

Test report from Statistics Norway's cognitive and usability testing of questions and questionnaires

Appendix B of WP4 Deliverable 3 of the MIMOD project

Dag F. Gravem

Nina Berg

Frode Berglund

Karianne Lund

Katharina Roßbach

January 2018

Contents

1	Introduction.....	3
2	Testing mixed-mode questionnaires.....	3
3	WP4 test design and objectives	4
3.1	Test Methods used.....	5
4	Unimode tests – the ICT survey.....	6
4.1	The CATI test questionnaire	6
4.2	The CAWI test questionnaire.....	6
4.3	CATI test findings.....	7
4.3.1	Behaviour coding.....	7
4.3.2	Retrospective cognitive interviews	10
4.4	CAWI test findings	11
4.4.1	Usability issues (see also WP5 deliverable 3).....	11
4.4.2	Use of “Don’t know” and “Do not wish to answer” buttons.....	11
4.4.3	Other findings from the retrospective cognitive interviews	12
4.5	Summary and discussion of CATI and CAWI unimode test findings.....	12
5	Mode specific tests – the “omnibus”	13
5.1	Questions/variables selected and tested for the mode specific “omnibus”.....	13
5.1.1	Labour Force Survey - HWACTUAL.....	13
5.1.2	Adult Education Survey – NFEREASON.....	16
5.1.3	Adult Education Survey - NFENBHOURS.....	17
5.1.4	EU-SILC – questions related to dwelling and dwelling costs HH021, HH060 and HH070 18	
5.1.5	EU-SILC – HH071 mortgage principal repayment and related questions	20
5.1.6	EHIS – AL1 on alcohol consumption	22
5.1.7	ICT survey – A1 and A2 on types of Internet connections	23
5.1.8	ICT survey – B3 and B5 on types of units used for Internet activities, and types of Internet activities	24
5.1.9	ICT survey – D2 on goods and services bought or ordered over the over the Internet for private use last 12 months	25
6	Discussion of unimode and mode specific designs tested	27
7	Discussion of test design and further research.....	27

1 Introduction

The main task of WP4 has been to offer recommendations on question and questionnaire design for mixed mode surveys, with an emphasis on mixed mode designs that include web. A discussion of whether to use mode specific or unimode design approaches for questions is central.¹

From WP4 deliverables 1 and 2, we know that all main modes are used for all the key ESS social surveys. From the reviews of the model questionnaires, technical specifications and other background materials in deliverable 3, we also know that one specific mode or type of mode (usually visual) often is presupposed or recommended by Eurostat.

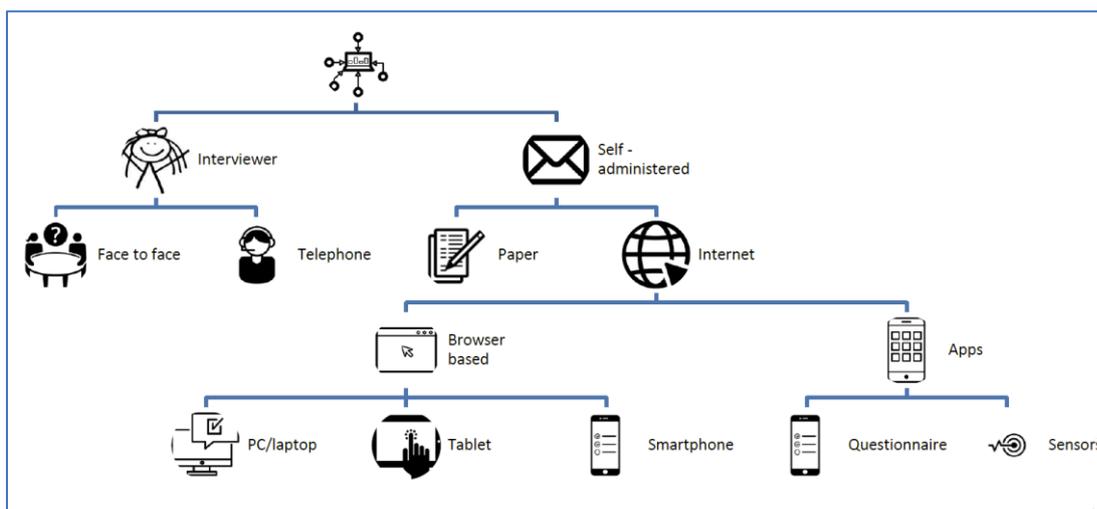
When applying Campanelli et al.'s classification criteria described in Deliverable 3, we also know that all the ESS questionnaires contain question types that can be more vulnerable to measurement differences between modes. Often, the mode recommended by Eurostat for a questionnaire will not be recommended for a specific question according to the classification criteria.

Expert appraisals using Campanelli's or other criteria can only *indicate* possible problems. There is a need for qualitatively and quantitatively testing these questions and questionnaires to gain more insight into how they perform in practice in different modes. Developing and testing specific questions could also shed light on the fruitfulness of unimode and mode specific approaches. How far can the unimode approach be pushed? In what cases, if any, can using the strengths of the different modes counteract measurement differences using a mode specific approach? These two questions formed the basis for the design of Statistics Norway's WP4 cognitive and usability tests.

2 Testing mixed-mode questionnaires

New modes, devices, data sources and mixes of these makes the task of question testing more complex, as noted and illustrated by figure 1 from Deirdre Giesen's presentation at the 2018 Quest workshop. Ideally, all modes should be tested for possible mode effects, but the scarcity of resources usually impedes this. (Giesen 2018.) The literature is still scarce on procedures and recommendations for the testing of mixed mode questionnaires, and Giesen's presentation will in the following serve as a benchmark for the description of Statistics Norway's WP4 tests.

Figure 1. "New modes and mixes". Illustration from Deirdre Giesen's presentation "Testing for mixed-mode and mixed-device data collection" at the 2018 Quest workshop in Wiesbaden



¹ Unimode = make questionnaires as similar as possible across modes. Mode-specific = adapt specifically to each mode. For further discussion, see WP4 deliverable 3.

The presentation outlines a five-step mixed mode testing strategy that takes resource constraints into account:

Step 1: Make a mode risk assessment, given questionnaire content, design and population: is there a danger of measurement errors in different modes?

Step 2: Decide if a focus on comparing modes is needed. If yes, assign comparable groups to different modes and test in a similar way by randomising and matching. Also consider whether additional types of test respondents are needed.

Step 3: Decide which modes to test, and when in the development process, outlining three approaches

- a) Testing “Worst case”/highest risk modes only, e.g. CATI and smartphone
- b) Testing iteratively with easily available modes like paper and interviewer as a first guide to programming for other modes
- c) Testing the highest risk modes, and do a desk review of other modes

Step 4: Test the relevant modes as realistically as possible, using modes and devices respondents are familiar with and likely to use in a fielded survey.

Step 5: In analyses and reporting, distinguish between usability and content findings, and reflect on mode-specificity regarding findings.

3 WP4 test design and objectives

Statistics Norway’s WP4 tests was divided into two main parts;

- 1) Testing a unimode approach for one single survey. Here, the entire Norwegian version of ICT survey questionnaire was tested in CATI, CAWI-PC and CAWI mobile versions. (The CAWI mobile tests were also used for WP5 analysis.)
- 2) Testing mode specific approaches for selected questions in a specially designed “omnibus” test questionnaires.² In the mode specific part, selected questions and question types from the LFS, AES, EU-SILC, EHIS and ICT surveys were tested in CATI and CAWI-PC versions.

The mode risk assessment in step 1 was carried out using the Campanelli criteria mentioned above. Rather than assigning comparable groups to the different modes, we used a test-retest design for both the unimode and the mode specific parts.

In the test-retest design, the test persons completed a CATI mode test via telephone first. About one week later they completed a CAWI re-test responding to the same questions at Statistics Norway’s cognitive lab using PC or their own mobile device.

The participants were not told in advance that they would be asked to respond to the same questions. Compared to using comparable groups, this design did result in some cases of memory effects influencing the results. On the other hand, it allowed us to ask the test persons to compare and evaluate how questions performed in each mode, and to compare their answers. Additionally, we had the chance of asking them which of the modes, if any, they preferred.

² An omnibus survey contains questions on a wide variety of subjects, in this case from multiple surveys

With reference to step 4 in Giesen’s test strategy, we recruited test persons between the ages of 25 and 35 under the assumption that this is a demographic group which is comfortable with using smartphone and web for a variety of purposes. We did this to obtain more robust results and avoid variations due to differences in familiarity with the technology. To further ensure this, the CAWI mobile tests were all done on the respondents’ own devices. The drawback of our approach is that the results of our tests cannot necessarily be assumed to be representative of the general population. However, many problems identified for test persons in this group will likely also be problematic, and more so, for population groups less familiar with modern communication technology.

The test persons were chiefly recruited through invitations posted on Statistics Norway’s Facebook page. Additionally, six of the test persons were recruited internally among new employees at Statistics Norway, and only responding in CAWI mobile mode. With one exception, the internal recruits had little to no experience with survey questionnaires in their daily jobs.

Table 1 summarizes the tests and questionnaire versions used, with a total of 28 tests and 18 test persons. For two of the CAWI follow-up tests (test persons 5 and 6), we were unable to schedule appointments within the test period but concluded that we have sufficient information from the completed tests to perform the analysis. Thus, there were 16 test-retests, two CATI only ICT tests, and six CAWI mobile only ICT tests.

Table 1. Summary of tests and questionnaire versions used

Test person	CATI test	CAWI (re-)test
1	ICT	ICT PC
2	ICT	ICT PC
3	ICT	ICT mobile
4	ICT	ICT mobile
5	ICT	Not conducted
6	ICT	Not conducted
7	Not planned	ICT mobile
8	Not planned	ICT mobile
9	Not planned	ICT mobile
10	Not planned	ICT mobile
11	Not planned	ICT mobile
12	Not planned	ICT mobile
13	“Ominbus”	“Ominbus” PC
14	“Ominbus”	“Ominbus” PC
15	“Ominbus”	“Ominbus” PC
16	“Ominbus”	“Ominbus” PC
17	“Ominbus”	“Ominbus” PC
18	“Ominbus”	“Ominbus” PC

3.1 Test Methods used

In the unimode CATI tests, we used behaviour coding and conducted a short retrospective interview focussing on pre-selected questions. In the unimode CAWI (re-)tests, eye tracking equipment was used, and new retrospective interview focussing on usability and user experience was conducted.

In the mode specific (omnibus) CATI tests, behaviour coding was also used, but no retrospective interview was done. In the CAWI re-tests, eye tracking was used, and a longer cognitive interview conducted focussing on usability as well as going through all the questions in some detail, with reference also to CATI responses.

4 Unimode tests – the ICT survey

Prior to the tests, we reviewed the ICT survey model questionnaire provided by Eurostat and identified several questions likely to have mixed-mode challenges, according to the Campanelli typology. The main factors possibly influencing quality were the following:

- A. Mark all that apply format
- B. Inherent difficulty due to concepts
- C. Use of instructions
- D. Use of Don't know
- E. Question length (Dillman)

4.1 The CATI test questionnaire

The Norwegian ICT survey is currently CATI only and embedded in the 2nd quarter of a social survey omnibus. A slightly adapted version of the ICT part of this omnibus was used in the CATI tests. In terms of structure and wording, it is very close to the model questionnaire. The verbosity and numerous technical terms in both questions and instructions are also found in the Norwegian CATI questionnaire. The many “mark all that apply” questions, however, have been turned into batteries of yes/no questions. This adaption has some advantages in terms of the attention given to each element, but several of these questions have very many elements, which may lead to satisficing, a lack of overview and possible breakoff.

4.2 The CAWI test questionnaire

The CATI version was in its turn used as the basis for the MIMOD CAWI test questionnaire. Some of the wording was changed to adapt to self-completion, e.g. referring to “you” instead of “the respondent”. Introductory texts were included in a bold, large font, whereas the questions were in normal size bold font. Interviewers’ clarifying information was included on-screen in non-bold. “Don’t know” and “Do not want to answer” options with greyed-out labels were included for all questions. (See figure 2.)

Figure 2. Layout of mobile adapted and PC versions of ICT questionnaire

Statistisk sentralbyrå
Statistics Norway

Vi starter med noen spørsmål om tilgang til og bruk av informasjons- og kommunikasjonsteknologi. De første spørsmålene gjelder ting husholdningen har hjemme. Ting man har på hytte, i feriebolig eller lignende skal ikke regnes med.

Har du tilgang til Internett hjemme, uansett type utstyr?
Hvis husholdningen har tilgang, men ikke bruker den svarer du 'Ja'. Hvis husholdningen kan få tilgang, men ikke abonnerer på tjenesten svarer du 'Nei'.

Ja
Nei
Vet ikke
Ønsker ikke å svare

Så kommer noen spørsmål om hvilke typer internettforbindelser som brukes hjemme. Først kommer to spørsmål om bredbåndstilknytning.

Bruker husholdningen en fast bredbåndsförbindelse, for eksempel via ADSL, kabel-TV, fiberkabel eller satellitt?
Bredbånd er en betegnelse brukt om signaloverføring av en viss minimumshastighet (vanligvis raskere enn ISDN).

Ja
 Nei
 Vet ikke
 Ønsker ikke å svare

Bruker husholdningen mobil bredbånd med minst 3G over mobiltelefonnettverket hjemme Enten via mobiltelefon, smarttelefon, bærbar PC eller nettbrett og lignende?
Mobil bredbånd omfatter bare mobilnettverk med tredje generasjons hastighet, kalt 3G eller bedre. Det gjelder UMTS (3G), HSDPA eller Turbo 3G (3G+) eller LTE (4G). Mobiltelefon eller bærbare PC-er og nettbrett som bare brukes utenfor hjemmet eller via trådløse ruter hjemme skal ikke regnes med.

Ja
 Nei
 Vet ikke
 Ønsker ikke å svare

For converted “check all that apply” questions that were turned into batteries of questions, different approaches were tried out. For some batteries, the entire question stem and ending were in bold fonts, in other batteries, the question stem was in non-bold.

The layout was different in PC and mobile format, though both had paging navigation. In the PC version, two or three questions were visible per page, but only one question was visible per page in the mobile version.

On PC, radio buttons and checkboxes were used for response options, and users had to use forward/backward navigation buttons to move between pages. In the mobile version, finger-friendly vertical buttons were used. When picked, a thicker black frame was added to the response option.

In the CAWI mobile version, responding to a question with only one response option took you automatically to the next question. On checkbox format questions, of which there was only a handful, the navigation buttons had to be used. Visually, there was no way of telling whether it was possible to pick more than one option, however. If the test person chose “Don’t know” or “Do not wish to answer”, the navigation buttons had to be used regardless of whether one or more than one substantial response was allowed.

4.3 CATI test findings

4.3.1 Behaviour coding

The CATI interviews were behaviour coded by two independent coders according to a scheme with four dimensions: Answer quality, Interruptions, Clarification and Interviewer behaviour. The dimensions and their respective categories are presented up in table 2.

Table 2. Behaviour coding scheme

Dimension	Categories
Answer	Adequate/codable
	Inadequate/incodable
	Qualified/“I think ...”
	Refusal
	Don’t know
Interruption	Question interrupted by respondent before fully read
Clarification	Interviewer initiated
	Respondent initiated
Interviewer behaviour	Small changes made to the question
	Big changes made to the question

A reconciliation of the behaviour coding showed that 59 of the 120 questions for at least one of the six test persons were coded as not as not adequately answered and/or in one or more of the other categories presented in table 2. Of these 59 questions, there were eight where some problem was reported in half the tests or more: A2b, A2c, A2d, B7, D1, D3, E1c and E1d. These questions will be discussed in the following paragraphs.

4.3.1.1 A2b, A2c and A2d

Bruker husholdningen mobilt bredbånd med minst 3G over mobiltelefonnettet hjemme Enten via mobiltelefon, smarttelefon, bærbar PC eller nettbrett og lignende?
Mobilt bredbånd omfatter bare mobilt nettverk med tredje generasjons hastighet, kalt 3G eller bedre. Det gjelder UMTS (3G), HSDPA eller Turbo 3G (3G+) eller LTE (4G). Mobiltelefon eller bærbare PC-er og nettbrett som bare brukes utenfor hjemmet eller via trådløse ruter hjemme skal ikke regnes med.

Ja
Nei
Vet ikke
Ønsker ikke å svare

Figure 3. ICT survey A2b

“Does the household use a mobile broadband connection with at least 3G over the mobile phone network at home? Yes/No”

A2 b, c and d are the three last of the answer options from question A2 on different Internet connections. The behaviour coding confirms assumptions from the expert evaluation that the many technical terms and complex sentence structure makes comprehension difficult. Also, many respondents will not have the technical insight to know what connection they do have. The narrowband connection questions A2c and A2d stand out with several “Don’t know” responses.

4.3.1.2 B7

Nå kommer noen spørsmål om bruk av internett og delingsøkonomi.
Har du i løpet av de siste 12 månedene brukt nettsteder eller apper til å ordne innkvartering, for eksempel en leilighet, et rom eller et feriehus, fra en annen privatperson?
Flere svar er mulig.

Ja, nettsteder eller apper som er spesielt laget for at privatpersoner skal kunne bestille og leie ut innkvartering, som for eksempel AIRBNB
Ja, andre nettsteder eller apper, som Facebook
Nei
Vet ikke
Ønsker ikke å svare

Figure 4. ICT survey B7

“And now for some questions about use of Internet and the sharing economy. Have you, during the last 12 months, used websites or apps to arrange an accommodation, e.g. an apartment, a room or a holiday house from another private individual?”

- a) Yes, websites or apps especially designed for private persons to order and offer accommodation, like AirBNB
- b) Yes, other websites or apps, like Facebook
- c) No”

This is a good example of a question originally written for a visual mode not being sufficiently adapted to an aural mode. A simple “yes” to this question requires the interviewer to probe whether this was AirBNB or similar, and/or other websites or apps. After a few interviews, the CATI interviewer started to add the response options to compensate for this, making the question longer but more comprehensible.

4.3.1.3 D1

“Now for some questions on your private use of the Internet for online purchases. All online purchases are included, regardless of the device used, such as desktop or portable PC, handheld units such as tablets and mobiles/smartphones. Does not apply to manually typed e-mails, SMS or MMS.

When did you last use the Internet for buying or ordering goods or services for private use?

- a) Within the last 3 months
- b) Between 3 months and a year ago
- c) More than a year ago
- d) Have never bought or ordered goods or services over the Internet”

This was designed as an open question in CATI format with no mention of reference periods, making it unnecessarily difficult for some respondents. As a result, the interviewer started to read the response options as well, like for B7. On one occasion, the interviewer omitting the last option, taking for granted that the test person at some time had ordered something for private use.

The long introduction makes the cognitive task of remembering all the different conditions hard, but there is no evidence of an impact here. In the retrospective interviews, however, the relevance of the information in the introduction was called into question by one test person in the cognitive, as it appeared redundant for the interpretation of the question.

4.3.1.4 D3

“You have answered that you have bought or ordered goods or services over the Internet for private use during the last 12 months. From whom did you buy or order this? Was it ...

- a) **Norwegian sellers?** Yes/no
- b) **Sellers from the EU countries?** Yes/no
- c) **Sellers from the rest of the world?** Yes/no
- d) **Sellers whose country of origin is unknown?** Yes/no”

Here, the test persons were unsure of which country the sellers were from. One test person mentioned IKEA, which is established in Norway, but originally a Swedish company with the international headquarters currently in the Netherlands. Additionally, recall is also an issue here, as it can be difficult to remember every purchase, and hence every seller, from the last 12 months.

4.3.1.5 E1c and E1d

“[Which of the following identification procedures for online services for private use have you used during the last 12 months? Online services can be Internet banking, public services online, or buying or selling goods online]

(...)

- c) **Have you used a security token, such as security token BankID? Do not include mobile phone BankID.** Yes/No
- d) **Have you used an electronic identification certificate stored on your own machine, or card used with a card reader?** E.g. Buypass with a smartcard and card reader. Yes/No”

Here, difficult terms and obscure and/or outdated processes were the main causes of the experienced problems. The introduction is long and complicated and has been translated very closely to the technical source text.

4.3.2 Retrospective cognitive interviews

After each CATI interview, a short cognitive interview was conducted based on a semi-structured interview guide. The moderator first asked about the test persons' overall impression of the survey, before going over to probing pre-selected questions expected to be challenging for the respondents. Other questions where problems were detected during the initial interview were also probed. How many questions each test person received during the cognitive interview depended on time constraints and the reactions and answers given by the test persons during answering the survey. On average, the test lasted about 30 minutes with 15-20 minutes for the CATI interview and 10-15 minutes for the cognitive interview.

4.3.2.1 Overall impressions

TP3 said the total impression of the survey was very good. TP2, TP5 and TP6 said it was ok, while the remaining two said some of the questions were easy and some questions were difficult. TP1 stated that especially the technical questions were challenging since some terms were not known to the test person. TP5 stated that the interviewer did a lot of talking because of the long sentences in the questions. TP6 stated that questions which contained "other" were more challenging because they were not very precise and (s)he suggested to have examples.

4.3.2.2 Introductions and clarifying information

5 out of 6 test persons received a general question of "Do the long texts (in introductions, questions and instructions) make it easier or more difficult to understand the question and give an answer?" TP3, who was very satisfied with the survey, did not receive this question. TP1 stated that the length was ok. TP2 suggested to split long questions and admitted that with long questions, (s)he lost focus because of the large amount of information. TP4 stated that the lengthy texts sometimes made it easier and sometimes more difficult to understand the question. TP5 stated that the long sentences required long, continuous listening from the test person and time to process the information. The same test person stated that if the question is short and the introduction is long, it is still easily understandable. Difficulties arose for TP5 when the question itself was very long. TP6 stated the questions are very long and suggested to name the reference period just once and not in every single question in order to shorten the long texts.

4.3.2.3 Questions selected for further probing based on expert review

These questions that had been selected for further probing were A1, A2a, A2b, B5b, B5c, B7, and E1a-g. Of these, A2b, B7, Ec and Ed were also among the ones that had been identified as especially problematic in the behaviour coding, to a large extent confirming the expert appraisal and what has been discussed above. A summary of retrospective interview findings for each of these questions is found in Appendix A.

4.3.2.4 Other findings

Towards the end of the cognitive interview, the moderator asked the test persons if it had been easy or difficult to differentiate between internet use for private purposes versus job-related purposes. Especially since the questions referred to different units and covered different types of internet use. 5 out of 6 test persons were asked to comment on this. TP1, TP2 and TP4 stated that it was easy to differentiate between the two different types. TP5 stated that it was ok to differentiate between job and private usage of Internet, but that the job-related questions were more difficult than the ones about private usage. TP6 said it was ok for her/him to differentiate between job and private usage but that (s)he could imagine that other test persons might find it more difficult to differentiate between the two.

Finally, the moderator asked the test persons if they have any suggestions improvement other than those which are already named. TP1 stated that for many questions (s)he could have answered faster but (s)he did not want to interrupt the interviewer. TP2, TP3, TP4 and TP6 had no additional comments. TP5 emphasized that the formulations should be improved, especially that one does not include words which no one uses.

4.3.2.5 *Summary of retrospective cognitive interviews*

To sum it up, difficulties of understanding terms were often related to terms that are used by experts, but not the average person (i.e. IT terms), and words that are old-fashioned in Norwegian language. Generally, the test persons found that the ICT survey consisted of many long sentences which made it more difficult during the interview. First, they needed to wait until the interviewer had finished reading the questions. Second, the survey questions, introductions and instructions contained a lot of information, that was often repetitious and redundant, which meant more time was needed to process the question.

4.4 CAWI test findings

The four CAWI re-tests were conducted about a week after the telephone interview, two of which in the PC CAWI version, and two using mobile CAWI. In addition, 6 CAWI mobile tests also intended for WP5 analysis were conducted. All these ten tests are included in the below analysis.

4.4.1 Usability issues (see also WP5 deliverable 3)

The overall usability was satisfactory, although mobile version test persons experienced some problems. Test persons with smaller iPhone models had to scroll vertically on the most verbose questions, to see all the response categories, but this did not lead to any measurement errors. However, in one case, a respondent did not see the “next” button and was not able to navigate to the next page, needing help from the moderator. The selection buttons could have been a bit wider and more finger-friendly.

The loading time between pages varied between tests. This led to some navigation confusion in the mobile version, as some users tried to click on the “next” button before the automatic loading of the next page. In one case, a test person was unsure of whether this had led to him/her skipping a question.

One of the test persons remarked that she/he would have liked a clearer indication of what had been answered before the automatic loading of the next page, to be sure that the correct response option had been chosen.

The fact that it was not possible to distinguish the few questions with more than one response option from the one option questions in the mobile format led to some initial navigation confusion for three of the eight mobile version test persons. They picked one of the options, expecting to continue automatically to the next page. When this did not happen, they all believed they had not actually picked an option: the black frame around the option was not enough to make this clear. After having unintentionally deselecting the response, they re-selected the response category and clicked on the “Next” button to proceed.

4.4.2 Use of “Don’t know” and “Do not wish to answer” buttons

The “Do not wish to answer” button was not used by any of the test persons. The “Don’t know” button was however used on several questions, notably of two types: questions with different technical terms, and the final element of question batteries asking about “other” uses of/purchases through the Internet.

Figure 5. Question with “Don’t know” selected

On question A2 on types of Internet connections used at home, three of ten test persons used “Don’t know” for one or both narrowband connection types (A2c, A2d), due to unfamiliar and irrelevant terms. In the telephone interviews, three of six test persons responded “Don’t know” to the same questions, even though it was not offered as an explicit response option.

On question D2 on goods and services bought or ordered over the internet, five of ten test persons answered “Don’t know” to D2o on “other goods or services”. Being the last of 15 different categories of goods and services, the response task of both remembering all the previous categories, among them several difficult and unintuitive ones, and then assessing whether anything else had been procured, is very complicated.

In the retrospective interviews, two of the test persons commented that she/he interpreted the greying out as indicating that we would prefer that respondents did not use them. Two others were unsure of whether it was actually possible to select them, although one of them actually did. Another said that “after a while, I just noticed the [non-grey] response options, and kind of forgot about the two last ones.”

4.4.3 Other findings from the retrospective cognitive interviews

In general, the questions that were perceived as most problematic in the CAWI interview were the same that were problematic in the CATI interview: types of Internet connections, purchases of goods and services online.

The overall verbosity and repetitions were an issue, and several test persons commented that in the question batteries they “looked for what had changed”, regardless of whether question introductions were bold or non-bold.

Another test person commented that the transitional introduction from one topic to another were awkward, as “there really wasn’t a change of topic” – it all had to do with ICT.

4.5 Summary and discussion of CATI and CAWI unimode test findings

The main problems with the ICT survey questionnaire that was used for the MIMOD tests did not have to do with mode specific issues: they were rather related to difficult, unfamiliar and technical terms, and the length and verbosity of the questions. The test results indicate that including explicit greyed-out “Don’t know” response options is justified for the questions with difficult and technical words – they did not seem to result in casual “Don’t know” responses, and the questions seemed to perform similarly in CATI and CAWI. However, this should be further tested on other groups of respondents quantitatively as well as qualitatively.

One factor that significantly contributed to the length and verbosity of the test questionnaire, is the transformation of checkbox questions in the model questionnaire to series of yes/no questions. On the positive side, this transformation makes a CATI/CAWI unimode approach possible for the questions, and it also make the questions better adapted to mobile CAWI (See WP5’s deliverable 3

Har du kjøpt eller bestilt andre varer eller tjenester over Internett de siste 12 månedene?

Ja

Nei

Vet ikke

Ønsker ikke å svare

for further discussions on this). On the negative side, the transformation does not allow for using one of the strengths of large screen visual modes: a presentation of all response options. For this reason, some of these questions from the ICT survey were selected for the mode-specific testing (see below).

All in all, both Eurostat’s model questionnaire and Statistics Norway’s implementation of it suffer from insufficient operationalization of theoretical variables. In the case of the model questionnaire, this is perhaps justifiable, provided that NSIs are allowed enough freedom to do proper operationalizations and user-friendly questionnaires. We recommend that Statistics Norway take the findings from the MIMOD tests into consideration in future revisions of the ICT survey questionnaire to better achieve this goal.

5 Mode specific tests – the “omnibus”

For the mode specific tests, we selected questions that could be problematic in CATI and/or CAWI mode, but where the strengths of one or both modes also could be used to aid the respondent. This aid could take the form of either clarification, by providing context, structuring information retrieval or aiding in information processing. In the case of CAWI, these possible help functions involved space-demanding layout. Therefore, the tests were only done in CAWI PC format. The resulting limitation of our mode specific approach for web questionnaires will be discussed below. The different questions were assembled in an “omnibus” survey questionnaire, made with Axure prototyping software.

5.1 Questions/variables selected and tested for the mode specific “omnibus”

The questions for the “omnibus” are all taken from Statistics Norway’s social surveys, which apart from the CAWI AES are all single-mode CATI. We consider them either as key or typical questions for each survey, and some are generalizable to other surveys or topics. The questions/variable selected are presented in table 3, with references to their names in model questionnaires or survey documentation.

Table 3. overview of questions/variables selected for mode-specific tests

Survey	Question/variable	Description
Labour Force Survey	HWACTUAL	Hours actually worked in reference week
Adult Education Survey	NFEREASON	Reasons for participating in learning activity
Adult Education Survey	NFENBHOURS	Hours spent participating in learning activity
EU-SILC	HH021, HH060, HH070	Questions on dwelling and dwelling costs
EU-SILC	HH071	Mortgage principal repayment
EHIS	AL1	Alcohol consumption
ICT survey	A1, A2	Types of Internet connections
ICT survey	B3, B5	Internet activities
ICT survey	D2	Internet purchases for private use

As the main focus was on testing specific questions, we will present the results question by question and not mode by mode like we did for the unimode tests above, with an emphasis on the CAWI tests. Behaviour coding was used in the CATI tests, but qualitatively rather than quantitatively, as a limited set of questions was pre-selected.

5.1.1 Labour Force Survey - HWACTUAL

The HWACTUAL variable is the number of hours the respondent has worked during the reference week, and a key question in the LFS. We tried out a day-by-day calculator for hours actually worked

during the reference week. The idea is to use the strengths of web mode in visual presentation and calculation for increased precision.

As described in WP4 deliverable 3, Statistics Finland have previous quantitative testing experiences with a mode-specific design with a single question in CATI mode, and an hour calculator in CAWI mode. On the other hand, the ONS in the UK have had negative experiences with the same CAWI design in cognitive tests and has decided not to include it in their revised LFS questionnaire. Because of these different results, we decided to try to test it in a Norwegian context as well.

The CATI version was a single question: “How many hours did you work during last week?” The CAWI version is shown in figure x. Clarifying information in non-bold reads: “Include any paid and unpaid hours worked, and flexitime. Answer in hours and minutes per day.” This differs from the regular Norwegian CATI question, which is a sequence asking detailed questions about absence and extra hours worked, but is similar to the Finnish and UK questions.

Figure 6. LFS HWACTUAL Day-by-day calculator for determining hours worked last week.

Hvor mange timer arbeidet du i forrige uke?		
Ta med eventuell betalt og ubetalt arbeidstid, samt flekstid		
Oppgi svaret i timer og minutter per dag		
	Timer	Minutter
Mandag	<input type="text" value="7"/>	<input type="text" value="30"/>
Tirsdag	<input type="text" value="8"/>	<input type="text" value="00"/>
Onsdag	<input type="text"/>	<input type="text"/>
Torsdag	<input type="text"/>	<input type="text"/>
Fredag	<input type="text"/>	<input type="text"/>
Lørdag	<input type="text"/>	<input type="text"/>
Søndag	<input type="text"/>	<input type="text"/>
Totalt	<input type="text" value="15"/>	<input type="text" value="30"/>

5.1.1.1 Findings

Of the six test persons who participated in the mode specific tests, four had regular paid work, one of whom was on maternity leave. In the CATI interview, two of the test persons quickly answered an adequate “37 and a half hour” (which is standard tariff full time), one answered a qualified “40, I think”, and the person on maternity leave answered an adequate “0”. The quick answer of the first two respondents could be an indication of satisficing, reporting contracted rather than actual work hours.

In the CAWI retest, the two test persons who had responded “37 and a half hour” in CATI both started filling in 7 hours and 30 minutes for each day of the week. One of them did so for Saturday and Sunday as well, discovered in the sum field that the total was 52 hours and 30 minutes, and removed the figures for Saturday and Sunday, ending up with 37 hours and 30 minutes. During the retrospective interview, this test person said “Do you want to know how much I am paid for, or how much I actually work? I interpret it as how much I am *paid* for”. Thus, unpaid extra hours were left unreported.

The other test person who had reported 37 hours and 30 minutes in the CATI interview filled in 7 hours and 30 minutes for Monday through Friday, before changing the hours for one of the days to 8

hours and 30 minutes. In the retrospective interview, she too said that she only included hours she got paid for, and had included one hour of overtime. She did not include flexi hours.

The test person who had reported “40, I think” in the CATI interview followed another strategy. She rounded up or down to the nearest hour for each day, disregarding the minutes boxes (figure 7). The hours added up to 41 hours. In the retrospective interview, she said that she noticed the sum field but did not pay much attention to it. This was confirmed by the eye tracking recording. In the retrospective interview, she said that she would have estimated it to 40 hours if the question had been identical to the CATI version.

Figure 7. Test person rounding off hours worked to the nearest hour, without entering minutes.

The screenshot shows a survey form with the following content:

Hvor mange timer arbeidet du i forrige uke?
Ta med eventuell betalt og ubetalt arbeidstid, samt fleksitid
Oppgi svaret i timer og minutter per dag

	Timer	Minutter
Mandag	<input type="text" value="8"/>	<input type="text"/>
Tirsdag	<input type="text" value="10"/>	<input type="text"/>
Onsdag	<input type="text" value="8"/>	<input type="text"/>
Torsdag	<input type="text"/>	<input type="text"/>
Fredag	<input type="text"/>	<input type="text"/>
Lørdag	<input type="text"/>	<input type="text"/>
Søndag	<input type="text"/>	<input type="text"/>
Totalt	<input type="text" value="26"/>	<input type="text"/>

NESTE

5.1.1.2 Evaluation

Both the CATI and the CAWI questions were asked without the context of contracted versus actual working hours found in the regular LFS. If this had been added to the test, it would possibly have made it clearer that both paid and unpaid work should be included. This information could also have been included in the question text, instead of just as clarifying information.

However, the risk of satisficing day-by-day instead of for the week remains. For workers with no variation in daily hours, the task is likely to be perceived as unnecessary and adding to the response burden. Based on test results and experiences from other countries' NSIs, this mode-specific solution cannot presently be recommended. Further testing, including testing of other approaches is advisable.

5.1.2 Adult Education Survey – NFEREASON

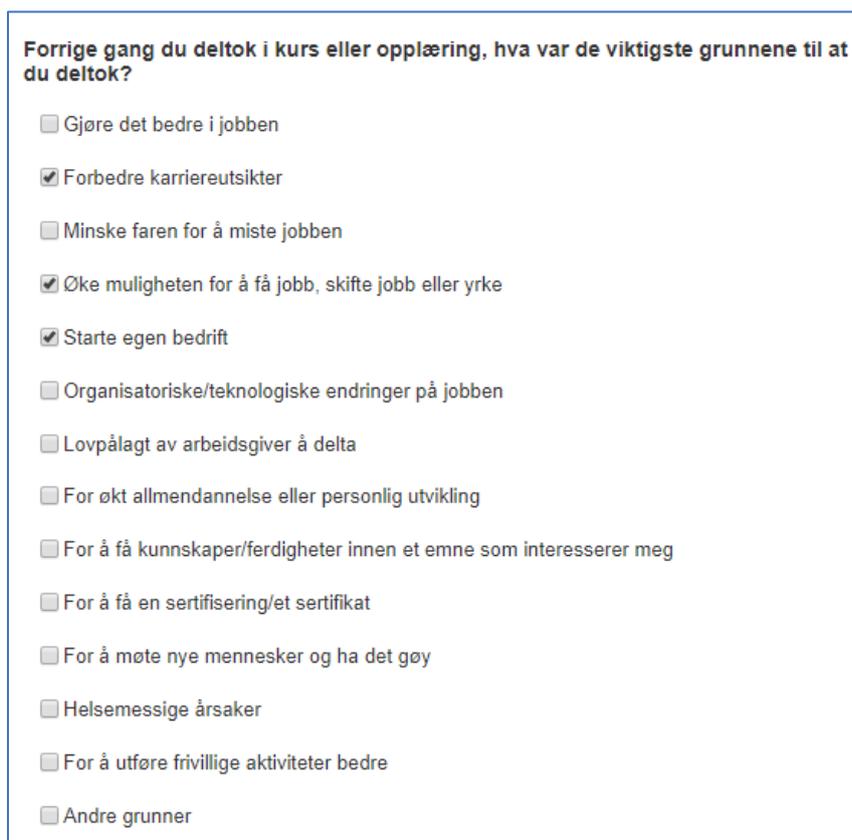
From the Adult Education Survey, we first selected the NFEREASON variable, about the most important reasons for participation in a non-formal learning activity. In the AES model questionnaire, it is intended as a mark/check all that apply question, presupposing a visual stimulus. As the question has 14 response options and includes a ranking task, reading out all the response options in CATI before asking the respondent to rank them appears unfeasible.

Statistics Norway has previously asked this as an open question with interviewer coding in CATI. With our test-retest design we wanted to investigate possible measurement differences between CATI and CAWI. Interviewer effects are a concern in CATI, and primacy effects a possible problem in CAWI mode.

In the AES, the question is asked about an activity randomly selected by the computer, from the last 12 months. To simplify, we instead referred to the last non-formal learning activity the test person participated in (which could also have been more than 12 months ago).

In the CATI, the open question was “Last time you participated in a course or training, what were the main reasons for your participation?” followed by interviewer coding. In the CAWI tests, the question was presented with checkbox response options as shown in figure 8.

Figure 8. AES NFEREASON check all that apply question



Forrige gang du deltok i kurs eller oppl ring, hva var de viktigste grunnene til at du deltok?

- Gj re det bedre i jobben
- Forbedre karriereutsikter
- Minske faren for   miste jobben
-  ke muligheten for   f  jobb, skifte jobb eller yrke
- Starte egen bedrift
- Organisatoriske/teknologiske endringer p  jobben
- Lovp lagt av arbeidsgiver   delta
- For  kt allmendannelse eller personlig utvikling
- For   f  kunnskaper/ferdigheter innen et emne som interesserer meg
- For   f  en sertifisering/et sertifikat
- For   m te nye mennesker og ha det g y
- Helsemessige  rsaker
- For   utf re frivillige aktiviteter bedre
- Andre grunner

5.1.2.1 Findings

In the tests, five of the six test persons reported *more* reasons for course participation in the CAWI version than in the CATI version. For three of them, there was a match between the interviewer’s coding and the test persons’ CAWI responses on at least one reason. Eye tracking data revealed that in CAWI, the test persons generally read or skimmed through the list of response options, selecting options either as they went along, after having read through the whole list, or as a combination.

One example of measurement differences is the newly employed test person who answered “to learn how to do the job” in CATI. This is not covered verbatim by the response options, and the interviewer chose to code it as “Other reasons”. When the test person responded to the question in the retest, she selected the first option “To do the job better”. In the retrospective interview, she commented that none of the options was exactly right, but that “To do the job better” was most similar. She further commented that “to learn the job” will be relevant for many newly employed respondents.

5.1.2.2 Evaluation

With available response options, the response process is very different from open questions, allowing for. The test situation can have contributed to the test persons investing more time and cognitive effort in the task than they would in a normal CAWI situation. Still, it is likely that the mode-specific approach tested will result in more and different responses.

To compensate for this, interviewer probing and branching could be introduced in the CATI version, e.g. “Were there job-related or personal reasons, or both, for your participation in this learning activity?” with follow-up questions for further specification of reasons. Such a solution could on the other hand be introduced in CAWI as well, and be better suited for mobile web, and remove the need for mode-specific question design.

5.1.3 Adult Education Survey - NFENBHOURS

The second question selected from the AES concerns how many hours of instruction the respondent received in connection with the last non-formal education. The model questionnaire contains a discussion of different ways of operationalizing this question; by asking per day, week or even month, leaving this up to each country. In the Norwegian AES, it has been possible to choose between number of hours, or number of days with a follow-up on the average number of hours per day. In the 2016 AES, 52% of the respondents chose to report hours, and 48% days, so this seems justified.

Due to a programming error, the question was not asked in the CATI tests. The sequence intended was taken from the CATI interview, closely following the model questionnaire: “We would now like to know how many hours of training you received. Is it easiest for you to report this in hours or days of training?” If days is selected: “How many days did you receive training?” and “On average, how many hours per day did you receive training?” The test person would then have been presented with a control question on the calculated figure: “We have calculated that you received a total of x hours of training. Does this seem right, or do you think a different total would be more correct?” If different total is chosen, the respondent would have been asked “How many hours of training would you estimate that you received?”

In the 2016 AES, 96% of the respondents who got the control question answered that the sum total seemed right, so although we did not get to test the CATI version qualitatively, there is quantitative evidence that seems to suggest that it works.

Like in HWACTUAL, the idea for the tests was to try to use the strengths of the visual, computer-assisted CAWI mode by calculating and presenting the total number of hours in context. The CAWI question put the elements from the CATI question sequence in a grid question to attempt using the strengths of the visual stimulus, and is shown in figure 9 – with the sum field filling the function of the control question.

Figure 9. AES NFENBHOURS with calculation days * average number of hours

Vi ønsker nå å vite hvor mange timer opplæring du fikk.
Hvordan er det lettest for deg å oppgi dette?

I antall timer

I antall dager

Hvor mange dager fikk du opplæring?

Hvor mange timer fikk du opplæring per dag i gjennomsnitt?

Totalt antall timer med opplæring

5.1.3.1 Findings

Three test persons preferred to report the number of hours, and two preferred to report the number of days. The last test person did not get the question, as she had not participated in any learning activity during the last 12 months.

In the retrospective interviews, the first test person who wanted to report hours had actually participated in training over several days. The courses were separate, but related, and she had to estimate and add hours from them. She entered 25 hours, before modifying it to 15. The second test person to report hours said that she had to think about how much of her workday was *not* part of the course, thus subtracting rather than adding, and ending up with five hours. The third hour-reporting test person gave a rough estimate of six hours.

The first test person to report days of training responded five days of training, and eight hours on average per day. In the retrospective interview, he said that eight hours probably was too much on average, and that he would have reported fewer hours if only asked to report hours. The second test person reported two days and five hours on average per day. In the cognitive interview, she expressed that “average hours” was a difficult concept, and that she was unsure of whether breaks should be subtracted.

5.1.3.2 Evaluation

As in the LFS HWACTUAL question, the hours calculator was intended as an aid in the cognitive process. The tests demonstrate, however, that people’s actual processes of recalling and processing information does not necessarily match the processes that can be foreseen in a questionnaire. Having to calculate average hours per day can be more difficult than adding and estimating hours per day. Adjusting days or hours on the basis of the sum field would involve division and subtraction and add further complexity to the cognitive process. With the results from the testing of HWACTUAL in mind, introducing a day-by-day calculator is not necessarily a good idea either.

5.1.4 EU-SILC – questions related to dwelling and dwelling costs HH021, HH060 and HH070

The variables/questions selected from the EU-SILC are economy-related factual questions about the respondent’s dwelling. Problems related to information recall, from memory as well as other sources such as invoices and bank transactions, and information processing are possible sources of measurement errors/differences.

We wanted to gain some insight into whether the visual CAWI mode would provide some useful context or overview, and whether the CATI interviewer could help in motivation or calculation efforts. Figure 11 shows the first questions of this sequence: A filter question on tenure status is

followed by questions on rent monthly, quarterly or yearly, and then by questions on what the rent covers and does not cover. This also serves as a filter question, followed by questions on (additional) costs for electricity and fuels for heating.

Figure 11. EU-SILC questions on tenure status and housing costs. Dwelling owner sequence.

Så kommer noen spørsmål om din bolig

Eier du eller noen i husholdningen boligen som selveier, gjennom borettslag eller boligaksjeselskap, eller leier eller disponerer du/dere boligen på annen måte?

Eier som selveier
 Eier gjennom borettslag/boligaksjeselskap
 Leier eller disponerer på annen måte

Betaler du/dere husleie eller fellesutgifter?

Ja
 Nei

Hvor mye betaler du/dere i husleie/fellesutgifter?

kroner per måned
 kvartal
 år

Omfatter husleien/fellesutgiftene ...

	Ja	Nei
... elektrisitet?	<input checked="" type="radio"/>	<input type="radio"/>
... annen oppvarming?	<input type="radio"/>	<input checked="" type="radio"/>
... varmtvann?	<input type="radio"/>	<input checked="" type="radio"/>

Hvor mye betaler du/dere for elektrisitet og fast eller flytende brensel?

kroner per måned
 kvartal
 år

5.1.4.1 Findings

The question on dwelling status (HH021 – owner-occupier, owning through housing cooperative, tenant or other) was perceived as easy to answer in both modes, although one test person commented that it was unnecessarily long and repetitious in CAWI. There were no measurement differences between CATI and CAWI.

On HH060, a yes/no question on whether the test persons paid rent (“husleie”) or joint costs (“fellesutgifter”) or not, there were some unexpected measurement differences and errors. One of three owners reported paying such costs in the CATI test, but *not* in the CAWI test. Two of three tenants reported *not* paying such costs in the CATI test, but reported paying them in the CAWI tests.

The cognitive interviews revealed that all these test persons in fact *did* pay rent or joint costs. The cause of the differences seems to be that the terms “husleie” and “fellesutgifter” were mutually unclear. Tenants generally pay “husleie”, and owners pay “fellesutgifter”. One of the tenants said that she only read “fellesutgifter” when completing the CAWI questionnaire, but did not notice “husleie” and hence answered no. The other tenant referred to the term “månedlig leie”, which is the same, but was not perceived as such by her. The owner seems to have missed the term “husleie” in the CATI question, a possible recency effect.

When it comes to the different questions for calculating housing costs for variable HH071, comparability is limited by the erroneous responses to HH060. For the remaining sums of costs, measurements were the same, or marginally different, with one exception: The test person was very unsure of joint costs and made different estimates in the CATI and CAWI tests respectively. This however was most likely a random error with imprecise measurement in both CATI and CAWI.

5.1.4.2 Evaluation

Whether the measurement errors on the question about rent/joint costs is related to mode and/or type of dwelling status is difficult to assess. More probably, the findings point to general problems with use of difficult and unfamiliar terms, and a lack of tailoring the questions to dwelling status. There is no evidence that the visual display gave a better overview of total housing costs in CAWI than in CATI mode, but this could potentially be developed further for future tests, e.g. a summary table showing the calculation of total housing costs from different questions.

5.1.5 EU-SILC – HH071 mortgage principal repayment and related questions

We also wanted to test another instance of a visual calculator aid for the EU-SILC questions on mortgage principal repayment. The rationale is that this could possibly compensate for recall/information retrieval issues likely to occur in interviewer-administered modes.

In cases where the respondent has more than one mortgage, the Norwegian CATI questionnaire has one loop of questions for each mortgage, beginning with the largest: “How much is left of the largest mortgage?” “How much do you pay in total for this mortgage per month?” “How much of this sum is payment of interests?” “What is the interest rate for this mortgage?” and then asking the same loop for the second and third largest mortgage.

In the CAWI test version, the layout was adapted to whether the respondent had one or more mortgages, as seen in figure 12. Rather than asking three loops with five pieces of information (amount, payment, interest, interest period and interest rate), the information was collected in three tables, the second of which calculates mortgage payment - interest = principal.

Figure 12. EU-SILC HH071 mortgage principal repayment – single and multiple mortgage versions

Har du/dere boliglån eller annet lån med sikkerhet i boligen nå?
Regn med eventuelle slike lån brukt til annet enn bolig
Regn ikke med fellesgjeld

Ja
 Nei

Hvor mange slike lån har du/dere?

1
 2
 3 eller flere

Hvor mye gjenstår av dette lånet?
 Kroner

Hvor mye betaler du/dere for dette lånet per måned?
Oppgi 0 dersom du ikke betaler ned lånet
Regn med eventuelt lån til innskudd, andel, aksje eller andre overtagelsesbeløp
Regn ikke med nedbetaling av fellesgjeld over husleien

<input type="text"/>	Kroner totalt	-	<input type="text"/>	Hvorav renter	=	<input type="text"/>	Avdrag
----------------------	---------------	---	----------------------	---------------	---	----------------------	--------

Hva er rentesatsen på dette lånet?
 %

Har du/dere boliglån eller annet lån med sikkerhet i boligen nå?
Regn med eventuelle slike lån brukt til annet enn bolig
Regn ikke med fellesgjeld

Ja
 Nei

Hvor mange slike lån har du/dere?

1
 2
 3 eller flere

Hvor mye gjenstår av disse lånene?

Største lån Kroner
Nest største lån Kroner
Tredje største lån Kroner

Hvor mye renter og avdrag betaler du/dere for disse lånene?
Regn med eventuelt lån til innskudd, andel, aksje eller andre overtagelsesbeløp
Regn ikke med nedbetaling av fellesgjeld over husleien

	Kroner totalt	Hvorav renter	Avdrag	Termin
Største lån	<input type="text"/> -	<input type="text"/> =	<input type="text"/>	Velg ...
Nest største lån	<input type="text"/> -	<input type="text"/> =	<input type="text"/>	Velg ...
Tredje største lån	<input type="text"/> -	<input type="text"/> =	<input type="text"/>	Velg ...

Hva er rentesatsen på disse lånene?

Største lån %
Nest største lån %
Tredje største lån %

5.1.5.1 Findings

Only the three dwelling owners got the questions on mortgages. They all had mortgages; one, two and three respectively. The owner with one mortgage answered substantially differently in the CATI and CAWI tests only on the question of interest rate. In the CAWI retrospective interview, he commented that the interest rate had gone up, but he was not sure of how much.

The owner with two loans answered the same mortgage payment sums in the CATI and CAWI tests for both loans, but reported twice as much in interest payment for the largest loan in the CAWI test. As she reported the same sum in interest payment for the second largest loan, this was probably not due to her confusing the principal and interest sums. In the retrospective interview, she said that the table gave her a good overview of the numbers, which could mean that the CAWI measurement was most precise for her.

The owner with three mortgages initially answered “one main mortgage” in the CATI interview. The interviewer tried to probe, as the answer could indicate that there were additional mortgages, but ended up coding it as one mortgage. When completing the CAWI test, however, the test person chose the category “3 or more mortgages” first. She started filling in how much was left of the loans in the resulting tables (figure 12, right), but quickly went back to the previous question and answered “2” instead.

After reporting the size of two mortgages, she again went back to the previous question and answered “1”. When she filled in the new resulting table on monthly payments (figure x, left), eye tracking revealed that she spent time both reading the question, the labels “Kroner totalt” (total payment) “Hvorav renter” (whereof interest) and “Avdrag” (Principal). Still, she interpreted the table erroneously, filling in the principal in the “Kroner totalt”. As she estimated both the principal and the interest payment to be NOK 15000, the resulting sum in the “Kroner totalt” was 0. (In the CATI test, she had estimated total monthly payments to be 30000.)

In the retrospective interview, it became clear that the three mortgages were in the same bank, with slightly different interest rates and conditions, and as such it made sense to report them as one loan.

5.1.5.2 Evaluation

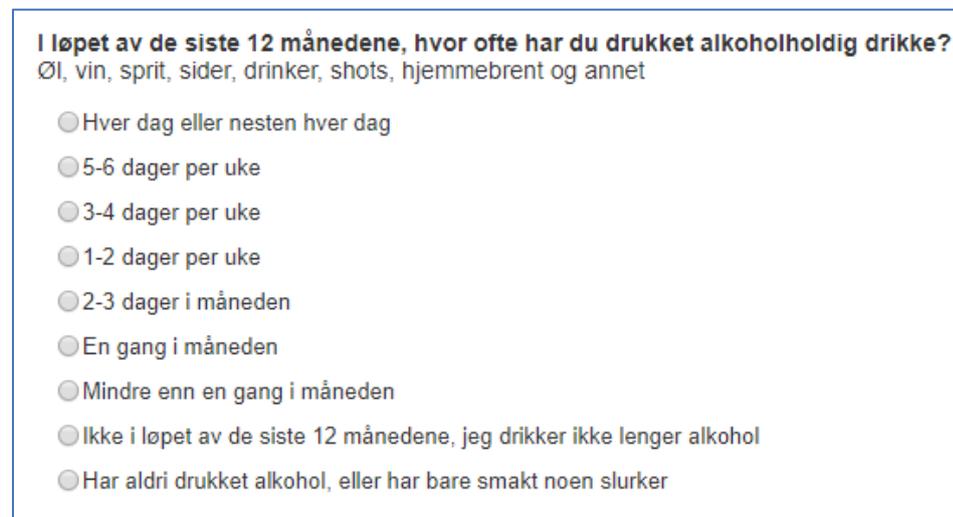
The three-mortgage test person's behaviour can be interpreted as satisficing, but it could also be argued that our efforts at aiding and structuring the questions led to an unnecessarily large response burden. Also, the subtracting calculation in the single-mortgage table question was not sufficiently clear and self-explanatory. The approach was inspired by the logic of business questionnaires, but respondents in social surveys will be even more heterogenous in terms of mathematical skills and ability to cognitively process tabular questions than business survey respondents. Such tools may help some respondents but make the response task more complex or result in measurement error for others.

5.1.6 EHIS – AL1 on alcohol consumption

The question selected from the European Health Interview Survey concerns alcohol consumption during the last 12 months. Question AL1 has nine different response options ranging from “Every day or nearly every day” to “Have never tasted alcohol”. As the response options are lengthy and a challenge for the working memory, the question is better suited for visual modes than aural.

The Norwegian CATI EHIS follows the recommendations described in WP4 deliverable 3, with a branching question asking first whether the consumption is weekly, monthly or more rarely, with follow-up questions for all these three categories. The CAWI version tested asked AL1 as a single question. The aim of the tests was to shed light on whether the branching/non-branching approach affected the judgment and selection phases of the cognitive process differently.

Figure 13. AL1 single question on alcohol consumption in CAWI



I løpet av de siste 12 månedene, hvor ofte har du drukket alkoholholdig drikke?
Øl, vin, sprit, sider, drinker, shots, hjemmebrent og annet

- Hver dag eller nesten hver dag
- 5-6 dager per uke
- 3-4 dager per uke
- 1-2 dager per uke
- 2-3 dager i måneden
- En gang i måneden
- Mindre enn en gang i måneden
- Ikke i løpet av de siste 12 månedene, jeg drikker ikke lenger alkohol
- Har aldri drukket alkohol, eller har bare smakt noen slurker

5.1.6.1 Findings

Of the six test persons, five reported the same frequency of alcohol consumption in the CATI and CAWI tests. The sixth test person was pregnant, and answered “not at all” in the CATI test, based on her current situation. Judging from the recording of the CATI interview, it is possible that she did not hear that the reporting period was the last 12 months. In the CAWI test, she answered 1-2 days per week, based on her usual consumption before pregnancy.

5.1.6.2 Evaluation

Regardless of what may have influenced the judgment of information against the response options, the question presupposes an evenly distributed drinking pattern uninfluenced by seasonal or other variations. The specific situation of the test person only fortified the mismatch between actual behaviour and the reference period. Mode effects may not necessarily be the chief source of measurement error on questions like this.

5.1.7 ICT survey – A1 and A2 on types of Internet connections

The unimode tests of question A2 indicated that respondents have difficulty separating the four types of internet connection, as well as problems with difficult terms. “Don’t know” was used by several test persons; in probably all those cases the substantial answer should have been “No”.

To lessen response burden, the question was tested as an open-ended question in CATI with interviewer coding. In CAWI, the same was intended to be achieved by presenting all four types of connections as response options and removing most of the technical terminology, as a simplified version of the model questionnaire question.

Figure 14. ICT CAWI mode specific A1/A2

Hva slags internettforbindelse bruker husholdningen hjemme?
Her er det mulig å sette flere kryss

- Fast bredbånd, for eksempel via ADSL, kabel-TV, fiberkabel eller satellitt
- Mobilt bredbånd med 3G eller høyere
- ISDN eller modem med vanlig telefonlinje
- Mobil internettilkobling med 2G eller lavere
- Ingen, har ikke internett hjemme

5.1.7.1 Findings

In the CATI tests, the interviewer coded that “Fixed broadband connections” was used at home by all six test persons, based on their open responses. No other types of Internet connection were coded, but the interviewer did not probe for such information either. In the CAWI tests, the six test persons all responded that they used “Fixed broadband connections” at home as well. Additionally, two test persons answered that they used “Mobile broadband” at home. In the retrospective cognitive interview, a third test person was unsure of whether to select this option, and a fourth said that she used mobile broadband, but not at home. When the fixed broadband connection is temporarily unavailable, the mobile broadband will be used, mostly on smartphones, but not normally or regularly.

5.1.7.2 Evaluation

The approach we tested will likely result in some measurement differences, but the best way of designing the question will depend upon whether the occasional use described above is to be measured or not. If not, a better solution will probably be single questions in unimode asking whether the respondent has a fixed broadband connection or not.

5.1.8 ICT survey – B3 and B5 on types of units used for Internet activities, and types of Internet activities

On these questions, the test persons had reacted negatively to the repetitive aural/visual information in the unimode tests. Here too, we tested asking a single question with a check all that apply format rather than a series of yes/no questions.

Figure 15. ICT B3 and B5 CAWI mode specific

På hvilke av disse enhetene har du brukt internett de siste 3 månedene?
Her er det mulig å sette flere kryss

- Stasjonær PC
- Bærbar PC
- Nettbrett
- Mobiltelefon/smarttelefon
- Spillkonsoll, nettbokleser, smartklokke eller annet mobilt utstyr
- Ingen, har ikke brukt internett de siste 3 månedene

Hvilke av følgende aktiviteter har du brukt internett til privat de siste 3 månedene?
Her er det mulig å sette flere kryss

- Sende eller motta e-post
- Bruke Skype, FaceTime eller andre telefoni-/videosamtaleapper
- Delta på Facebook, Twitter, Instagram, Snapchat eller andre sosiale nettverk
- Søke etter informasjon om varer eller tjenester
- Høre på musikk via nettradio, Spotify, Tidal eller andre strømmetjenester
- Se på TV direkte eller i opptak fra kringkastingsselskaper
- Se filmer eller serier på Netflix, HBO Nordic eller andre bestillingstjenester
- Se video på Youtube eller andre delingstjenester
- Spille eller laste ned spill
- Søke etter helserelatert informasjon, for eksempel om skader, sykdom, ernæring eller andre helsetiltak
- Avtale en legetime eller time på sykehus
- Selge varer eller tjenester på Finn, Ebay eller lignende
- Utføre banktjenester i nettbanken

5.1.8.1 Findings

On question B3 on types of units used to access the Internet, the test persons gave the same answers in the CATI and CAWI test, with one exception. The exception was an instance where a test person reported having used the Internet on an e-reader device in the CATI interview, but did not select this option in the CAWI test. During the retrospective interview after the CAWI test, she said that she

should have selected this option. She commented that she did not use the term “nettbokleser”, which is the one used in the response option, but rather “lesebrett”. This mismatch, combined with the rather superficial reading through of the response options that was revealed by the eye tracking recording probably contributed to the measurement error.

On question B5, the six test persons answered “yes” to a total of 64 response options between them in the CATI test. Of these, 61 were also selected in the CAWI test. Further, two options were selected in CAWI but not in CATI. The three instances where an option was not re-selected in CAWI was “playing or downloading games”, “finding information about goods or services”, and “searching for health-related information”. The three instances where a response that had not been chosen in CATI was selected in CAWI, was “Listening to music via web radio [etc]”, and “finding information about goods or services.” In the retrospective cognitive interviews, one of the test persons commented on the vagueness of the term “finding information about goods or services”. She said that “usually, I want to find out whether an article exists or not, I am not searching for information about it”. Also, she noted that the three-month reference period could be problematic for activities that were rarely carried out. Many activities were carried out daily and were no problem to respond to. Another test person commented that games would be downloaded for the children, and that it was unclear whether it had to be downloaded for own use or not.

5.1.8.2 Evaluation

The match between responses in CATI and CAWI mode should be considered sufficiently good for these questions. Activities that are carried out more infrequently will be more prone to measurement errors, as will activities that are vaguely specified. Both these sources of errors are unrelated to mode, and improvements in wording and relevance of activities would benefit both CATI and CAWI. On these questions, the tests indicate that a mode-specific design for CAWI on these questions can lead to a reduced response time/burden, without compromising data quality.

5.1.9 ICT survey – D2 on goods and services bought or ordered over the over the Internet for private use last 12 months

This question was also perceived by several respondents as repetitive and boring in the unimode version. Also, many respondents were unsure of whether they had purchased “other” goods or services. This is due to working memory issues, as it is difficult to remember all the 14 categories, and then retrieve and process the information. Unclear response categories may also add to the problem. A visual presentation on one page could add context and possibly be helpful. (Though the best solution would more likely be to have fewer and clearer categories.)

Figure 16. ICT D2 CAWI mode specific

Hvilke av følgende varene og tjenestene har du kjøpt eller bestilt over internett til privat bruk de siste 12 månedene?
Her er det mulig å sette flere kryss

- Mat eller dagligvarer
- Forbruksvarer som møbler, leketøy og lignende, unntatt elektronisk utstyr
- Legemidler
- Klær eller sportsutstyr
- PC-maskinvare, inkludert skjerm, tastatur, mus
- Annet elektronisk utstyr, inkludert kamera, radio, TV, stereo
- Abonnementer på TV-kanaler, bredbånd, fasttelefon, mobiltelefoni, kontantkortpåfyll eller andre teletjenester
- Innkvartering i forbindelse med ferie
- Reisebilletter (fly, buss, båt, tog) billette eller andre tjenester i forbindelse med reiser
- Billetter til arrangementer
- Filmer eller musikk, inkludert strømmetjenester som Netflix, HBO, Spotify, TV2sumo osv.
- Bøker, tidsskrifter eller aviser, inkludert e-bøker
- Elektronisk læremateriell
- Dataspill, TV-spill, programvare og programvareoppdateringer
- Andre varer eller tjenester

5.1.9.1 Findings

On this question, a total of 46 “yes” responses were recorded in CATI, whereof 41 were also selected in CAWI. Additionally, three responses were selected in CAWI but not in CATI.

Eye tracking revealed different reading strategies, as some test persons only went through the list once, whereas others went through it once, before reviewing their responses and adding or removing options.

During the retrospective interviews, one of the test persons commented that several of the categories were unclear and lumping different things together, e.g. TV subscriptions, pay-per-view and different phone costs. She also thought the order was somewhat non-intuitive order, e.g. the “Software” etc. category (14 from the top) should be closer to “PC hardware” etc. category (5 from the top). Another test person commented that the “other goods or services” category could be split in two, as she frequently bought cleaning services online.

5.1.9.2 Evaluation

The responses to D2 did not match quite as good between CATI and CAWI as was the case for B3 and B5. This, however, could also partly be caused by the longer reference period of 12 months. The different reading strategies evidenced by eye tracking indicates that the mode-specific way of presenting all response options as a list will give some overview. By not forcing respondents to actively evaluate whether they have procured “other goods and services”, the response burden is reduced. If getting an answer to this question is of great importance, it would be advisable to ask a yes/no question with reference to the above question: “In addition to the ones in the categories you

selected in the question above, did you buy or order any other goods or services over the Internet during the last 12 months?”

On this question as well, the tests indicate that the categories’ content and order could benefit from a review, to make the question better understandable regardless of mode, and regardless of a unimode or mode specific strategy.

6 Discussion of unimode and mode specific designs tested

The aim of running the tests described in this report was to gather more information on how unimode and mode specific question designs would work for key and typical questions from ESS surveys that could be vulnerable for measurement effects due to different modes being used.

From the unimode tests of the ICT questionnaire, perhaps the main finding, however, was the room for general improvement of the questions regardless of mode. Cutting down on the total amount of text/spoken words, simplifying the language, making response categories more relevant and intuitive, and in some cases fewer, would likely reduce both response burden and measurement errors. Some of this improvement work can and should be done by Statistics Norway, through better operationalization and adaptation on our national version of the ICT questionnaire. Some of the more basic problems, however, should be addressed by Eurostat in the opinion of these authors.

For the mode specific tests of selected ESS questions, results must be said to be mixed. The more ambitious experiments with calculations in tables were less successful. The business survey way of thinking that inspired the questions is not necessarily suited for respondents in a cross-sectional sample. (Nor is it always for business respondents, but that is another discussion.) The cognitive and practical (e.g. retrieving documentation on work hours or mortgages) processes involved in answering the questions we had selected can differ from respondent to respondent, and not necessarily match the process and sequence of operations presupposed in our tested questions. Another problem is that several of the tools we tested would work poorly on mobile phones or other small screens.

On some of the ICT survey questions where a check all that apply questions is presupposed in the model questionnaire, there is evidence that this format can be kept in visual modes, but be replaced with a series of yes/no questions in aural modes. This requires that the response categories are clear and relevant to the respondents, and that there are not too many of them. The relevance and number of response categories on several of the ICT questions is another issue that should be addressed by Eurostat rather than the individual NSIs. The number of response options is another issue that make questionnaires less suited for mobile phone screen sizes.

This is not to say that the more ambitious mode specific approaches should be abandoned. There is a need for continuous testing of possible solutions to address the challenges we are facing as questionnaire designers and data collectors, mode specific as well as unimode. New and emerging technologies, like voice recognition and bots, could potentially bridge some gaps between modes, but also create new measurement issues that have yet to be charted and dealt with. This could also mean that we will need to develop new test methods.

7 Discussion of test design and further research

Our aim of testing on the one hand a whole ESS survey, and on the other hand several key questions from various sources, and at the same time provide input for WP5 on mobile adaption of questionnaires, meant that we had to make some compromises to avoid having too many dimensions to consider during analysis. We did this by only testing among people in a limited age

range, with a certain proficiency in the use of ICT tools. The tests where we found the more promising results should also be tested among respondents that are less proficient, but still on devices that they use or are familiar with.

The combinations of test methods used – expert evaluation before testing, behaviour coding for CATI test interviews, eye tracking for CAWI tests, and retrospective cognitive interviewing all contributed to a better insight into how the questions worked. The behaviour coding classification could have been better specified and discussed among coders. The eye tracking equipment worked best in the CAWI PC tests. On the small screen it was more difficult to assess what was being focussed on, and the contraption used for holding the respondents' smartphones in place made the situation somewhat unnatural. E.g. the test persons were not free to turn their phones to achieve a landscape view.

For future test-retest designs of mixed-mode questionnaires, we should consider waiting longer than one week between the test and the re-test. During the CAWI re-test, several test persons commented that they recognized questions, and some said that they tried to think back to what they had answered in the CATI test. We should also consider adding some questions in the tests that will not be a part of the re-test, and vice versa. On some of the questions we tested, we should also have added more questions on the same topic, to make a more realistic context. This would for instance apply to the question from the LFS on actual working hours.