survey's data quality, testing of survey questionnaires in a business environment is key to the success of a survey. Cognitive interviewing provide a greater understanding of some of these issues and allows survey takers to make the necessary improvements to a questionnaire, improvements which generally increase the clarity of the questions, reduce response burden and consequently result in increased data quality.

Does mode matter? First results of the comparison of the response burden and data quality of a paper business survey and an electronic business survey.

Deirdre Giesen[28, 29]
Statistics Netherlands,[30] Division of Methodology and Quality

In 2006 Statistics Netherlands conducted a pilot of an electronic version of the Structural Business Survey (SBS) questionnaire. This paper describes the results of a first evaluation of the effects of the electronic data collection on perceived response burden and data quality. The first results indicate that the electronic form works well. These results are preliminary, as the analyses were conducted while the pilot was still in the field.

1. Introduction

The Structural Business Survey (SBS) questionnaires measure a large number of indicators of the activity and performance of Dutch businesses. In 2004 a redesign of the SBS started. The general goal of the redesign was to reduce the costs of the survey for our organization as well as the respondents and to remain or even improve the quality level of the statistics produced. Part of the redesign was an improvement of the data collection by redesigning the content of the questionnaire and the development of an electronic questionnaire (see also Giesen en Hak, 2005 and Snijkers et al., 2006). In 2006 a pilot was conducted in which 7800 of the 70000 business in the sample were requested to fill out the form electronically.

This paper describes a first evaluation of this pilot. This evaluation was conducted in the spring and summer of 2006, while the pilot was still in the field. The goal of this preliminary evaluation was to assess as soon as possible if any changes were necessary in the development of the 2007 questionnaire. In the next section the pilot and the electronic questionnaire will be described. The following section covers the methods used for the evaluation. Finally, the results and conclusion will be presented.

2. The pilot e-SBS

Pilot groups

The design of the SBS distinguishes about 180 different groups of establishments, according to size class and industry. For each group different versions of the questionnaire and different sample and follow-up techniques are specified. For the pilot, five of these groups were selected to receive an electronic questionnaire. All other groups received the traditional paper questionnaire.

---

[28] The author can be contacted at igin@cbs.nl

[29] The author would like to acknowledge the valuable contributions made to this reseach by the members of the project team: Bart Fieten, Thieu van Kasteren, Frans Kerssemakers, Jos Logister, Martin van Sebille, Wout Slotegraaf , Max Storms, Jo Tonglet, Rachel Vis & Bud de Wittt.

[30] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

Establishments in these pilot groups may have had previous experience with the SBS questionnaire, larger firms have a higher sampling probability and the largest firms receive a questionnaire each year.

In table 1 some characteristics of the pilot groups are summarized.

Table 1: Groups for the electronic pilot

Questionnaire	Size Class[31]	N
Construction	4-9	1320
Retail	4-9	584
Employment agencies	0-3	2158
Manufacturing	5	1958
Welfare & Child Care	0-9	1780
Total Pilot group		7800

Approach

The pilot groups received a letter in March 2006 inviting them to download an electronic questionnaire on the internet. In this letter codes were provided to log in. A paper questionnaire was available on request. However, to stimulate electronic response the letter did not mention that and how the respondents could request the paper form. In May 2006 a first reminder letter was sent out, with the same information as the first letter. In the end of June a second reminder was sent with a paper questionnaire. Due to technical problems it was not possible to repeat the individual login information in this letter. The letter did mention the possibility to download an electronic questionnaire and referred to the previous letter for the necessary information to do so.

Questionnaire

The pilot questionnaire is an off-line questionnaire, that has to be downloaded by the respondents. The questionnaire consists of several screens and each screen consists of several, related, questions.

[31] Size class depends on the number of employees and ranges from 0 (0 employees) to 9 (500 and more employees).

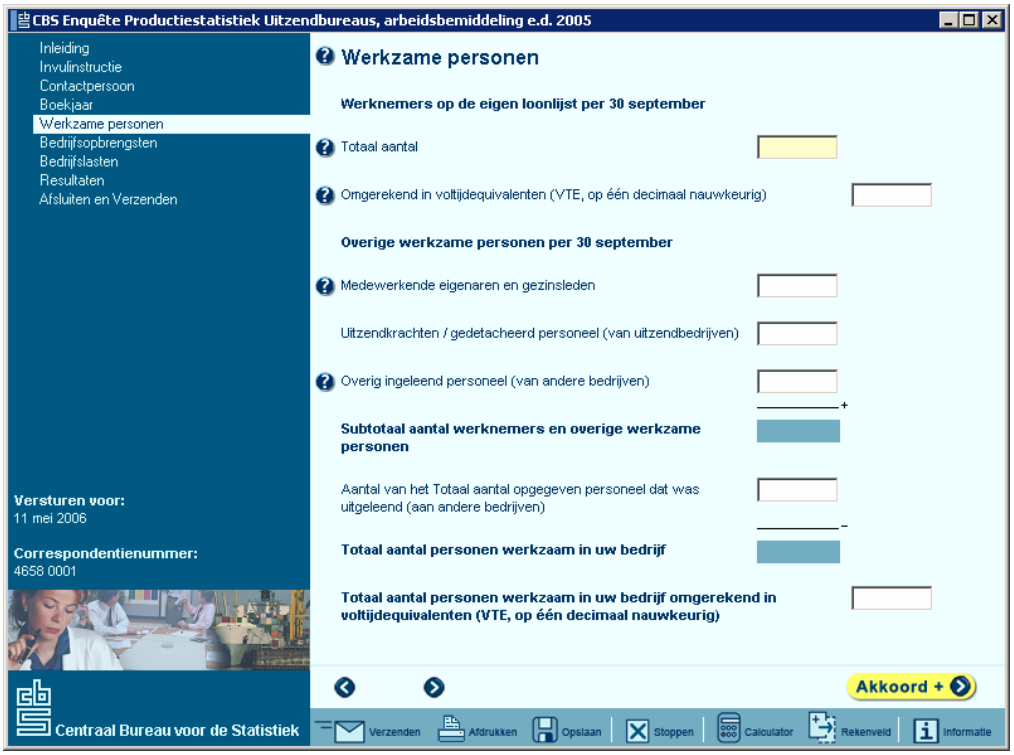


Figure 1: Example of screen of the e-SBS

See figure 1 for an example of a screen of the e-SBS. Three parts can be distinguished in each screen:. The left part of the screen is a foldable index that can be used to navigate through the screens of the questionnaire. On the bottom of the screen is a bar with different buttons, for example for saving and printing the questionnaire. The main screen consists of the questions. When additional explanatory information is available for a question, a question mark is placed before the question text. By clicking on the question mark the additional information is displayed next to the question.

Respondents can navigate freely through the questionnaire and can leave fields blank. This freedom of navigation allows respondents to complete the information that is available, while waiting for information form other sources. Questions that are not applicable can be left blank. As we know that often many items are not applicable, this was regarded more respondent friendly than a approach in which each item has to be answered. However, before the questionnaire can be sent back each screen has to be approved by clicking on the yellow so called “ok button”. Once a screen has been approved a green checkmark appears in the index header related to this screen. It is possible to change items after the “ok button” has been clicked. The checkmark than disappears and the respondent has to declare the screen “ok” again to be able to send the questionnaire.

The questionnaire asks for financial details that in the end have to add up to the trading results of the firm. The electronic questionnaire automatically adds up these numbers. Before the questionnaire can be sent back a few checks are made: at least one figure for both turn over and costs must be provided and the relations between a few variables are checked.

3. Sources and methods used for the evaluation

Following the methods used in previous evaluations of the Structural Business Surveys (f.e. Giesen & Hak, 2005) we used a mixture of qualitative and quantitative methods. The following were used

- Telephone interviews with early respondents and respondents with doubtful data (n=17)
- Field visits with retrospective interviews and observation of the response process (n=8)
- Analysis of audit trails (log files of completed questionnaires)
- Data on the use of the website
- Analyses of unit and item non response
- Information of the respondent helpdesk

Telephone interviews

The main goal of the telephone interviews was to detect quickly if there were any serious problems in the questionnaire that had to be solved immediately. For the telephone interviews we selected respondents who had sent in their questionnaires several times and who had given unlikely responses. For the interviews a protocol was used and a short report was written for each interview. 17 respondents were interviewed

Field visits

The goal of the field visits was to assess whether there were any parts of the questionnaire that caused a high response burden or had a negative impact on the data quality. Eight respondents were visited. In four cases the respondent was observed while filling out the questionnaire and in the other four cases we talked about the response process retrospectively, after the respondent had filled out the questionnaire. For both types of visits an extensive protocol was used and detailed reports were written. Two visits were video taped for demonstration purposes. Originally it was planned to visit two respondents from each pilot group, however we did not succeed in finding respondents from employment agencies in time.

Audit trails

In the so called 'audit trails' information about the respondents use of the electronic questionnaire is logged. For example, the trails contain information about the number of times certain buttons have been used or the questionnaire has been printed. Only information up to the moment of sending back the questionnaire is logged. For the evaluation we analysed the 3349 audit trails that were available in July 2006.

Use of the website

Some information is available about the use of the website where the questionnaires could be downloaded. Of certain parts of the website the number of views and visits is recorded. Every time a part of the website is clicked on is counted as a view. Several views from the same computer within 20 minutes are defined as a visit. This information can be used as an indicator how often the different parts of the website were visited. The data analysed concern the data available in June 2006.

Response information

To investigate the unit response, and the item response data of the pilot groups are compared to similar groups in the year before (when all respondents received a paper questionnaire) and to other respondents groups in 2006 who got the paper questionnaire. More detailed information about the operationalizations used is given in the results paragraph.

Help desk information

All incoming and outgoing respondent contacts with relation to the SBS questionnaires are documented, with a short description of the contents of the contact and a code. During the pilot period the helpdesk staff was asked to use additional codes. In table 2 the codes used for the analyses are summarized:

Table 2: helpdesk calling codes used for the evaluation

Code	**Meaning**
Standard codes	
71	Technical question
57	Change from electronic to paper questionnaire
19	General remark
11	Question about content of the questionnaire
Additional codes for pilot	
INST, FILL, SEND	To be used in the description of item with technical questions. Indicating respectively problems with *ins*tallation of the software, *fil*ling out the questionnaire or *send*ing the questionnaire.
RB	Response burden, for each contact in which the respondent complains about the burden caused by the questionnaire.
WP	Wants Paper: if a respondent asks for a paper questionnaire but can be persuaded to try the electronic instrument

In the hectical practice of the helpdesk, it proved very difficult for the staff to use the additional codes. For the evaluation the items registered until August were analyzed. We restricted the analyses to contacts initiated by the respondent. From these contacts the following items were select to be read carefully: all items with the codes 57 and 71 and all items that included the additional codes or words that referred tot the electronic SBS. Also we talked to members of the helpdesk staff to check whether their experiences matched the information we read in the contact items.

4. Main results

The results of the evaluation are described with respect to the following topics: 1) approach (introductory letter, download procedure etcetera) 2) the content of the questionnaire 3) usability of the questionnaire and 4) the quality of the data collected.

4.1 Approach

In our interviews most respondents said that the way they had been approached by Statistics Netherlands was correct. Only 4% of all respondents from the pilot group asked actively for a paper questionnaire. As is shown in table 3, the highest level of requests for a paper form is found in the retail group.

Table 3: reasons for asking a paper questionnaire

Questionnaire	% asked for paper questionnaire
Building	3,6
Retail	7,9
Employment agencies	4,3
Manufacturing	3,1
Welfare & Child Care	4,3
Total Pilot Group	4,1

In the contact items noted by the helpdesk we find information about the reasons respondents give for requesting paper. In the 232 items coded the following reasons were found

- 31% 'just prefers paper'
- 27% had problems downloading the questionnaire (f.e. no rights to install software)
- 18% did not have internet access
- 9% did not have a computer
- 7 % said they did not know enough about computers to fill out an electronic questionnaire
- 7% could not run the questionnaire on their computer because their configuration did not match the requirements (mostly apple computers).

Overall, few problems were found concerning the downloading and installing of the questionnaire. In the first 5 months of the survey the helpdesk had only noted 19 phone calls with technical questions. Some small problems were found in the download procedure. For example, it was confusing that the user number in the letter was presented as two groups of numbers with a space bar, whereas this space bar could not be entered on the screen while logging in.

An important problem discovered concerned the identification of individual respondents. Each questionnaire is unique and contains identification information. However, this was not clear from the letter or the questionnaire. Within one enterprise different SBS questionnaires may be sent out to different business units. During the pilot in a large building company the network manager installed one questionnaire on the company's network that was used by three business units. Three different questionnaires were sent in, but in our systems this was registered as three submissions of the same questionnaire.

The website where the questionnaires should be downloaded also contained documents with tips on how to complete the questionnaire and instruction on how to download the questionnaire. In our interviews with respondents we learned that they hardly looked at these documents. The data on the use of the website confirms this. About 20% of the people who download a questionnaire also look at the tips.

Table 4: use of the web site

Link	Number of visits
Questionnaires	
Welfare and Childcare	2298
Manufacturing	2078
Employment Agencies	1761
Construction	1412
Retail	678
Additional information	
Download instruction	224
Tips for questionnaire Welfare and Childcare	512
Tips for questionnaire Manufacturing	402
Tips for questionnaire Employment Agencies	352
Tips for questionnaire Construction	316
Tips for questionnaire Retail	205

4.2 Content of the questionnaire

In the development of the electronic questionnaire we found that the "translation" from all texts of the paper questionnaire (including headers, instruction, references to other questions) are a source of error. Unfortunately, the questionnaires used for the pilot still contained some errors. For all complex questionnaires it is difficult to make sure that they are error free. But in electronic questionnaires new sources of errors are introduced. For example, in a field in the Manufacturing questionnaire it should have been possible to enter a negative value but the field only accepted positive values.

4.3 Usability

All respondents we interviewed were positive about the usability of the electronic questionnaire. The design of the questionnaire is developed mostly according to the principles that are used by the Dutch Tax Office. Most respondents are familiar with the electronic tax forms and some explicitly mentioned that they recognized the Tax design and appreciated the similarities.

Most respondents preferred the electronic instrument over the paper version. The main advantages respondents mention are that the electronic questionnaire automatically adds up all items and that it is easier to correct errors.

The observations of the response process indicate that the response burden is significantly lower for the electronic questionnaire than for the paper ones (that were observed in previous studies). In the SBS questionnaire almost all items have to add up to the total costs and revenue of the

establishment. Often, at the end of the questionnaire the trading result that follows form the questionnaire does not match the company's figures and all items have to be checked for errors. The electronic instrument first of all prevents calculation errors within the questionnaire. However, many errors stem from items that have been wrongfully stated more than once or that have been forgotten. With the electronic instrument it is easier to navigate through the questionnaires to look for errors. Also the effects of the corrections of items on the trading result are visible at once.

The field visits also showed some points where the usability could be improved. The main problems found were:

- Respondents hardly read the question clarification (presented behind the question mark button).
- The question clarification could not be printed. This was problematic for respondents who printed out the empty questionnaire to fill it out themselves on paper or to ask others to fill out (parts of) the questionnaire.
- Some respondents seemed to use the "OK button" as a means to navigate through the questionnaire in stead of as conscious decision that the page had been fully completed. In one case we observed that a respondent by accident sent in a questionnaire in which certain pages were not completed yet.
- Respondents expected more controls in the questionnaire.
- Respondents asked for routing to make the questionnaire shorter.
- Depending on the screen configuration of respondents it could occur that they had to scroll vertically to see all items on a page. We observed how a respondent did not see that the page had to be scrolled.
- Respondents do not understand what the use Calculation aid is and don't use it. The Calculation aid allows respondents to specify the posts that are used to calculate an item. In the paper questionnaire respondents often write down sub-items in the margins of the questionnaire and the Calculation aid was meant as an alternative for that.
- It should be possible to send in an improved questionnaire. This was possible in the pilot, but the receiving systems were not able to process these questionnaires.

The audit trails provide interesting information about how often certain features of the questionnaires were used. Mostly this information confirms impressions from our observations. For example, the data show that only 60% of the respondents ever used a clarification button. This means that 40% of the respondents fill out the questionnaire without any additional information about the definitions used by Statistics Netherlands. The data also show that the calculation aid is indeed hardly ever used. In table 5 the data on use of the questionnaire features are summarized. Note that the audit trails only contain information about the use of a questionnaire up to the moment that it is sent. Often, respondents print the questionnaire after sending. This is not registered by the audit trails.

Table 5: Data on use of questionnaire features from the audit trails

	% ever used	If used: how often (mean)
Buttons		
Question clarification	60	12
General information	14	1
Calculation aid	8	4
Calculator	6	2
Stop	19	2
Save	33	2
Approve	100	42
Page forward	81	20
Page backward	75	24
Index	64	27
Print	49	2
Activities		
Changing fields	89	11
Sending via e-mail	95	1
Sending via internet	6	1,2

4.4 Data quality

In this first quick evaluation of the pilot we looked at the unit and the item response to assess the data quality. Given the time frame for this project it was not possible to look at the plausibility of the responses.

Unit response

In the pilot study, mode has not been randomly assigned, which makes it difficult to assess the effect of mode on the response. The pilot groups consist of specific branches of industries and size classes. Thus, any differences found between the pilot groups and the other groups might be caused by other characteristics than just the mode of the questionnaire. It is also possible to compare the pilot groups in 2006 with similar groups in 2005 (when all respondent received a paper questionnaire). Unfortunately, between these two years more things have changed than just the mode of data collection. An important change was that the timing of the reminder letters has been advanced in 2006.

In table 6 data are presented that allow us to compare the pilot groups both with similar groups in the year before as well as with the other groups in the same year. As the data collection was still going on when this evaluation was done, we had to restrict the analyses for the 2006 data to the

situation at 105 days after the start of data collection. At that point in time 38% of the pilot groups had returned their questionnaire, and 36% of the other groups had done so. Interestingly, the overall response in 2006 at 105 days is notably higher with 36% than the response at that same time the year before (22%). In 2005, a similar level of response was reached twenty days later. Probably this is caused by the fact that the first reminder letter was sent out much earlier in 2006. The response figures here give confidence that the introduction of the electronic mode has at least not negatively impacted the unit response. Possibly, further analyses can even reveal a positive influence.

Table 6: Unit response

	% questionnaires returned		
	2005		2006
	at 105 days	at 125 days	at 105 days
Pilot group (e-forms)	24%	32%	38%
Others (paper forms)	22%	37%	36%
Total	22%	36%	36%
N	61858	61858	65447

Item response

The observations in the field had led to some concern about the item nonresponse in the electronic instrument. Therefore we studied the overall item nonresponse in the pilot questionnaires and the item nonresponse of items that might be on the part of the screen where respondents have to scroll to. Item nonresponse is generally high for all SBS questionnaires as many items are not applicable to every respondent. In Table 7 the overall item nonresponse of the pilot groups in 2006 is compared to the item nonresponse of the similar groups in 2005.

Table 7: Item nonresponse

Questionnaire	Number of items	Item nonresponse 2005	Item nonresponse 2006
Construction	175	57%	57%
Retail	150	49%	52%
Employment agencies	95	73%	76%
Manufacturing	141	51%	50%
Welfare and childcare	137	67%	69%
Total	698	58%	60%

The data do not indicate a large increase in item nonresponse with the introduction of the electronic questionnaire.

In table 8 the item non response is given for variables that might not be visible without scrolling. Whether this is the case depends on the respondent's screen configuration. Overall we do not see that these items show a higher item non response. It must be noted that we don't know whether

the respondent had to scroll for an item. If we could restrict the analyses to respondents who did have to scroll a difference might be found. Only for the employment agencies a clear difference was found. In 2006, with the electronic questionnaires the item non response for Item AFSCHRG110000 (a depreciation item) increased from 46% to 74%. Further analyses are needed to assess whether this large increase in item nonresponse is indeed caused by mode. It does not seem likely that only one item shows this effect.

Table 8 Item nonresponse of possible "scroll" items

Item	Item nonresponse	
	2005	2006
Building		
VERKOOP21100	92,8	93,3
Manufacturing		
ONTVANG100000	87,7	87,8
OPBRENG110000	86,9	84,4
BEDRLST349900	34,4	37,4
OPBRENG111000	95,5	95,5
OPBRENG113000	98,6	98,8
BEDRLSH349900	96,6	97,1
Retail		
VERKOPH211200		
Flowers and Plants	55,7	62,9
Animal necessities	96,4	91,8
Other non-food articles	78,1	64,2
Employement agency		
AFSCHRG110000	45,8	74,4

Conclusions and recommendations

The main conclusion from this evaluation is that we can proceed with the further development of the electronic SBS questionnaires for all respondents in 2007. There is no reason to believe that large scale implementation of the electronic questionnaire will cause large problems with either the respondents, our helpdesk or the data quality.

Many smaller and larger recommendations for the further development of the presentation and content of the questionnaire followed from the evaluation. An important conclusion is that the development error free electronic questionnaires requires more time and effort than the development of paper questionnaires.

References

Giesen, D & Hak, T. (2005) *Revising the Structural Business Survey: From a Multi-Method Evaluation to Design.* Paper presented at the Federal Committee on Statistical Methodology Research Conferences, Arlington, Virginia, November 14-16 2005. http://www.fcsm.gov/05papers/Giesen_Hak_IXB.pdf

Snijkers, Ger, Onat, Evrim, Vis, Rachel, Tonglet, Jo & 't Hart, Robert (2006) The Dutch Annual Business Inquiry: Developing and Testing an Electronic Form. Proceedings of the 10th International Blaise Users Conference https://www.blaise.com/ibuc2006papers/papers/251.pdf

What the eye doesn't see: A feasibility study to evaluate eye-tracking technology as a tool for paper-based questionnaire development

Lyn Potaka

Acknowledgment: The author would like to acknowledge the involvement of Peter Brawn (Access Testing Centre) in producing the work discussed in this paper.

Eye-tracking technology is often used as a tool to help evaluate website designs. However, fewer studies have examined the use of eye tracking technology as a tool for questionnaire evaluation. This paper discusses a qualitative study in New Zealand, which tracked eye movements to evaluate the layout and presentation of a self-complete paper Census questionnaire. Respondents' visual behaviour was recorded, using eye-tracking technology, to identify which elements on the form were drawing respondents' attention and which elements were being ignored. 16 respondents were observed, and their eye movements were recorded as they filled out the Census questionnaire during 30 minute interviews. Follow-up questions were also asked. Results from this study were consistent with findings from studies using cognitive interviewing, and helped confirm that respondents can skip over important information on the form. This study further confirmed the importance of questionnaire layout in assisting respondents to complete questionnaires correctly and supported other research that suggests eye-tracking technology can be a useful tool in informing the design of self-complete questionnaires.

1. Introduction

Eye movements and their link with cognition have been of interest to cognitive and social psychologists for many years. Various systems have been developed to track and measure people's eye movements and have become more sophisticated over time (Ellis, Candrea, Misner, Craig, Lankford, & Hutchinson, 1995). With these advances in technology, the use of eye tracking systems has expanded and this technology is now becoming popular as a means to evaluate web site designs and computer usability (Ellis et al, 1995).

Eye gaze technology is also of interest to survey organisations, in helping to inform the development of survey questionnaires. Although survey researchers already use this technology to investigate visual design effects, this work has largely involved studies using electronic questionnaires (see for example, Tourangeau, Couper & Galesic, 2005). Although the results of usability research on electronic questionnaires may be applied to the layout of questionnaires in general, it has been noted elsewhere that reading from a computer screen can differ in important ways from reading text on paper (see for example, Spool, Scanlon, Schroeder, Snyder & Deangelo 1999; Neilson, 2000). However, there have been few published examples of eye gaze technology having been used to evaluate the layouts of self-complete paper questionnaires.

One notable exception is the study conducted by Redline and Lankford (2001) in which they specifically examined the utility of eye gaze technology using self-complete paper questionnaires and concluded that this equipment was potentially useful for this purpose (Redline & Lankford, 2001).

2. Background

Words printed on a questionnaire do not work in isolation to communicate meaning to respondents, and the visual design and layout of the questionnaire is also important in conveying meaning (Dillman, 2001; Redline & Lankford, 2001). Therefore, any evaluation of a questionnaire needs to consider these aspects together.

However, the way that respondents see and process information when reading from paper is not easy to measure via traditional approaches to questionnaire evaluation, such as cognitive testing. Reading behaviour is difficult to observe reliably using this approach and self-reports on preferred layouts can sometimes be misleading (eg. Spool et al, 1999; Andre & Wickens, 1995). Therefore the questionnaire design team at Statistics New Zealand were keen to evaluate eye gaze technology as a way to assure the quality of Statistics New Zealand's self-complete paper questionnaires. A small-scale feasibility study was initiated for this purpose and the 2006 Census form designs, under development at the time, were seen as appropriate vehicles for this research.

At the time the study took place, no laboratories with eye tracking equipment were available for hire within New Zealand and therefore the work was conducted in collaboration with Access Testing Centre in Sydney, Australia. Statistics New Zealand questionnaire designers worked in partnership with the centre's Perceptual Psychologist (Peter Brawn), a specialist in the use of eye-tracking equipment.

The study was necessarily a small-scale project, intended to evaluate the feasibility and value of this technology within the Statistics New Zealand environment. As such, funding on this project was limited, and evaluations were restricted to a small number of key issues.

Although the primary objective for this research was to evaluate this technology itself, a secondary objective was to draw conclusions about the design of the forms under development for Census 2006, and in particular the visibility of key elements on those forms. The three features of particular interest were: Reminder Bubbles, Routing Instructions and Alpha-Numeric boxes.

Reminder Bubbles

Reminder Bubbles were first introduced to New Zealand Census forms in 2001, as a way to remind respondents that they should mark their answers as a horizontal line across the oval marking space. This horizontal marking format was used to minimise scanning errors which would occur when respondent marks extended beyond the defined area. However, as respondents often reverted to ticking behaviour part way through the form, it was hoped that occasional reminders would encourage the correct marking behaviour.

These reminders appeared in 'bubbles' or oval shapes with a lighter background than the main question area (as shown in Figure 1). Bubbles were strategically placed throughout the form wherever there was sufficient space and positioned to the right of question text and response categories. However, the questionnaire developers were concerned about the visibility of these bubbles because they appeared outside of the main navigational path (on the left hand side of the column) and so would often remain outside the respondent's immediate foveal view, or attentive field of vision. This is the area where vision is sharp and estimated to be 9 characters in width, or two degrees either side of the eyes' fixation point (Kahneman, 1973).

Figure 1: Reminder Bubble

Routing Instructions

Another element of interest to researchers was the salience of routing instructions. Routing instructions were used throughout the form to signify when a respondent's did not need to answer subsequent questions. However, such instructions are always problematic in a self-complete questionnaire as a proportion of respondents miss the instructions or route incorrectly. These errors in navigation are described as errors of omission (when a respondent acts upon a routing instruction incorrectly and skips intervening questions), and errors of commission (when a respondent fails to act upon a routing instruction and does not skip when they should). Although the number of these errors can be reduced through a variety of design strategies (see, for example Redline & Dillman 2001; Redline, Dillman, Dajani & Scagg, 2003) distinctiveness is known to be a key factor in optimising the likelihood that such instructions will be seen and acted upon. This research demonstrates that routing instructions are most effective when they are made salient through variations in the use of size, colour and shape (Jenkins, Dillman, Carley-Baxter & Jackson, 1999).

.

For most questions on the New Zealand Census forms routing instructions were presented alongside each individual option that required the respondent to skip. However for some questions this layout became visually confusing as multiple instructions competed for the respondent's attention (as shown in Version A, Figure 2).

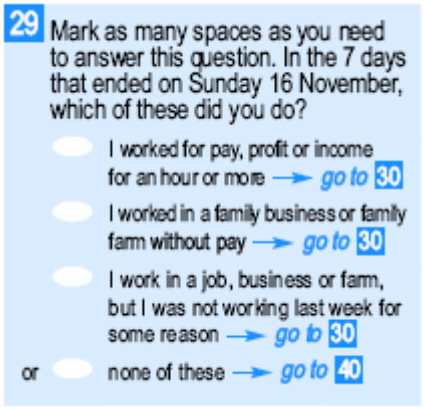


Figure 2: Version A. Routing instructions for all response options

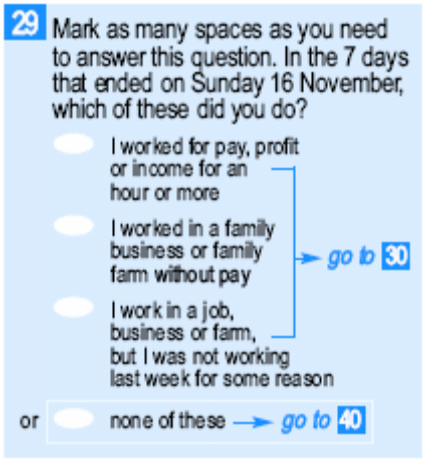


Figure 3: Version B. Bracketed routing instruction

To reduce the density of text, a new format was introduced using shorter line lengths for response text and a single routing instruction for all options routing to the same question number. This routing instruction used a bracketed arrow to encompass all of the applicable categories (as shown in Version B, Figure 3). Although this design reduced the visual complexity of the question, the designers were concerned that respondents would not always understand this format, particularly when selecting options which appeared in the middle of the bracketed area. In addition this format increased the visual distance between the relevant option and the relevant instruction. This meant that instructions would sometimes appear outside of the respondent's immediate foveal view.

Alpha-Numeric boxes

In the 2006 development alpha-numeric boxes became a standard for all write-in questions on the forms. In previous censuses, only a very small number of questions had used these boxes. There were a number of compelling reasons to include alpha numeric boxes, most importantly the faster capture and processing of results. However, the boxes were visually dominating due to their size and the large amount of white space required. In addition, there were often difficult choices to make in deciding the number of boxes that should be included. To maximise the quality of response, a sufficient number were required so that respondents could record their answers in full, without truncation or abbreviation. However limited space meant that boxes had to be kept to a minimum.

For these reasons the developers wanted to test two versions of the alpha numeric boxes. In one version boxes were 'indented' and appeared beneath the relevant category (as shown in Version A, Figure 4). This was the preferred option as boxes did not obscure later response options. This design also kept the navigational path clear and uncluttered, while conveying more clearly to respondents that the write-in answers were linked to a particular response. However, this format meant that there would be fewer boxes in each row. The second version tested (Version B) had boxes 'aligned' hard against the left hand margin, so that the maximum number of boxes would

fit within each row (as shown in Version B, Figure 5). The developers were particularly concerned that response options appearing underneath the alpha numeric boxes would be missed with this version.

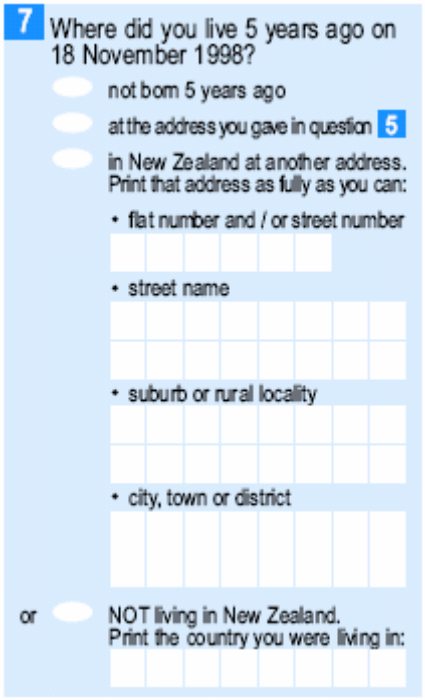


Figure 4: 'Indented' Version

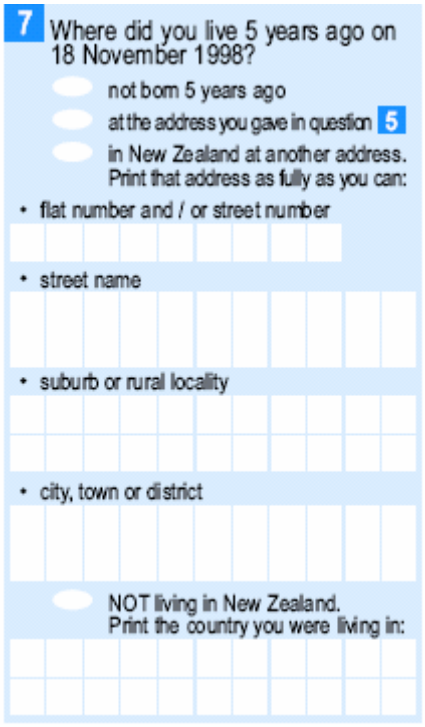


Figure 5: 'Aligned' Version

3. Methodology

3.1 Materials

Eye gaze hardware and software, including optical unit, head tracking device and computer recording equipment were used for this study.

Two different versions of a four page 'individual' census questionnaire, consisting of 47 questions about a range of topics were used. Version A and Version B differed in their visual layout for alpha-numeric boxes and routing instructions (as explained in the earlier section of this paper).

3.2 Respondents

Eye gaze data was obtained from 16 respondents who had answered an advertisement in a Sydney newspaper asking for volunteers. Because the study was originally designed to test the feasibility of using eye-gaze technology to evaluate forms, no strict sampling criteria was applied. The only restriction placed on the recruitment of respondents was that all participants needed to have been New Zealand citizens or have previously lived in New Zealand, so that the form would make sense to them. The final sample consisted of an even split of male and female respondents aged between the ages of 18 and 55.

3.3 Procedure

Respondents were randomly assigned to one of two groups, each using a different version of the questionnaire and interviewed separately.
For each interview, the eye tracking camera was positioned underneath the computer monitor and the paper Census questionnaire was presented as four individual pages, placed vertically in front of the computer screen.

Respondent interviews were approximately half an hour in length and each interview involved:

- a brief introductory explanation, including assurances of confidentiality and time to connect up the software and calibrate the respondent's eye
- approximately 10 minutes for the respondent to complete either version A or version B of the questionnaire while eye movements were tracked and displayed on a computer screen in an adjacent observation room
- 10 minutes of follow-up questions with an interviewer to determine the respondent's subjective experience of completing the questionnaire.

4. Results & Discussion

4.1 General

Several factors made this research less than ideal. The questionnaire was placed in an upright position so that respondents' eye movements could be tracked, and this was not a natural way to fill out a questionnaire. There was also some loss of data when respondents leaned in too close to the questionnaire, which limited conclusions.

However, the results of this study reinforced findings from Statistics New Zealand's cognitive research work on the Census forms and confirmed findings from other research (eg. Redline & Lankford, 2001; Dillman, 2001). Respondent behaviours, such as re-reading questions and navigating backward in the form to check previous answers, could be observed when watching respondents' eye movements in real time. This information, combined with dwell time data was useful in helping the developers to determine which questions were causing respondents the most difficulty.

Real-time observations also suggested that respondents didn't always read all of the information presented to them for each question, before providing an answer. For example, respondents' eyes sometimes only tracked across parts of the question stem before moving on to the response categories and commonly skipped over any instructions accompanying the questions. On other occasions respondents appeared to superficially scan the response options before quickly selecting an answer. These observations also suggested that some respondents are more likely to read carefully than others, which concurs with other research (Jenkins & Dillman 1993; Krosnick & Alwin, 1987).

4.2 Routing Instructions

In general, the results of this study indicated that the routing instructions were performing adequately. For this study, no errors of omission and very few errors of commission were recorded. There was an equal level of error for both versions of the questionnaire.

However, it was interesting to note that, where errors were recorded, respondents had sometimes observed the routing instruction but simply did not skip as required. This finding would therefore support the conclusion from earlier work (Redline & Lankford, 2001) that respondents who do not immediately act on instructions may fail to recall them later.
It was difficult to determine why respondents failed to route immediately. In follow-up questioning, some respondents suggested that this was because they had wanted to 'check' that they had selected the correct category. These respondents said that they had wanted to read the remaining options to be sure that no other options applied.

This result was inconclusive, but suggested to the designers that the traditional format, where routing options are presented alongside each individual option, may be a better choice to encourage immediate routing and may give respondents greater confidence that they are routing correctly.

4.3 Reminder Bubbles

Analysis of the eye gaze patterns for reminder bubbles was informative. These reminder bubbles were regularly missed by respondents. However, bubbles presented for some questions were more likely to be missed than others. For example, the bubble appearing in question three on the individual form (the first question where a bubble appears) was observed by almost all respondents in the sample. This finding might suggest that there are characteristics of individual questions which will increase or decrease the likelihood that bubbles will be observed.

Such characteristics might include the position on the page or the complexity of the individual question. It is possible that respondents may perceive the bubbles early in the questionnaire and then, after reading the information, respondents may make a judgement that they can safely ignore this information later in the form. This would support other reports on the ways that respondents read and process instructions (Wright, 1980; Dillman, 2001). Alternatively, the density of the text within some questions may mean that reminder bubbles are sometimes competing with other information and may consequently be ignored. The difficulty of the question is a further factor that may influence whether bubbles are observed. For difficult

questions respondents may become absorbed with the answering task and fail to attend to peripheral information.

The conclusion from these results was that reminder bubbles could be useful in some situations, but should be reserved for non-essential information.

4.4 Alpha-Numeric Boxes

An evaluation of respondents' eye movements for questions that contained alpha-numeric boxes confirmed that respondents sometimes failed to observe response options which appeared below the alpha-numeric boxes (as shown in figure 6). However, this occurred for both the 'indented' version and the 'aligned' version of the form.

Designers were unable to draw any firm conclusions from this finding due to the small number of respondents involved in the study and the likelihood that results were biased by differences between the two comparison groups. For example, respondents completing the 'indented' version of the form were more inclined to 'skim' question text, and respondents in the 'aligned' group were more likely to read questions thoroughly. This fits with previous work that suggests that respondents differ in their motivation to read and answer carefully (Jenkins & Dillman, 1993; Krosnick & Alwin, 1987).

However, the data also revealed that respondents were less likely to miss options when they were actively seeking out an answer from amongst a list of options. For example, the 'live alone' category shown in figure 6 was missed less often than other questions. This finding would indicate that when a respondent already has a preformed answer they will make a more thorough visual search of the information to find the appropriate category. However, when answering questions where they have no preformed answer, such as non-factual or attitude questions, there may be a greater risk that respondents will miss options appearing underneath the alpha-numeric boxes.

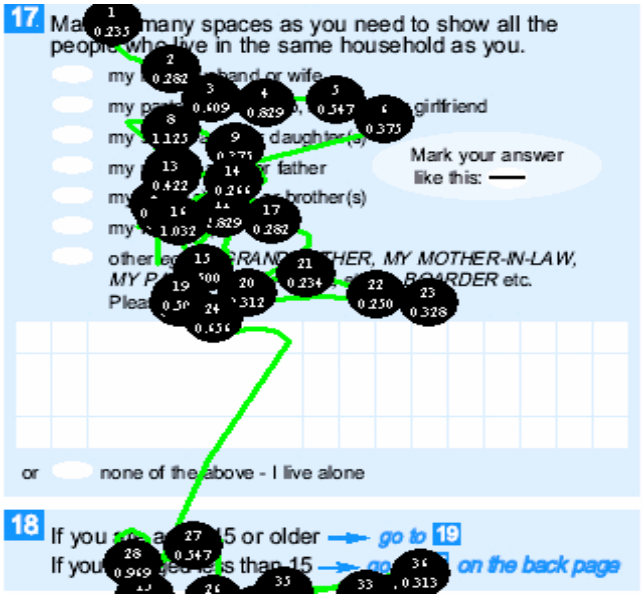


Figure 6: Example of respondent's eye movements showing no fixation on instruction bubble or response option appearing underneath alpha-numeric boxes.

5. Conclusions

5.1 Forms Design

This study helped to confirm the importance and impact of visual design on the quality of response and supported existing knowledge and research in this area.

However, the small number of respondents involved in this study limited the conclusions that researchers were able to make and a direct comparison of two different design formats was not appropriate for a sample of this size. However the work was useful in allowing researchers to draw some conclusions about the designs that would work most effectively.

Although there were some interesting indications that certain question types may have contributed to the results (such as the positioning, complexity and density of questions), it is clear that further work is required to confirm these hypotheses and to further identify the question characteristics most likely to influence results.

5.2 Eye-tracking Technology

The author of this paper would concur with the conclusion reported by Redline (2001), that eye-gaze technology is a promising tool for the evaluation and design of paper self-complete questionnaires. In particular, this technology would appear to be most useful in helping to evaluate individual designs. Such an evaluation can tell designers which questions are most likely to be read, which instructions are most likely to be missed and how best to present the required information on the page.

However, the tool also has a great deal of potential in helping to expand questionnaire design knowledge generally. Such a tool can help to identify the question characteristics which may impact on reading behaviour and help to identify the visual designs that are likely to work most effectively for paper questionnaires.

Although the technology retains some of the limitations noted in Redline's earlier paper (2001), such as a loss of data when respondents lean in too close, it is likely that these issues will be resolved over time.

Information obtained from a study of eye movements may not provide questionnaire designers with the same richness of information that traditional evaluation techniques, such as cognitive testing can provide. Overall however, eye gaze research can complement and enhance knowledge gained from more traditional methods. It can also be useful for convincing for project sponsors because it provides objective, unbiased data.

In summary, this study further confirmed the importance of questionnaire layout in assisting respondents to complete questionnaires correctly and supported other research that suggests eye-tracking technology can be a useful tool in informing the design of self-complete questionnaires.

References:

Andre A D & Wickens C D (1995*). When Users Want Whats Not Best For Them*. Ergonomics in Design, 3 (Oct) 1

Dillman D A (2000). *Mail and Internet Surveys: The Tailored Design Method.* 2nd Ed. New York: John Wiley & Sons.

Ellis S, Candrea R, Misner J, Craig C S, Lankford, C P, Hutchison T E (1998). *Windows to the Soul? What Eye Movements Tell Us About Software Usability.* Paper presented at the Usability Professional's Association, Washington DC.

Jenkins C & Dillman D (1993). *Combining Cognitive and Motivational Research Perspectives for the Design of Respondent-Friendly Self-Administered Questionnaires*. Paper presented at American Association for Public Research Annual Meeting, St Charles, Illinois (May).

Jenkins C, Dillman D, Carley-Baxter L, & Jackson A (1999*). Making Invisible the Visible: An Experiment with Skip Instructions on Paper Questionnaires.* Paper presented at the meeting of the American Association for Public Opinion Research, Orlando, FL.

Kahneman, D. (1973). *Attention and Effort.* New Jersey: Prentice Hall.

Krosnick, J & Alwin, D F (1987). Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement, *Public Opinion Quarterly*, *51*.

Neilson J (2000). *Designing Web Usability***.** Indianapolis: New Riders Publishing.

Redline C & Dillman D (2001). The Influence of Alternative Visual Designs on Respondents' Performance with Branch Instructions in Self-Administered Questionnaires. Groves R, Dillman D, Eltinge J, Little R (eds). *Survey Nonresponse*. New York: John Wiley and Sons, Inc.

Redline C, Dillman D, Dajani A & Scagg M, (2003). The Effect of Altering the Design of Branching Instructions on Navigational Performance in Census. *Journal of Official Statistics, Vol 19, 4*

Redline C & Lankford C (2001). *Eye-Movement Analysis: A New Tool for Evaluating the Design of Visually Administered Instruments (Paper and Web)*. Proceedings of the Section on Survey Methods Research. American Statistical Association.

Spool J M, Scanlon T, Schroeder W, Snyder C & Deangelo T (1999). *Web Site Usability: A Designer's Guide*. San Diego: Academic Press.

Tourangeau R, Couper M P & Galesic M (2005). *Use of Eye-tracking for Studying Survey Response Processes*. ESF Workshop on Internet Survey Methodology, Dubrovnik.

Wright, P (1980). *Strategy and Tactics in the Design of Forms*. Visible Language, XIV 2, pp. 151-193.

Roundtable Discussion: Preliminary Steps in the Direction of a Research Agenda for QUEST's Second Decade

Jim Esposito, Moderator [32]
Debbie Collins, Group A Spokesperson and Benoit Allard, Group B Spokesperson

A. Introduction

Looking back over the content of various roundtable discussions we have had over the past ten years, the QUEST group certainly has covered a lot of territory. For example, at **QUEST1997** (Statistics Sweden, Orebro), we had very productive group discussions that focused on three general sets of issues:

- Policy and client communications concerning questionnaire development
- A professional profile for questionnaire-evaluation practitioners
- Methodology with specific regard to testing/evaluating questionnaires

Then at **QUEST2001** (US Bureau of the Census, Washington, DC), during a workshop that was organized to address three general themes (questionnaire-evaluation practice, pragmatics and theory), we again enjoyed very productive group discussions on the following issues/topics:

- Cognitive interviewing
- Archiving
- Questionnaire evaluation theory

As a community of practitioners, we have made substantial progress in many of these areas over the past ten years (e.g., QDET conference and monograph; JOS Special Edition; Statistics Sweden's manual on questionnaire development and evaluation; and a variety of other papers, articles and books). But the 2007 workshop had a different composition: A new generation of practitioners has joined the QUEST community and, so, this workshop was viewed as a good opportunity to ask attendees to share their thoughts (within the context of roundtable discussions) on what they believe our general agenda should be for the next ten years—if only to reinforce the importance of research issues proposed earlier and/or research efforts currently underway. This "agenda" would be useful as a means of focusing attention on important research issues and topics—*and should not be viewed as prescriptive or binding in any way*. With that introduction as preface, attendees were challenged to chart a course for the future.

B. Roundtable Structure

The roundtable was structured as follows:

- Review of roundtable objective and process (moderator: 5 minutes). *Roundtable objective***:** To identify a set of important research issues/topics for the next decade.
- Break-out into two subgroups [A and B] for separate discussions (about 30 minutes): *Brainstorm* research issues/topics and *rank* issues/topics in order of importance.
- Full group re-assembles and subgroup spokespersons present a brief summary of their group's discussion (15 minutes, total): Presentation of highly ranked research topics.

[32] Please allow me to apologize here, at the outset, if I have overlooked or misrepresented any of the topics, issues or ideas proposed during the course of our roundtable discussion. My role as moderator made it difficult to take detailed notes and my memory for extemporaneous discussions is not quite as good as it used to be.

- Compare/contrast summary reports and note commonalities/differences (full group discussion: 10 minutes): Identify issues/topics for which there is general agreement as to importance.
- Discussion of tentative agenda for second decade (15 minutes): Full-group discussion of possible initiatives (or work in progress).

C. Research Topics/Issues as Identified and Ranked by Groups A and B

As noted, roundtable participants broke-out into two subgroups for the purpose of identifying and ranking important research topics/issues. The results of that process appear in **Table 1**. [Note: See the addendum for brief descriptions of a subset of those topics/issues identified in Table 1 and subsequently elaborated upon by roundtable participants.]

D. Discussion

As one can see from Table 1, even though the two groups were working independently, there was a considerable degree of overlap in the topics/issues identified and in the rankings of these topics/issues (e.g., an interest in business surveys and establishment response models, cognitive interviewing standards and procedures, mixed-mode effects and research, and the triangulation of findings from various evaluation methods); and, as one might expect, there were some notable group differences as well (e.g., an interest in ways to accelerate the questionnaire evaluation process, interviewer effects, and standards "enforcement"). Clearly, there is general agreement as to what some of our priorities might be and no shortage of work to be done. That said, what might an agenda look like and how might we proceed?

At the top of our agenda, as noted above, would be the following topics/issues:
- Business surveys (and related topics like web surveys and visual design)
- Cognitive interview methodology
- Mixed-mode effects and research
- Combining/triangulating findings from evaluation-research methods (including data analysis)

Now, given the topics/issues listed above, how might interested members of the QUEST community respond to this non-binding agenda? One way of doing so would be for members to make greater use of the QUEST website. Posting research papers would help members to identify individuals with common research interests and might shed light on how different survey research organizations have approached and solved particular problems. At the very least, such postings would help members to identify other researchers who have expertise or knowledge with respect to specific survey types (e.g., household vs. establishment surveys) and administration modes (e.g., CATI, CAPI, SAQ and web), with respect to specific survey issues (e.g., dependent interviewing; item and unit nonresponse; optimal question design; questionnaire translation methods; usability testing; visual design), with respect to specific evaluation methods (e.g., cognitive interviewing; behavior coding; eye-tracking techniques; split-panel testing) and with respect to specific survey content (e.g., health/disability issues; labor force issues).

[*Text continues after the table that appears on the next page.*]

Table 1: Ranked List of Research Issues and Topics [Groups A and B]

Rank**	Research Topics/Issues [**Group A**]
H	▪ Business surveys
H	▪ Combining methods (e.g., usability research; cognitive interviews; et cetera)
H	▪ Confidentiality/security of interview data
H	▪ Mixed mode research (strengths and weaknesses)
H	▪ Questionnaire evaluation reports: What content to include?
H	▪ Standardization of questionnaires
H	▪ Visual design
H	▪ Cognitive interviews (i.e., analysis; practical standards)
	▪ Accelerating the questionnaire evaluation process (e.g., quick-turnaround, rapid-response testing)
	▪ Archiving
	▪ Educating and/or training sponsors/non-experts about evaluation methods
	▪ Interdependence of researchers/practitioners and subject-matter experts
	▪ Qualitative vis-à-vis quantitative evaluation data
	▪ Quality indicators for monitoring data collection
	▪ QUEST website: Tool for research
	▪ Sampling (e.g., criteria for selecting/recruiting research participants)
	▪ Standards "enforcement"
	▪ Telephone surveys: Future of this mode.
Votes/Rank	**Research Topics/Issues [Group B]**
6 [H]	▪ Data analysis
6 [H]	▪ Web surveys
5 [H]	▪ Cognitive interviews: Need standards.
3 [H]	▪ Cognitive interviews (i.e., interviewing behavior; training)
3 [H]	▪ Evaluation methods: Validation, triangulation and generalization
3 [H]	▪ Generalization: Bad—good question
3 [H]	▪ Mixed-mode effects
2	▪ Businesses, establishments
2	▪ Choosing research subjects/participants
2	▪ Respondent process/subgroups/psychological process
1	▪ Concepts review
0	▪ Definitions of "our methods" (e.g., lexicon)
0	▪ Interviewer effects

**** Note:** Unfortunately, the tally of votes for Group A was not recorded; however, the ranking of the various topics/issues was preserved. High-priority topics/issues for both groups are designated with the letter "H".

Another means of advancing this agenda would be to start laying the groundwork for QDET II, a follow-up to the international conference on **Q**uestionnaire **D**evelopment, **E**valuation and **T**esting methods that that was the inspiration of one of QUEST's esteemed emeritus members, Jennifer Rothgeb. Before doing so, however, we need to have something to offer that goes above and beyond what researchers reported at the first QDET conference. It will be important to be able to demonstrate to sponsors and to other survey practitioners that our evaluation methods are effective not only in identifying where questionnaire design problems exist but also in establishing that the design changes we make subsequently are effective in reducing measurement error. That is a daunting challenge, and one that we have not quite mastered as yet.

Whatever the future holds, the QUEST community seems well positioned to respond and to contribute. We have produced a considerable body of research to date and we have an enthusiastic and talented group of individuals working on a broad range of survey-related issues. A lack of imagination is the only obstacle that could prevent us from moving forward.

ADDENDUM: Brief descriptions of topics/issues identified during roundtable brainstorming. [33]

Archiving. We should consider developing an archive of evidence documenting the performance of survey questions and questionnaires. Such an archive would contain findings from pretests, and could be modeled along the lines of the Q-Bank. [DC, Group A].

Cognitive Interviews (i.e., analysis; practical standards). We need some mechanism for documenting our practices, potentially agreeing on protocols for their conduct, for example, minimum standards for analysis. [DC, Group A]

Cognitive Interviews: Need Standards and *Cognitive Interviews (i.e., interviewer behavior; training).* Though there is some documentation on the practice of cognitive interviewing (most notably, the 2004 Gordon Willis book), there are not yet universal standards for the testing process—such as developing the protocols and probes, the mechanics of embedding the probes into the instrument, knowing how and when to probe, duration of interviews, and how to conduct analysis. Development of such standards across agencies may serve to bring the best of current practice into focus, and aid in the training and the continuing evaluation of other staff within a given agency in order to maintain consistent high quality testing results. [JP, Group B]

Combining methods (e.g., usability research; cognitive interviews; et cetera). Traditionally, usability testing and cognitive testing have belonged to different branches, often performed by different people within the organization (usability tied up to IT and cognitive testing stemming from the psychology/sociology field). We find that there is a lot to gain by combining the two traditions and the knowledge related to them. [BH, Group A]

Concepts review. By this I meant "review of old concepts" or, if one wants to be blunt, "hacking at sacred cows". While new/revised questionnaires undergo testing before they get implemented, some older surveys are stuck with old questions that are resistant to change. These venerable surveys (e.g. Census, Labour Force Survey in the case of Statistics Canada) have an influence on

[33] These descriptions have been reproduced with only minor editing. Our thanks to those individuals who were kind enough to forward descriptions after the workshop.

the new ones, for reasons of comparability. If questionnaire testing can demonstrate that these old questions can be improved, can it also demonstrate that the improvement will outweigh the inconvenience of losing comparability with the old benchmarks? [BA, Group B]

Definitions of "our methods" (i.e., the lexicon). It would still be good for us to have a lexicon; something manageable, maybe 20-25 pages versus 140 pages (e.g., the European Handbook). We often say that when people refer to "cognitive interviewing" they mean different things. Everyone could have a copy before to read and then make changes or add new terms/words to the list. For example, "usability test" is not in the last version from 2003. [BH, Group B]

Educating and/or training sponsors/non-experts about evaluation methods. This point means that survey sponsors often have unrealistic views about time schedules in testing. And in multicultural and language projects like those sponsored by the EU [European Union], a questionnaire design process often starts among people (e.g., administrators) who have very little knowledge of questionnaire design and, yet, still they produce a first draft. This makes it very difficult to continue with testing and developing. [PG, Group A]

Interdependence of researchers/practitioners and subject-matter experts. The design and evaluation of survey questionnaires proceeds much more efficiently and effectively when there is an ongoing and respectful collaboration between *subject-matter experts* (e.g., survey sponsors), who are knowledgeable about the content a particular survey domain (e.g., cancer risk factors; labor force issues; civic engagement), and *survey methodologists* (e.g., researchers and practitioners), who are knowledgeable about the mechanics of designing and evaluating survey questionnaires. We need to do a better job: (a) communicating our respective roles and responsibilities to subject-matter experts, and (b) trying to understand what *their* concerns and constraints might be. [JLE, Group A]

Mixed-mode effects and mixed-mode research. There has been a "perfect storm" brewing in the in the sea of survey-based research for some time now. Face-to-face interviewing has become prohibitively expensive, landline telephone usage is declining, and self-administered forms of interviewing (e.g., web-based surveys) are becoming more viable. To counteract the negative impacts of this storm, more-and-more survey organizations are moving to mixed-mode surveys. Are we prepared? What do we know (as a community of practitioners) about designing and evaluating such surveys? [JLE, on behalf of unidentified persons in Groups A and B]

Qualitative vis-à-vis quantitative design-and-evaluation methods. Too often it seems, we, and the survey sponsors we support, tend to think of these two methodological approaches in "either/or terms"; often the distinction is made between "soft" data and "hard" data. We need to convince ourselves, and the sponsors we support, that survey design-and-evaluation research is more about process and balance. As we proceed from questionnaire development and presurvey research to survey administration and postsurvey research (i.e., the overarching *process*), we tend to move from qualitative methods (e.g., ethnographic techniques; cognitive interviewing) to a mixture of quantitative methods (e.g., split-ballot testing) and semi-qualitative methods (e.g., focus groups; behavior coding; respondent debriefings; interviewer rating forms). It is unproductive and inefficient to think in terms of one approach *or the other*, because the information needed and available from various parties at each step of the process is different; to be successful, we need to adopt a more balanced approach to questionnaire-design-and-evaluation research. One of our goals as a community should be determine if such a view can be empirically validated. [JLE, Group A]

QUEST Website: Tool for research. Using this as a resource to post best practice guidelines, examples of studies, et cetera. It could be a tool to communicate with the wider research community, not just QUEST members, to inform them about our work, et cetera. [DC, Group A]

Respondent process/subgroups/psychological process. Regarding the "subgroups" part, I was wondering if respondents could be labeled according to their response process. For example, some respondents barely read the questions and go straight for the answer boxes to figure out what is being asked. Others will try to guess at the survey's intent at every question. If we could somehow establish these types of respondents, it might be useful to evaluate questions in terms of how well they do with each type, when we fall short of coming up with a question that works perfectly for everyone. [BA, Group B]

Standards "enforcement". QUEST as a catalyst in developing best-practice guidelines for the conduct and analysis of pretesting methods like cognitive interviewing. [DC, Group A]

Standardization of questionnaires. Should we standardize wording and, if so, which wording should we use? What evidence should be used on the performance of questions to merit them being seen as the "standard" way of asking about "x"? [DC, Group A]

Telephone surveys: Future of this mode. We need to know more about how the quality is affected by the change in using telephones, from "ordinary" landlines to cell phones. Nowadays respondents are walking in the street, riding the bus, doing the dishes (or whatever) while answering the questions. Then all of a sudden the battery is running low and you have to contact the respondent again at another time. Different respondents are more or less sensitive to all of this; but, in for example living conditions, it's rather obvious that the quality is affected by what is going on around the respondent during the interview. Another thing about telephones is the changed pattern of using the telephone. For younger generations the telephone is so much taken for granted and differs from the telephone behavior of older generations, which affects the contact strategies. [GD, Group A] We might also want to consider the practical utility and feasibility of conducting cognitive interviews *over the telephone* and invest some time and effort into developing protocols for such interviews. [DC, Group A]

Visual design. An assumption is that, in the future, people use more-and-more self-administration in surveys and in many other respects. So I think more research should be done in this area and especially how to integrate it within the overall pretesting process (e.g., what kind of knowledge is needed in a survey institution and how to organize it). The QUEST group might contribute in this area universally. [PG, Group A]

Wrap-up Discussion Summary

Jack Fowler

There were three topics of discussion: issues related to the 2009 QUEST meeting, QUEST-related activities before 2009, and needed research.

1. There were five main suggestions regarding the organization of QUEST 2009:
    a. Consider shortening presentations and lengthening the amount of time for discussion.
    b. Consider identifying some key session themes and asking participants who have done relevant work to submit proposals that would fit in with the theme. The idea would be to try to create some sessions in which the presentations were more integrated, complementary and focused on the same topics.
    c. Possibly spend less time on discussions of participants' organizations so there would be more time for presentation of research and what participants are actually doing.
    d. Consider some kind of wrap-up discussion at the end of each day, or at some other reasonable break points, to give participants a chance to work on developing generalizations about what we know and what we need to learn about a specific set of issues.
    e. Schedule the meeting for a full three days of meetings; attendees should be clear that they should plan to stay for the entire meeting.

2. With respect to QUEST related activities before the next meeting:
    a. There was interest in trying to make the QUEST website more useful by having members more regularly post relevant articles and perhaps internal working papers and methodological reports. The idea would be to turn it into more of a resource for those seeking information.
    b. There was discussion about whether it was time to try to put together another conference on question evaluation along the lines of QDET. There seemed to be a division of opinion about whether or not there was enough new material available to justify a conference like that before 2009.

3. Research priorities noted include:
    a. As part of the continuing effort to document the value of better question evaluation, we need more studies that document the value of question testing. Studies that show that the results of questions that have been tested produce measurement that is more accurate or valid are particularly needed.
    b. On a related note, we need to develop and use better measures of success. Record-check studies are ideal, but rare. Split-ballot studies, in the presence of strong theory, can sometimes be interpreted from the perspective of which measure is best. We need other, better measures of when questions have been "improved".
    c. These measures are needed not only to help demonstrate the value of question evaluation but also in order to do studies of different approaches to the kinds of testing that is done (for example, different ways of conducting cognitive interviews). Critical evaluations of alternative ways of testing are needed in order

to begin to develop some standards for what constitutes adequate question evaluation. How can one tell good, effective cognitive testing from inadequate cognitive testing of questions?

d. A related topic is how best to analyze results of question testing. Analytic protocols appear to be very idiosyncratic to the testing organizations, and there are no real guidelines about how to best analyze results. Studies of alternative approaches are needed.
e. Standards for question testing also need to reflect the fact that testing can be done at different stages of the process. The best protocols for developing questions for a brand new survey instrument may be different from those for testing questions being used in an ongoing survey.
f. Finally, there is an ongoing need for better understanding, and better protocols, for testing survey instruments that will be used in more than one mode of data collection: mail and telephone; mail and internet; and in-person interview.
g. And, there still is an ongoing need to understand how best to test questions to produce comparable data across languages.

LIST OF PARTICIPANTS

	Name	Organisation	E-mail address
1	Allard, Benoit	Statistics Canada	benoit.allard@statcan.ca
2	Beatty, Paul	National Center for Health Statistics	pbb5@cdc.gov
3	Blanke, Karen	Federal Statistical Office Germany	karen.blanke@destatis.de
4	Collins, Debbie	National Centre for Social Research	d.collins@natcen.ac.uk
5	Cosenza, Carol	Center for Survey Research, University of Massachusetts	Carol.Cosenza@umb.edu
6	Dale, Trine	Statistics Norway	trine.dale@ssb.no
7	Davidsson, Gunilla	Statistics Sweden	gunilla.davidsson@scb.se
8	Esposito, Jim	Bureau of Labor Statistics	Esposito.Jim@bls.gov
9	Fowler, Jack	Center for Survey Research, University of Massachusetts	Floyd.Fowler@umb.edu
10	Giesen, Deirdre	Statistics Netherlands	igin@cbs.nl
11	Godenhjelm, Petri	Statistics Finland	petri.godenhjelm@stat.fi
12	Henningsson, Birgit	Statistics Sweden	birgit.henningsson@scb.se
13	Hole, Bente	Statistics Norway	bente.hole@ssb.no
14	Kelly, Paul	Statistics Canada	paul.kelly@statcan.ca
15	Lawrence, Dave	Statistics Canada	dave.lawrence@statcan.ca
16	Levesque, Marcel	Statistics Canada	marcel.levesque@statcan.ca
17	McGee, Alice	National Centre for Social Research	a.mcgee@natcen.ac.uk
18	Notnaes, Tore	Statistics Norway	Tore.notnaes@ssb.no
19	Pascale, Joanne	U.S. Census Bureau	Joanne.pascale@census.gov
20	Potaka, Lyn	Office of National Statistics	Lyn.Potaka@ons.gsi.gov.uk
21	Vis, Rachel	Statistics Netherlands	rvcs@cbs.nl
22	Willimack, Diane	U.S. Census Bureau	diane.k.willimack@census.gov