COOPERATION ON MULTI-MODE DATA COLLECTION (MMDC)

MIXED MODE DESIGNS FOR SOCIAL SURVEYS - MIMOD

GRANT AGREEMENT FOR AN ACTION WITH MULTIPLE BENEFICIARIES

AGREEMENT NUMBER – 07112.2017.010-2017.786

Recommendations for key questionnaire elements, questions and question types in mixed mode settings

Related to the European Statistical System's person and household surveys

WP4 – Deliverable 3

Date: January 31, 2019

Dag F. Gravem (SSB)
Nina Berg (SSB)

WP4: Mixed-mode designs

Contents

> *“If there is even a small chance of mixing modes in the project, design the questionnaire for the possibility of mixed-mode data collection”*[1]

[1] Guideline 11.7 in *Internet, Phone, Mail and Mixed-Mode Surveys – The Tailored Design Method* by Don A. Dillman, Jolene D. Smyth and Leah Melani Christian

Recommendations for mixed-mode questionnaires and mixed-mode questions in the European Statistical System's person and household surveys

1. Introduction and summary

Currently, as well as historically, the social surveys of the European Statistical System (ESS) have been conducted in different data collection modes and using different technologies. The choices of each country have been shaped by ESS guidelines, but also economy, tradition, competence, geography, degree of development and other factors. Currently, there is a shift from interviewer-administered data collection to self-administered web data collection, but neither interviewing nor paper questionnaires are dead.

Deliverable 1 of MIMOD WP4 showed that on a European level, *all* ESS surveys are mixed-mode. Deliverable 2 demonstrated some of the possibilities and constraints of communication and follow-up strategies that accompany the modes and mode changes in data collection. These facts ought to have implications for how survey specifications and requirements are developed and implemented, allowing for flexibility while ensuring that measurement differences are minimized.

This deliverable will review key ESS surveys and some of their central questions and assess their fitness for the mixed-mode reality: The Household Information and Communication Technology survey (ICT), the European Health Interview Survey (EHIS), the European Standards of Income and Living survey (EU-SILC).

For each survey, the deliverable takes Eurostat's model questionnaires and other documentation as a starting point, before reviewing some national adaptations of questions. Then, results from cognitive and usability testing of survey questions carried out for the MIMOD project is presented. Further documentation from these tests is found in Appendix B. Lastly, some recommendations, as well as suggestions for further testing and development is offered. These recommendations are aimed at Eurostat as much as at the National Statistical Institutions (NSIs).

2. Theoretical frameworks

Many authors have written about how the choice of data collection mode(s) influences the quality of a survey. In the chapter "Survey Mode and Mode Effects" from *Improving Survey Methods* (2014), de Leeuw and Hox review the literature and research and discuss how mode can influence coverage, sampling and nonresponse error, as well as measurement error. They stress that mixed-mode designs can be used in such a way that the strengths of one mode compensates for the weaknesses of the other.

Roughly, mode effects can be divided into "pure mode effects" that have to do with the instruments themselves, and "other mode-related bias" which has to do with different response propensities and coverage in different modes for different demographic groups. This deliverable will deal with instrument-related "pure mode effects" only, leaving the consideration of "other mode-related bias" for other work packages. We will examine the strengths and weaknesses of the different modes regarding key questions and question types in the ESS surveys, as well as the surveys' general fitness for mixed mode data collection.

2.1. Unimode, mode-specific and generalized mode question design

In their chapter, de Leeuw and Hox discuss the fact that empirical mode comparisons generally find only small mode effects, except in the case of sensitive questions where self-administered modes produce better results. They argue, however, that this lack of observed effects could be explained by the fact that the researchers take great care to make the instruments as similar as possible in the different modes. This strategy is often referred to as a *unimode* design.

In practice, de Leeuw and Hox go on to argue, in daily practice, there are different design conventions. Often, the survey designers try to optimize the questions for each mode. The authors use the term *question format effects* for the measurement differences this can lead to. It could also be called measurement differences due to *mode-specific* question design. When such optimized question designs do not yield measurement effects, it could perhaps be argued that the designers have discovered the holy grail of mixed mode question design: *generalized mode*, where different stimuli yield the same perceived stimulus, and consequently the same response.

In our review of how the ESS questionnaires (and others?) have been implemented in practice, we will look for and review all these types of question design.

2.2. The DCSS project – Key factors for differential measurement

Other authors go more deeply into the causes of mode effects. A deliverable from the ESSnet Data Collection for Social Surveys using Mixed Modes (DCSS), the predecessor of the MIMOD project, investigated this (Körner et al. 2013). The authors present a typology of key factors for differential measurement in different modes, with an overview of possible measurement effects and question types was developed. This typology takes the *survey mode* as a starting point for analysis, describing how the following four main factors can lead to possible measurement effects:

1) the type of social interaction – e.g. interviewer involvement, respondent's control

2) type of communication – verbal, non-verbal, para-verbal, computer mediated

3) questionnaire design options – visual vs. aural stimulus

4) computer assistance – e.g. skip instructions, consistency and plausibility checks, coding

In sum, the authors single out the following possible measurement effects:

- Social desirability bias
- Satisficing
- Question order effects
- Social de-contextualisation
- Recency effects
- Primacy effects
- Measurement of deviating concepts
- Routing errors
- Item nonresponse
- Errors in completion
- Deviation from interviewer protocol

Several of these have to do with data collection protocol and question format effects, rather than the questions per se. Regarding which question *types* have the risk of mode effects is highest, the authors single out

- Sensitive questions
- Difficult questions
- Long questions
- Questions with long lists of response items
- Questionnaires with complex skip instructions

This is a rather short and unspecified list of question types, but the results follow from the authors' aim of synthesizing theoretical approaches to mode effects. For our purposes, we decided to go to one of Körner et al.'s sources.

2.3. Campanelli et al. – Question characteristics relevant to measurement error

A different, "bottom up" approach was taken in the early 2010s by Pamela Campanelli and a group of survey methodologists, who used *questions* as the starting point. In their "Classification of Question Characteristics Relevant to Measurement Error and Consequently Important for Mixed Mode Questionnaire Design" (2013), they point out 29 different characteristics of individual questions, grouped by a) Type of Task, b) Characteristics of task and c) Implementation of question (see appendix A for details). They provide examples and references to literature and offer recommendations on which modes to use for each of the characteristics. (See appendix A for details.)

One single question can have several characteristics, and thus be judged differently according to each one. A weakness of the Campanelli typology is that they do not distinguish between paper self-completion and web self-completion in their recommendations, thus downplaying the importance of the factor of computer assistance, although this is discussed for applicable question types.

Regarding mode recommendations, questions having one of the following characteristics are judged as *not recommendable* for at least one mode, with reference to the numbering in the quoted paper (See appendix A):

1) Sensitive questions
4) Subjective, non-sensitive scalar questions
6) Unconstrained textual/verbal open questions
10) Open questions with interviewer coding
14) Mark all that apply response format
16) Ranking questions
18) Visual analogue questions
20) Questions with a high number of response categories
24) Questions with edit checks
29) Questions with show cards

Although useful for guidance on different mode effect risk factors, the model is difficult to use for quantifying and determining a survey or a question unfit for mixed-mode. Many of the typical questions with mixed-mode issues that we have encountered can be classified as having two or more characteristics. One possible way of quantitative use could be to define a threshold for the amount or proportion of questions unfit for mixed mode.

Instead, we used it as an inventory for pinpointing problems in the reviewed ESS questionnaires, and for references to relevant literature and areas where further research is needed. The questionnaires were then reviewed using the Campanelli criteria as a checklist, in pursuit of questions that in different ways violated the criteria if implemented in CAWI mode. We also looked for questions that could be problematic in CATI mode, but where the visual and computing strengths of CAWI mode could possibly be used in a mode-specific way. If an example was found in more than one questionnaire, it was in some cases added for comparison in the tests.

3. Delimitation from WP5

The purpose of WP4 is to offer recommendations on mixed-mode implementation of questions, with an emphasis on CAWI, but regardless of screen size. The mission of WP5 is to determine to what degrees ESS surveys are fit for mobile CAWI. In practice, all CAWI surveys are mobile web surveys, as experience shows that some respondents will always try to complete questionnaires using mobile phones.

Blocking respondents from responding using mobile phones is generally not recommended, as response rates and representativeness will decline, as well as complicating follow-up phases of data collection. Rather, a decision should be made on whether to make a mobile optimized *version* of the questionnaire, or to make a mobile first/mobile friendly questionnaire. Discussions of such decisions belongs to WP5 rather than WP4, but are also of relevance for WP4: which surveys are suitable for mixed mode with a mobile first approach, and which are suitable for mixed mode with an optimized mobile web questionnaire?

4. Assessments of model questionnaires/specifications and national implementations. Identification and testing of possible mode-sensitive questions.

For the assessment of which surveys are suited for mixed-mode data collection, and of questions and question types that require revisions, different sources of information are available. For the ICT survey and the EHIS survey, there are model questionnaires, and for the AES there is a “reference questionnaire”. For the LFS, minimal requirements on variable level are available, and the EU-SILC has extensive methodological guidelines available, including mode recommendations. Other than the model questionnaires and variable specifications, national questionnaires are also available for review, and some references to these have been included. Based on these reviews using the Campanelli classification, certain questions have been selected for user testing. The results from these user tests will also be discussed in this chapter.

As previously stated, it differs between countries whether each of the ESS surveys is conducted as a stand-alone project or combined with national or other surveys. Consequently, there can be national differences in each survey’s suitability for mixed-mode. In addition to assessing the ESS surveys’ fitness for mixed-mode data collection, the fitness of some national implementations will also be discussed.

4.1. ICT survey

4.1.1. Description of legal basis/guidelines/model questionnaire

The ICT regulation (Commission Regulation (EU) 2017/1515 of 31 August 2017 implementing Regulation (EC) No 808/2004) does not contain any recommendation or limitation regarding data collection modes to be used. Of methodological documentation, a model questionnaire does exist. Although different modes are not discussed in the model questionnaire, it presupposes a visual format, e.g. by referring to the “tick all that apply” response format (see below). The MIMOD survey showed, however, that CATI interviewing is the most used mode, by 16 ESS NSIs. The visual mode CAWI is the second most used, by 15.

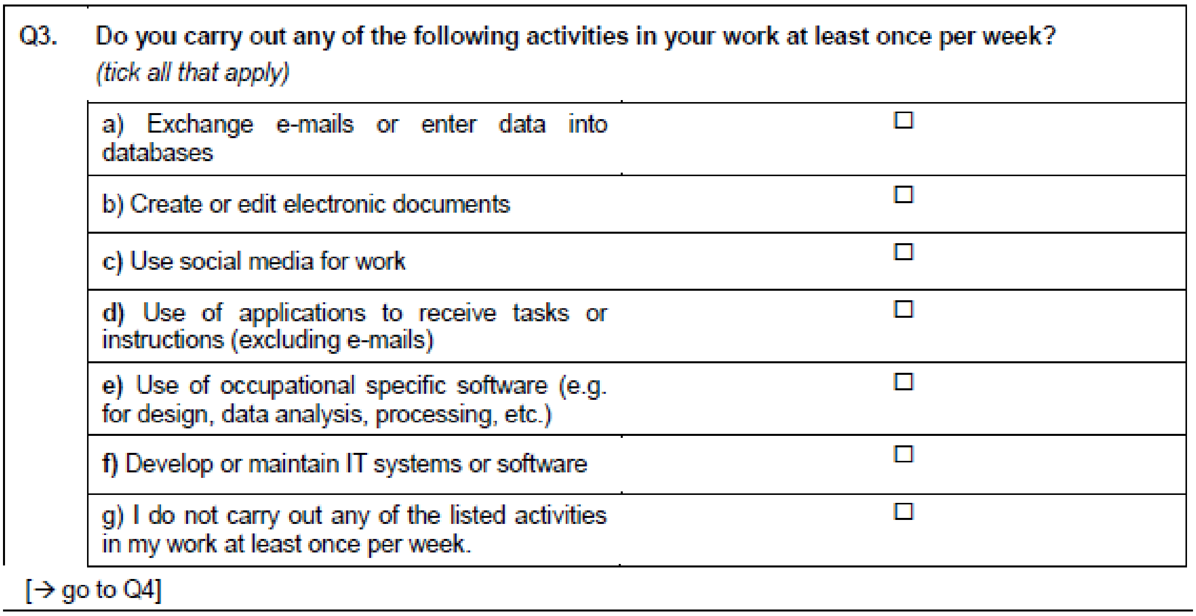
4.1.2. Mixed-mode related challenges in model questionnaire

The main issues identified in the ICT questionnaire are related to the following issues, with reference to the numbering in the Campanelli typology:

A. Mark all that apply format (14)
B. Inherent difficulty due to concepts (6)
C. Use of instructions and clarification (23)
D. Use of Don't know (25)
E. Question length (Dillman)

A. The model questionnaire uses the "tick all that apply" format rather than "yes/no". (Figure 1) This is visually possible in CAWI, PAP and CAPI using show card, but not in CATI mode. In CATI, such questions must either be asked as an open question, with interviewer coding, or be changed to a series of yes/no questions. The former could lead to underreporting or coding errors, whereas the latter could lead to a higher reporting than the check all that apply.

Figure 1. Example of "check/tick all that apply" recommendation from the ICT model questionnaire



Q3. Do you carry out any of the following activities in your work at least once per week?
(tick all that apply)

a) Exchange e-mails or enter data into databases	□
b) Create or edit electronic documents	□
c) Use social media for work	□
d) Use of applications to receive tasks or instructions (excluding e-mails)	□
e) Use of occupational specific software (e.g. for design, data analysis, processing, etc.)	□
f) Develop or maintain IT systems or software	□
g) I do not carry out any of the listed activities in my work at least once per week.	□

[→ go to Q4]

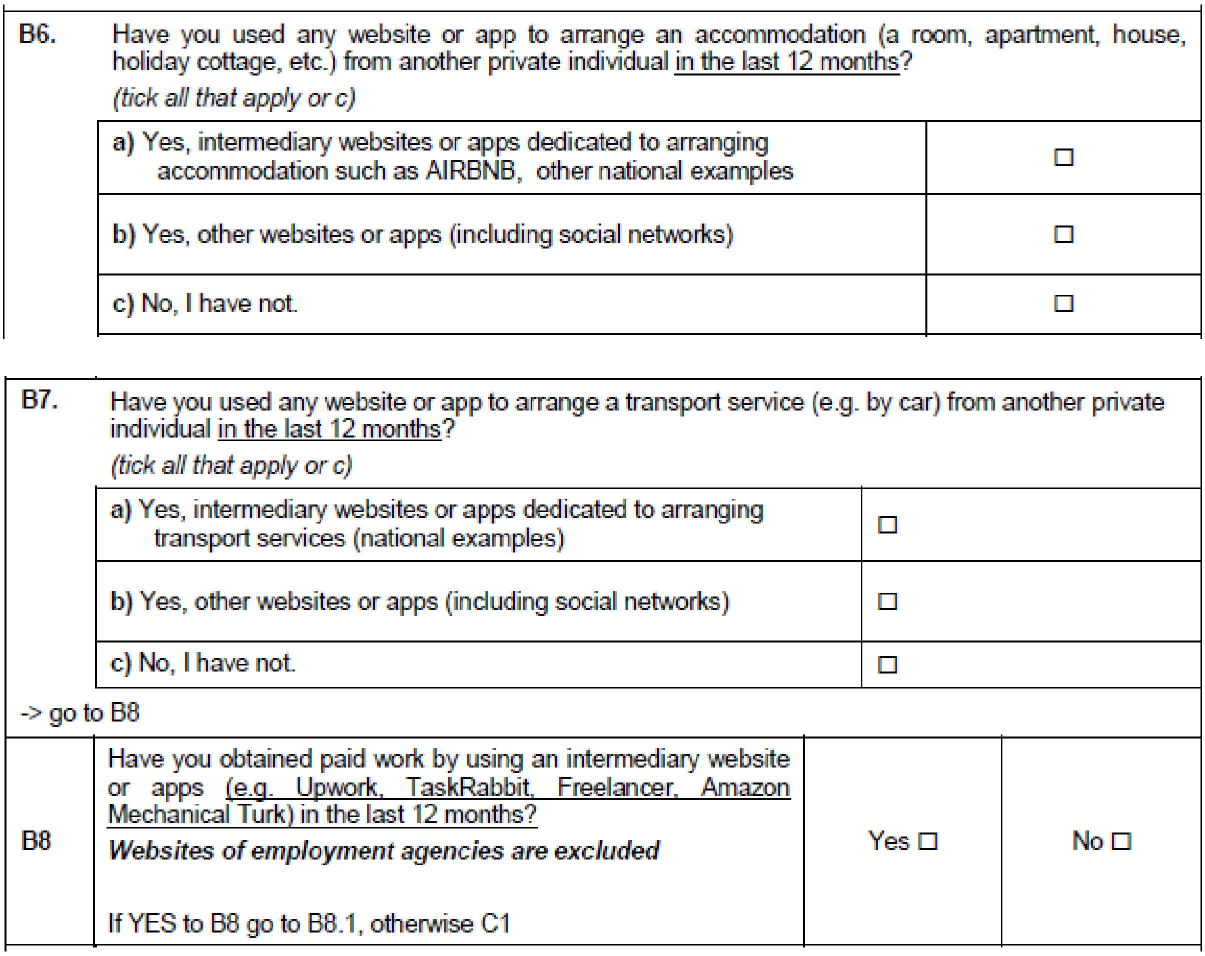
This lack of specificity means that there is a range of different solutions possible and necessary for the countries implementing the ICT survey depending on which mode(s) they use.

The most problematic question of this type is probably D2, asking which types of goods the respondent ordered over the internet in the last 12 months. With 15 response categories, many of which are thematically close, there is a clear danger of primacy effects with a "tick all that apply" format, as well as an excessive burden. One possibility is to replace this with branching questions

B. and C. Difficult concepts and use of instructions. The ICT survey contains very many technical terms, and several of them look more like theoretical variable descriptions than fully operationalized questions, see e.g. figure 1 above with the phrase "create or edit electronic documents" rather than "use Word, Excel or similar applications".

On other questions, examples or clarification is available in parenthesises (figure 2). There are no guidelines regarding how this should be implemented in aural modes. In question B8 on obtaining work by using an intermediary website or app, Amazon Mechanical Turk and other examples are mentioned in parenthesis. The information that employment agencies are excluded is not in parenthesis, however.

Figure 2. Examples of instructions and difficult concepts from the ICT model questionnaire



B6.	Have you used any website or app to arrange an accommodation (a room, apartment, house, holiday cottage, etc.) from another private individual <u>in the last 12 months</u>? *(tick all that apply or c)*	
	a) Yes, intermediary websites or apps dedicated to arranging accommodation such as AIRBNB, other national examples	□
	b) Yes, other websites or apps (including social networks)	□
	c) No, I have not.	□

B7.	Have you used any website or app to arrange a transport service (e.g. by car) from another private individual <u>in the last 12 months</u>? *(tick all that apply or c)*	
	a) Yes, intermediary websites or apps dedicated to arranging transport services (national examples)	□
	b) Yes, other websites or apps (including social networks)	□
	c) No, I have not.	□
-> go to B8		

B8	Have you obtained paid work by using an intermediary website or apps <u>(e.g. Upwork, TaskRabbit, Freelancer, Amazon Mechanical Turk) in the last 12 months?</u> ***Websites of employment agencies are excluded*** If YES to B8 go to B8.1, otherwise C1	Yes □	No □

In the question B6 on the relatively recent phenomenon of "sharing economy" online marketplaces for accommodation, AIRBNB is explicitly mentioned in one of the response categories. In the corresponding question B7 on transportation, Uber is however not mentioned. The way these questions are presented in the ICT model questionnaire are only suitable for visual self-completion modes, and ought to be reworked for interviewer-administered modes.

D. Use of "Don't know". On four of the model questionnaire questions, "Don't know" is offered as a non-substantial response category. In practical implementations, a non-explicit Don't know in CATI is sometimes combined with an explicit, visible Don't know option in CAWI. Campanelli recommends that spontaneous Don't know is not allowed if interviewer-administered and self-completion modes are combined, to avoid measurement differences. On many of the questions in the ICT questionnaire, "Don't know" appears to be a perfectly valid response, however, and not due to satisficing – several of the factual questions involve difficult and theoretical terms and instructions, as discussed above.

E. Question length. This is not explicitly discussed by Campanelli et al., but a central recommendation from Don Dillman et al. in mixed-mode surveys where CATI is involved, to "give priority to the short and simple stimuli needed for telephone". In aural CATI, there is a higher demand on the respondents' memory capacity for remembering complex sentences, additional clarifying information and long lists of response categories.

4.1.3. National implementations and experiences

In Norway, the ICT survey is done as a CATI interview, embedded in a 35-minute omnibus on a range of topics. Consequently, all check all that apply questions are yes/no, intended to be read out loud. The AIRBNB and Uber questions are however asked in the CATI suboptimal form described in the model questionnaire, with the long response categories being read out loud. Although this is against the principles of a short stimulus, interviewers debriefed informed that the questions worked fairly well. On the other hand, the interviewers also stated that the questions were generally too long.

The Netherlands does the survey in a mixed-mode CATI/CAPI/CAWI design. The Dutch questionnaire is implemented using a unimode strategy, where everything that is visible on the screen for CAWI respondents is also read out loud for CATI interviewees. CBS also implements Don’t know as part of this, reading out loud the Don’t know options on the questions where this is specified in the model questionnaire. They have also added Don’t know options for questions where they consider this to be relevant and necessary.

4.1.4. MIMOD user tests

Statistics Norway used the ICT survey as a case study for the unimode approach. The questions were presented in CATI and CAWI format as similarly as possible. Additionally, some questions were tested in a mode-specific version for CAWI. The tests confirmed many of the concerns that the evaluation according to Campanelli criteria had uncovered, but concluded that some main problems were not connected to mode:

The main problems with the ICT survey questionnaire that was used for the MIMOD [unimode] 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.

One factor that significantly contributed to the length and verbosity of the test questionnaire, was 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. 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. These tests indicate that the “check all that apply” format can be used in CAWI and achieve comparable results with open questions with interviewer coding in CATI. However, the feasibility of this approach will be limited by the number of response categories.

4.1.5. Key and typical questions – recommendations and suggestions for further testing

As there is not one dominant mode, either in terms of visual vs. aural, or interviewer vs. self-administered modes, the ICT model questionnaire should be reviewed and presented in a mode-neutral way. The “Check all that apply” questions should be considered replaced with yes/no sequences, as well as shortened or made more relevant. For some short ones, like the question on Internet connections, a check all that apply format could be acceptable in CAWI, while using a sequence of yes/no questions in CATI. Where this is not desirable, alternative ways of asking, such as branching, should be examined and tested. Long questions, and the placement (CAWI) and transmitting (CATI) of clarifying information should be reworked and tested and be standardized according to unimode principles as far as possible.

Allowing Don’t know should be considered for more questions. The concept of greying out Don’t know response options seems to prevent casual non-substantial responses and should be tested further in this and other surveys.

4.2. EHIS

4.2.1. Description of legal basis/guidelines/model questionnaire

For EHIS, a methodological manual with a model questionnaire exists. This is explicitly designed for face-to-face modes, which are described as the “preferred modes”. The responsibility of adapting the questionnaire to other modes is left to the participating countries, however:

> *Because several other survey modes like computer-assisted telephone interviews (CATI), computer-assisted web-based interviews (CAWI), self-completion mode as well as mixed-mode designs will be applied by the responsible national authorities, adaptations of the model questionnaire to the requirements of a specific survey mode may be necessary. (European Health Interview Survey (EHIS wave 3) Methodological manual. Eurostat 2018)*

Metadata on type of data collection is to be reported via the survey’s INTMETHOD variable, which has nine different values, and is collected for each interview. The manual states that this is for enabling an analysis of possible mode effects on the results. The guideline also contains an evaluation of which questions are suited and allowed for self-completion modes.

The document does contain guidelines for how to handle synonyms and clarifications, and how this should be handled in interviewer and self-administered modes:

> *a) Synonyms (for example: "Myocardial infarction" and "heart attack") and explanations of abbreviations ("GP" and "General practitioner"). The text doesn't have to be read in case of personal interviews (but can if needed) but may be useful to put in the questionnaire for self-completion mode.*
>
> *b) Clarifications or specifications (for example: "normal work" and "including both work outside the home and housework"; or "Asthma" and "allergic asthma included"). The text is considered to be part of the question and is supposed to be read in case of personal interview and be part of a question in case of self-completion mode.*

As part of the guidelines, 18 suggestions for show cards for face-to-face modes are included. These are both overviews of response categories and function as question batteries (e.g. CD, a list of diseases and chronic conditions where the respondent is to respond yes/no to each one, see below).

There is a separate document with further recommendations for CATI mode on certain questions: *Improvement of the European Health Interview Survey (EHIS) modules on alcohol consumption, physic al activity and mental health. Final Report* (Berlin 2011), from a Eurostat grant. It does contain some recommendation on adding filter questions and interviewer instruction, as well as probing protocols for questions with show cards.

4.2.2. Mixed-mode related challenges in model questionnaire

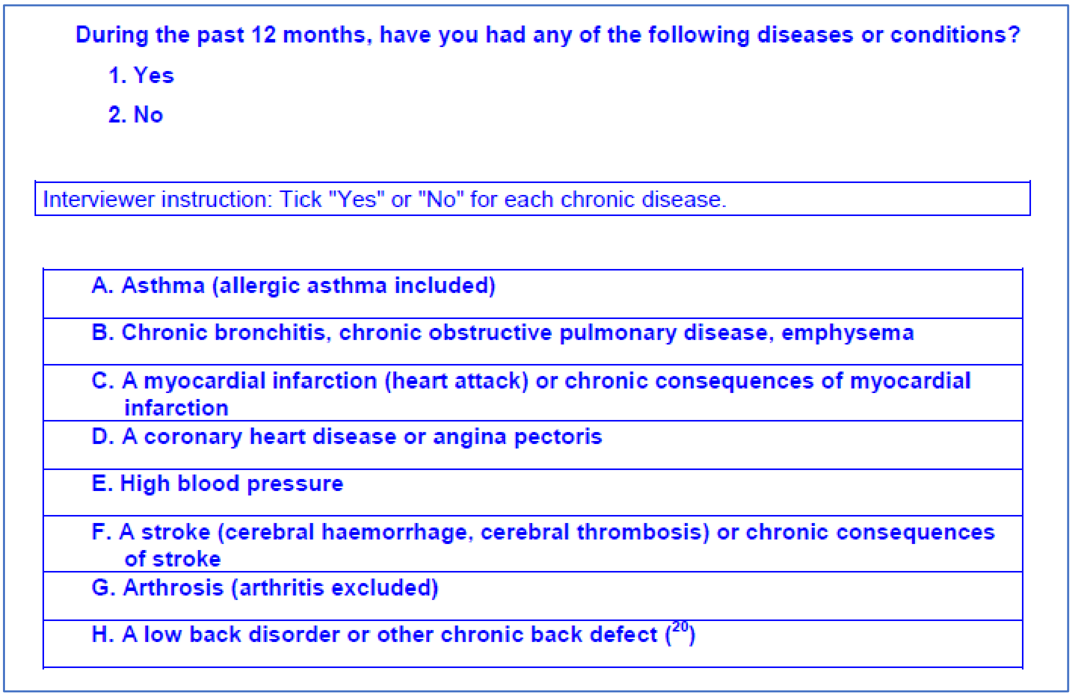
Applying the Campanelli criteria, the following issues are of particular importance for the EHIS questionnaire:

A. Sensitive questions (1)
B. Long “yes/no for each” question batteries (15)
C. Show cards (29)
D. Inherent difficulty due to difficult terms and calculations (5)

Regarding A. sensitive questions, this relates to the many questions on health and medical conditions, which could be underreported in interviewer administered modes. In the MIMOD survey, however, one of the responding NSIs stated that CAPI was the preferred mode because eye contact could verify or reveal whether a respondent has certain health problems. While this may be the case for some physical conditions, it will be difficult to assess for many others.

Regarding B. and C., the EHIS questionnaire can be said to be in the opposite situation of the ICT questionnaire. Where the ICT model questionnaire presupposes the “check all that apply” format, EHIS presupposes yes/no to each question. In the case of question CD (figure 3), the respondent is to be presented with a show card, but the interviewer is also expected to get a yes or no from the respondent for each.

Figure 3. Show card question with yes/no for each



During the past 12 months, have you had any of the following diseases or conditions?

1. Yes
2. No

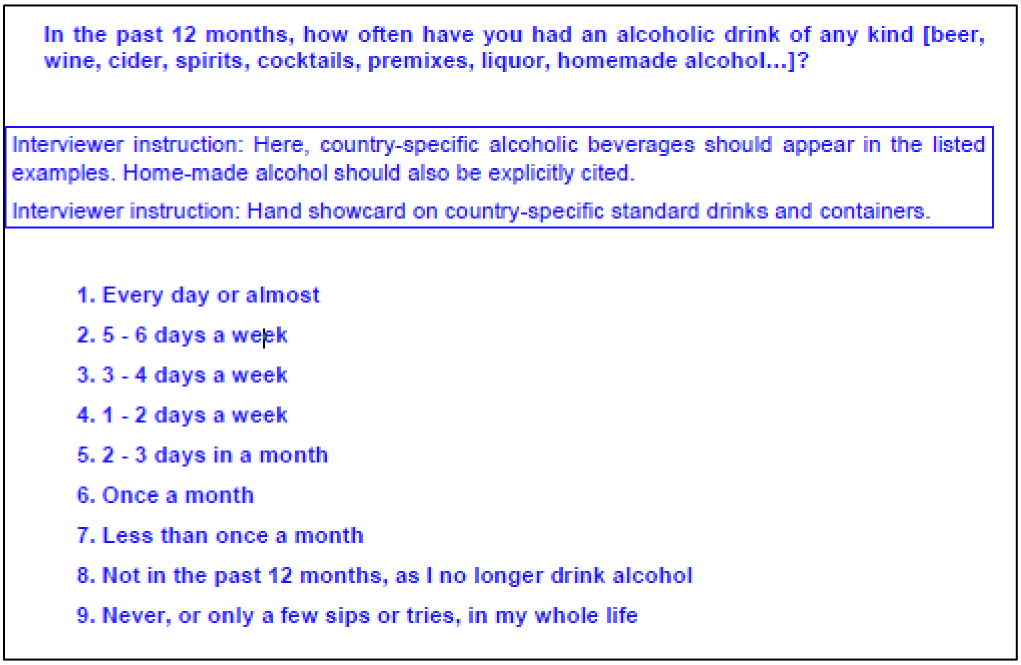
Interviewer instruction: Tick "Yes" or "No" for each chronic disease.

A. Asthma (allergic asthma included)
B. Chronic bronchitis, chronic obstructive pulmonary disease, emphysema
C. A myocardial infarction (heart attack) or chronic consequences of myocardial infarction
D. A coronary heart disease or angina pectoris
E. High blood pressure
F. A stroke (cerebral haemorrhage, cerebral thrombosis) or chronic consequences of stroke
G. Arthrosis (arthritis excluded)
H. A low back disorder or other chronic back defect ([20])

In CAWI mode, it is possible to present this both as a grid question with yes/no columns, in a checkbox format, or as a series of single-screen yes/no questions. Although the yes/no for each format recommended by Campanelli typology, its authors warn that too many such questions will increase completion time and response burden. This also means a risk of breakoff.

AL1 is an EHIS show card question with some different challenges. It is a single response only, but also implies potentially difficult calculations (figure 4):

Figure 4. AL1 - Single-response show card question, intended for face-to-face mode



In the past 12 months, how often have you had an alcoholic drink of any kind [beer, wine, cider, spirits, cocktails, premixes, liquor, homemade alcohol...]?

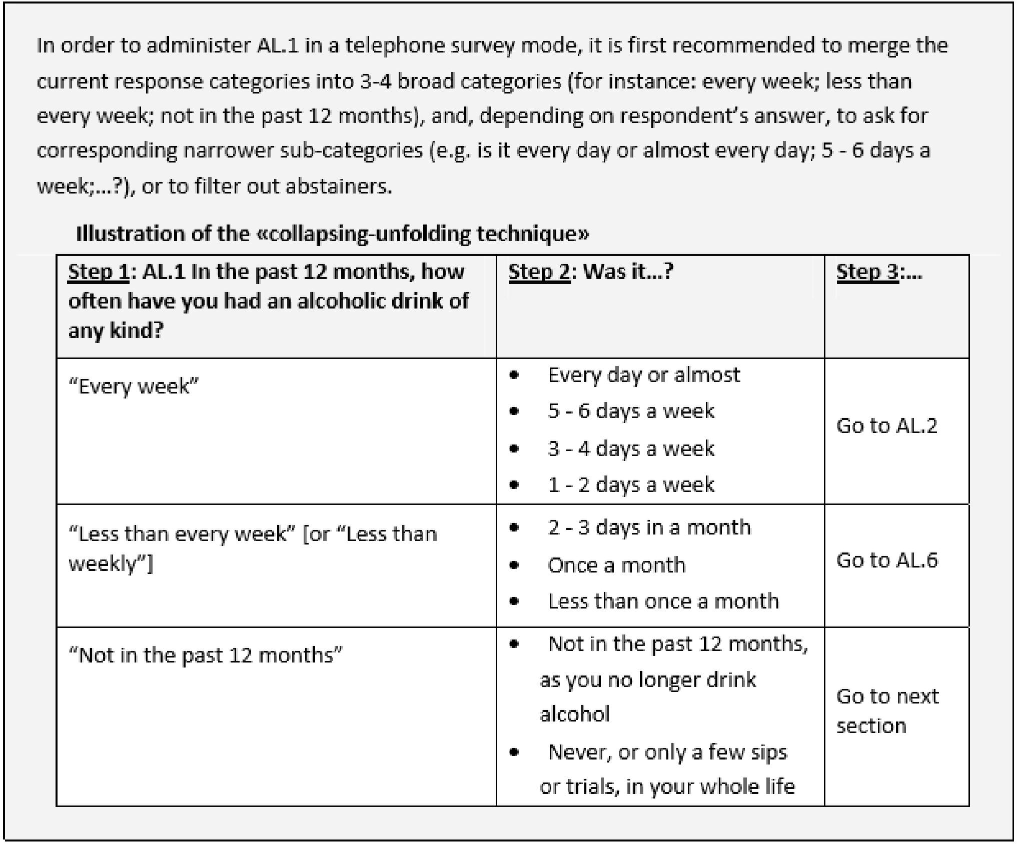
Interviewer instruction: Here, country-specific alcoholic beverages should appear in the listed examples. Home-made alcohol should also be explicitly cited.

Interviewer instruction: Hand showcard on country-specific standard drinks and containers.

1. Every day or almost
2. 5 - 6 days a week
3. 3 - 4 days a week
4. 1 - 2 days a week
5. 2 - 3 days in a month
6. Once a month
7. Less than once a month
8. Not in the past 12 months, as I no longer drink alcohol
9. Never, or only a few sips or tries, in my whole life

In the EHIS grant report with CATI recommendations, the authors recommend optimizing the question in the following way, since there is no visual memory support for the very long list of categories:

Table 78: Recommendations for the telephone survey mode...



In order to administer AL.1 in a telephone survey mode, it is first recommended to merge the current response categories into 3-4 broad categories (for instance: every week; less than every week; not in the past 12 months), and, depending on respondent's answer, to ask for corresponding narrower sub-categories (e.g. is it every day or almost every day; 5 - 6 days a week;...?), or to filter out abstainers.

Illustration of the «collapsing-unfolding technique»

Step 1: AL.1 In the past 12 months, how often have you had an alcoholic drink of any kind?	**Step 2: Was it...?**	**Step 3:...**
"Every week"	• Every day or almost • 5 - 6 days a week • 3 - 4 days a week • 1 - 2 days a week	Go to AL.2
"Less than every week" [or "Less than weekly"]	• 2 - 3 days in a month • Once a month • Less than once a month	Go to AL.6
"Not in the past 12 months"	• Not in the past 12 months, as you no longer drink alcohol • Never, or only a few sips or trials, in your whole life	Go to next section

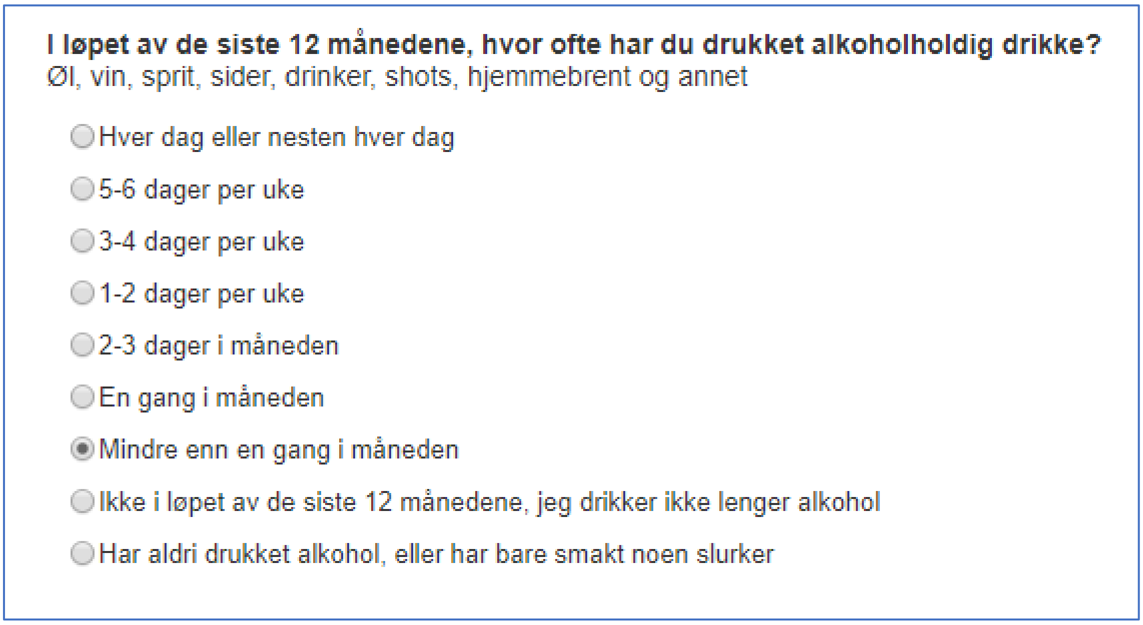
4.2.3. National implementations

The Norwegian EHIS questionnaire is a CATI only survey. It is a good example of an ESS survey that has been combined with a national survey, a factor that can influence comparability, mode choices and response burden. It is interspersed with about 150 National questions, some of which are follow-ups to the EHIS questions. Others are questions that are also found in the EU-SILC. Question AL1 has been extensively reworked and turned into a sequence to optimize it for CATI mode, but slightly different from the variant proposed in the grant report.

4.2.4. MIMOD user tests

The question on alcohol consumption was included as a question for the testing of mode specific variants in the tests done by Statistics Norway. In CATI, the question was tested with the branching version suggested by Eurostat, and in the CAWI re-test it was presented as one question with nine response options (figure 5) – the same as in the figure 4.

Figure 5. CAWI test layout of question on alcohol consumption



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

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. In the CAWI test, she answered 1-2 days per week, based on her usual consumption before pregnancy.

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. Mode effects may not necessarily be the chief source of measurement error on questions like this.

4.2.5. Key and typical questions – recommendations and suggestions for further testing

As shown above, the EHIS manuals and guidelines recommend and presupposes face-to-face modes, although some adaptations and discussions of the fitness of self-administered modes for certain questions are discussed. While the latter is commendable, the MIMOD survey shows that the reality is that self-administered modes and aural modes are widely used. These facts should be taken into consideration in the development of future EHIS manuals and guidelines.

For the long sequences of yes/no questions, mode and breakoff effects in self-administered surveys should be investigated, to see whether a reduction in numbers could positively affect response rates and response quality. Regarding sensitive questions, the general recommendation is to use self-administered modes. When using CAPI, one possibility is to let the respondent self-complete the (most) sensitive questions on the interviewer's PC/tablet. CATI of course prohibits this, unless the respondent can complete the sensitive questions in a web questionnaire or as paper self-completion. Such a multi-mode solution places higher demands on case management systems and other survey infrastructure, with risks of item nonresponse and reduced timeliness.

Based on the user tests, a mode specific design could viable for the question on alcohol consumption, with show cards in CAPI, one question in CAWI, and a branching question in CATI. However, the nine options (11 if Don't know and Refusal were included) violates the fitness criteria of a maximum of five response options formulated in MIMOD's WP5. The branching approach could be applied for mobile CAWI while keeping the one question format in PC CAWI. However, this may be considered unnecessarily complex in terms of programming and administration. A unimode branching approach could also be considered.

4.3. EU-SILC

4.3.1. Description of legal basis/guidelines/model questionnaire

The EU-SILC is a panel survey on income and various living conditions. The methodological guidelines states that five modes of data collection are possible for the survey: PAPI, CAPI, CATI, self-administered by respondent (presumably PAP) and CAWI, although the latter is not covered in the EU-SILC's legal basis. Further, priority is to be given to personal interviews (PAPI, CAPI) over the other modes of data collection. CATI has however been allowed on a "gentleman's agreement" basis for countries with person samples. The motivation given for this is that the interview *length* will be shorter as the whole household is not interviewed.

The document goes on to describe the various survey variables, and many of them have recommendations for implementation. These recommendations lack any discussions of how to implement the questions for different modes, let alone for a mixed-mode survey. Nevertheless, they presuppose interviewer-administered modes, as there are several references to response categories having to be read out loud. Therefore, the questions are intended as closed rather than as open questions with interviewer coding.

4.3.2. Mixed-mode related challenges

A key concern with the EU-SILC is the survey's length, especially when information on multiple household members is to be collected. Long surveys traditionally require interviewer involvement to obtain a high response rate. On the other hand, like in the EHIS survey, some of the questions are about sensitive issues and as such would be better suited for self-completion. These include questions on health and social and economic deprivation.

Questions about satisfaction are on 11-point scale with end-labelled categories. End-labelled questions (21) is one of Campanelli's characteristics. Campanelli et al. cite various literature that has found measurement differences, and measurement error for CATI respondents without a visual stimulus that can be offered by show cards or on-screen and forget the direction of the scale. Their recommendation is to be careful with the use of end-labelled questions. However, little research seems to have been made on 11-point scales, that will both be semantically difficult to label and distinguish, as well as difficult for the working memory to process in a CATI situation. Fully labelling these questions could do more bad than good. In a study on an 11-point end-labelled question on political affiliation on the left-right dimension, Gravem (2016) found only very small differences between a visual and a non-visual presentation of the scale in CAWI mode.

Factual questions on costs is another typical question type in the EU-SILC questionnaire. An important target variable is HH070 (monthly) Total housing cost for the household, comprised of rent, electricity, energy, insurance, mortgages etc. For many of these expenses, an average respondent would need to consult transaction records or other and make calculations and estimates. These kinds of questions would be covered as factual non-sensitive (2), with inherent difficulty due to recall (5) and as open questions requiring a number (7) in the Campanelli typology. Campanelli et al. recommends being "careful with visual layout" when including self-completion for questions requiring a number.

Asking about several precise sums that may need to be taken from external sources is a type of information request more commonly encountered in business surveys and would perhaps deserve a category of its own in the Campanelli typology. In business questionnaires, visual and calculational elements can be used to aid the respondent in understanding the larger picture and adjust sums to arrive at a reasonable total, e.g. monthly expenses.

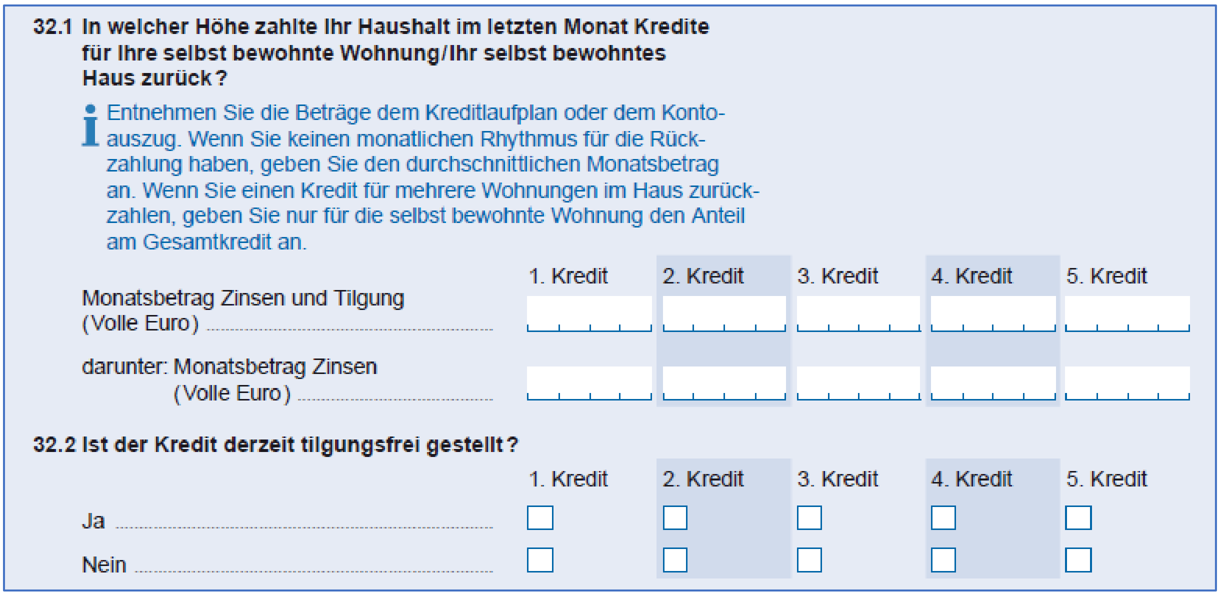
Successfully using such a strategy could mean that visual self-completion, i.e. CAWI mode would be the preferable mode. With self-completion, it would also be easier for the respondent to consult bank records, receipts or other external sources. On the other hand, properly trained interviewers could aid in both motivation, calculation and conversational interviewing to arrive at cost estimates. Theoretically, (some of) the drawbacks of each mode could be compensated by that mode's strengths, but this would require a mode-specific question design.

4.3.3. National implementations

The Norwegian EU-SILC is a CATI only survey. As it a panel survey, some key data on the household, housing and housing costs is used for dependent interviewing. Some attempts at adapting for CATI mode have been made: For the HH070 variable, filter questions on each type of expense is added, and it is possible to report costs per month, quarter or year for each of the expenses. 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 comparison, the German paper-based EU-SILC attempts to use visual stimulus to structure the reporting task of several mortgages, in figure 6. Question 32.1 which asks for monthly total payments ("Zinsen und Tilgung") and interest only ("Zinsen"). 33.2 asks whether the loan currently is without principal repayment.

Figure 6. German EU-SILC paper questionnaire questions on mortgage payments



32.1 In welcher Höhe zahlte Ihr Haushalt im letzten Monat Kredite für Ihre selbst bewohnte Wohnung/Ihr selbst bewohntes Haus zurück?

i Entnehmen Sie die Beträge dem Kreditlaufplan oder dem Kontoauszug. Wenn Sie keinen monatlichen Rhythmus für die Rückzahlung haben, geben Sie den durchschnittlichen Monatsbetrag an. Wenn Sie einen Kredit für mehrere Wohnungen im Haus zurückzahlen, geben Sie nur für die selbst bewohnte Wohnung den Anteil am Gesamtkredit an.

	1. Kredit	2. Kredit	3. Kredit	4. Kredit	5. Kredit
Monatsbetrag Zinsen und Tilgung (Volle Euro)					
darunter: Monatsbetrag Zinsen (Volle Euro)					

32.2 Ist der Kredit derzeit tilgungsfrei gestellt?

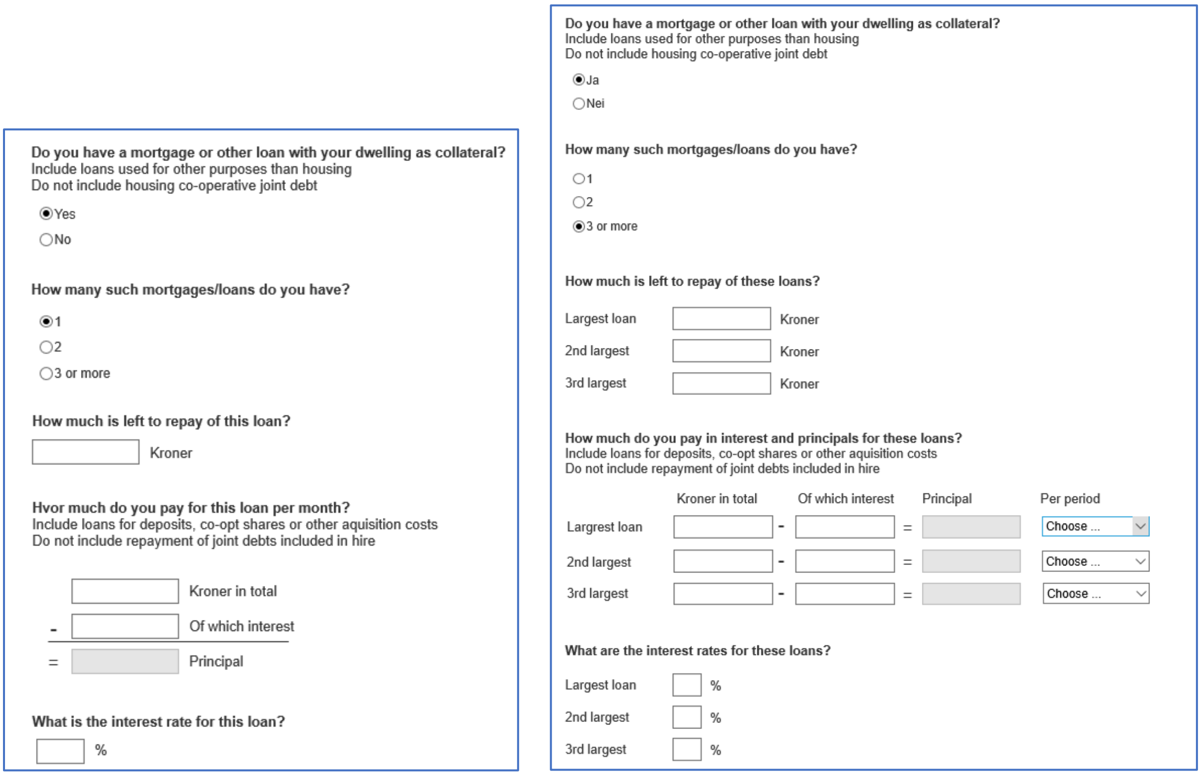
	1. Kredit	2. Kredit	3. Kredit	4. Kredit	5. Kredit
Ja	□	□	□	□	□
Nein	□	□	□	□	□

4.3.4. MIMOD user tests

In the Norwegian MIMOD user tests, we tested a visual calculator for the questions on mortgages, as a mode-specific solution using visual aids like in the German version of the question, but also by offering automatic calculations. In the CATI tests, the questions were asked as in the regular EU-SILC.

In the CAWI test version, the layout was adapted to whether the respondent had one or more mortgages, as seen in figure 7. 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 7. EU-SILC HH071 mortgage principal repayment – single and multiple mortgage versions



Do you have a mortgage or other loan with your dwelling as collateral?
Include loans used for other purposes than housing
Do not include housing co-operative joint debt

● Yes
○ No

How many such mortgages/loans do you have?

● 1
○ 2
○ 3 or more

How much is left to repay of this loan?

[ ] Kroner

Hvor much do you pay for this loan per month?
Include loans for deposits, co-opt shares or other aquisition costs
Do not include repayment of joint debts included in hire

	[ ]	Kroner in total
-	[ ]	Of which interest
=	[ ]	Principal

What is the interest rate for this loan?

[ ] %

Do you have a mortgage or other loan with your dwelling as collateral?
Include loans used for other purposes than housing
Do not include housing co-operative joint debt

● Ja
○ Nei

How many such mortgages/loans do you have?

○ 1
○ 2
● 3 or more

How much is left to repay of these loans?

Largest loan	[ ]	Kroner
2nd largest	[ ]	Kroner
3rd largest	[ ]	Kroner

How much do you pay in interest and principals for these loans?
Include loans for deposits, co-opt shares or other aquisition costs
Do not include repayment of joint debts included in hire

	Kroner in total		Of which interest		Principal	Per period
Largrest loan		-		=		Choose ...
2nd largest		-		=		Choose ...
3rd largest		-		=		Choose ...

What are the interest rates for these loans?

Largest loan	[ ]	%
2nd largest	[ ]	%
3rd largest	[ ]	%

Three test persons got the questions on mortgages, having one, two and three mortgages 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, to the right), but quickly went back to the previous question and answered “2” instead.

After reporting the size of two mortgages, se again went back to the previous question and answered “1”. When she filled in the new resulting table on monthly payments (figure 7, 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.

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.

4.3.5. Key and typical questions – recommendations and suggestions for further testing

The EU-SILC economy questions are perhaps the most interesting subjects for further testing. The “business survey” inspired approach from the user tests did not work out, but other ideas and sequences of retrieving and processing information should be tried out. Even more so than in the tests of the EHIS question on alcohol consumption, the approach tested by Statistics Norway requires a PC and is not suitable for mobile phone completion.

4.4. AES

4.4.1. Description of legal basis/guidelines/model questionnaire

The 2016 AES manual contains both a reference questionnaire and survey guidelines. The INTMETHOD variable - (interview method) of the AES has nine different values. It contains a general discussion of the pros and cons of different data collection methods and concludes by recommending interviewer administered modes, preferably CAPI, because of the complexity of the questions and need for interviewer assistance. Mixed-mode designs are very briefly discussed, with the cheaper modes of CATI or CAWI mentioned for follow-up.

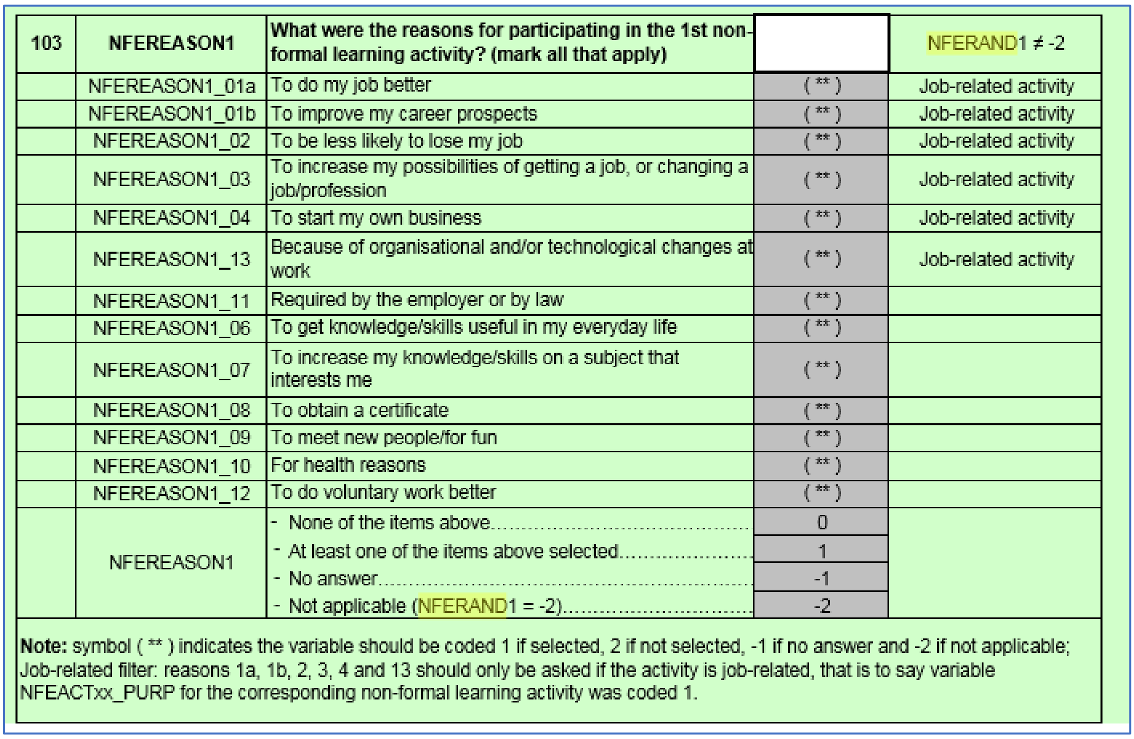
The survey guidelines contain some mode-specific recommendations for the random selection of learning activities for more detailed reporting, a key feature of the AES. The reference questionnaire does not mention different modes at all, but repeated references to the “mark all that apply” format, as well as very long questions, indicate that its authors have had a visual, self-completion questionnaire in mind rather than the recommended CAPI or CATI.

4.4.2. Mixed-mode related challenges

A. Mark all that apply, with long lists of response options (14)
B. Inherent difficulties with long and complex questions (5)
C. Inherent difficulty with long recall period (5)

A. and B. Long questions & many response options. As show cards instructions for (PAPI/CAPI) and read out instructions for CATI are left out, the latest reference questionnaire is less mode specific than the ones for previous waves (for PAPI/CAPI) have been. But this seems to be the only adaption towards mixed mode (CAPI/CATI). The questionnaire still has numerous questions with long text and very many response options with very long texts, like figure 8 shows, which requires visual aid and assistance to assure consistent comprehension in respondents. As it is, the questionnaire appears too demanding for CAWI, and not ideal for CATI, as the quality of the interviewers might affect the data.

Figure 8. AES example of mark all that apply and long list of response options

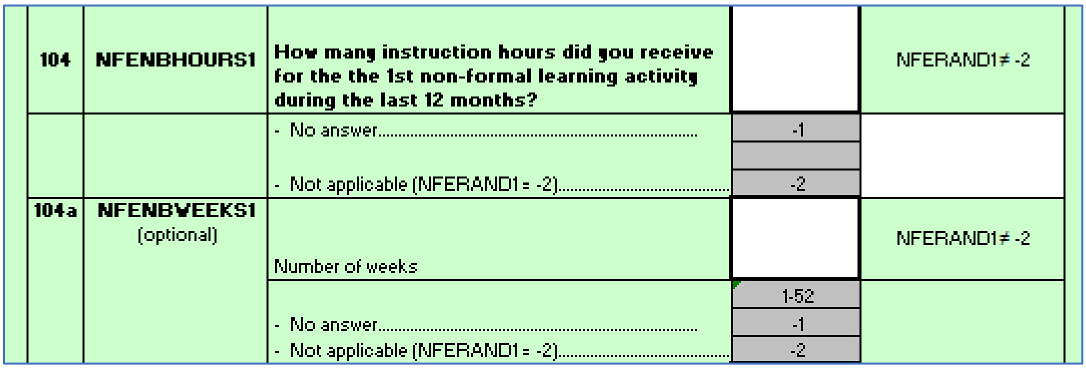


103	NFEREASON1	**What were the reasons for participating in the 1st non-formal learning activity? (mark all that apply)**		NFERAND1 ≠ -2
	NFEREASON1_01a	To do my job better	(**)	Job-related activity
	NFEREASON1_01b	To improve my career prospects	(**)	Job-related activity
	NFEREASON1_02	To be less likely to lose my job	(**)	Job-related activity
	NFEREASON1_03	To increase my possibilities of getting a job, or changing a job/profession	(**)	Job-related activity
	NFEREASON1_04	To start my own business	(**)	Job-related activity
	NFEREASON1_13	Because of organisational and/or technological changes at work	(**)	Job-related activity
	NFEREASON1_11	Required by the employer or by law	(**)	
	NFEREASON1_06	To get knowledge/skills useful in my everyday life	(**)	
	NFEREASON1_07	To increase my knowledge/skills on a subject that interests me	(**)	
	NFEREASON1_08	To obtain a certificate	(**)	
	NFEREASON1_09	To meet new people/for fun	(**)	
	NFEREASON1_10	For health reasons	(**)	
	NFEREASON1_12	To do voluntary work better	(**)	
	NFEREASON1	- None of the items above...	0	
		- At least one of the items above selected......................	1	
		- No answer...	-1	
		- Not applicable (NFERAND1 = -2).............................	-2	

Note: symbol (**) indicates the variable should be coded 1 if selected, 2 if not selected, -1 if no answer and -2 if not applicable; Job-related filter: reasons 1a, 1b, 2, 3, 4 and 13 should only be asked if the activity is job-related, that is to say variable NFEACTxx_PURP for the corresponding non-formal learning activity was coded 1.

C. Recall. Long recall periods (12 months) with questions about how hours of training (104. NFENBHOURS1) and cost of training (109. NFEPAIDVAL1) requires recollection and calculation, see figure 9. As non-formal learning activities are not always thought about in these terms, this information might not be readily available, hence these questions can be a challenge both with or without interviewer assistance, but more so without (CAWI).

Figure 9. AES example of long recall period



104	NFENBHOURS1	**How many instruction hours did you receive for the the 1st non-formal learning activity during the last 12 months?**		NFERAND1≠ -2
		- No answer...	-1	
		- Not applicable (NFERAND1 = -2)...	-2	
104a	NFENBWEEKS1 (optional)	Number of weeks		NFERAND1≠ -2
			1-52	
		- No answer...	-1	
		- Not applicable (NFERAND1 = -2)...	-2	

Some

4.4.3. National implementations

In Norway, the 2016 AES was intended as an interviewer administered survey, but changed last minute to mixed-mode CATI/CAWI due to lack of time/personnel and budget restrictions. The survey was set up as a mixed-mode CATI/CAWI survey, with embedded split sample mode sequence experiments with CATI->CAWI and CAWI->CATI as the respective sequences). The CATI and CAWI questionnaire were more or less identical. Little was done in terms of adapting the questionnaire to the two modes.

The intention was to study differences in response, net sample representativeness and possible measurement differences. Technical difficulties during data collection corrupted the controlled experiment, and the study became a concurrent – and not sequential – study, where it was possible for the respondent to answer either in CAWI or by CATI. Despite the non-experimental conditions, it is possible to look for indications of mode effects on some of the questions.

103. NFEREASON1 is an example of a question with many response categories, 13 in all (figure 8 above). It asks about the reasons for participating in a randomly selected non-formal learning activity, six of which are job related, but the job-related alternatives are only visible to employed respondents. In CAWI, it was presented as one question with 13 response options, and in CATI it was a series of yes/no questions rather than used as open questions with coding. Such a design could possibly lead to primacy effects in CAWI, and to more options being picked in CATI. Three of the thirteen categories have statistically significant differences between CATI and CAWI mode: the 7th through the 9th on the list (confusingly titled 11, 06 and 7 in the model questionnaire) – options in the middle of the list.

Table 1. NFEREASON 1 response options with statistically significant differences in CATI and CAWI

Reason for participating	CATI	CAWI
7. Required by employer or law	37.6	31.3
8. To get knowledge/skills useful in everyday life	15.9	20.2
9. To increase my knowledge/skills on a subject that interests me	23.1	33.9
N	675	741

There is no indication of any primacy effect, and the differences could be due to mode-related bias rather than instrument effects. The connection between CAWI response and a tendency to want to participate in non-formal education for everyday life and hobby type purposes could be investigated further by controlling for more demographic variables.

Question NFENBHOURS 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.

The respondents who chose to report days are then asked to report on average how many hours of training they receive per day. In the CATI interview, the respondent was then presented with a control question: "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?"

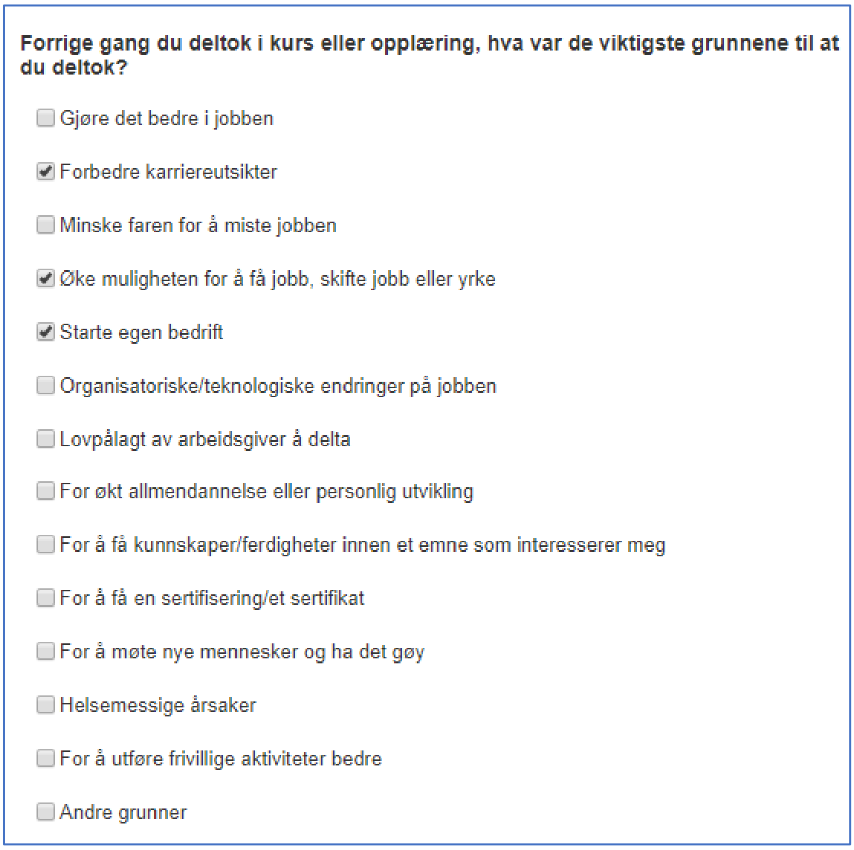
In the 2016 AES, 96% of the respondents who got the control question answered that the sum total seemed right, so there is quantitative evidence that seems to suggest that it works.

4.4.4. MIMOD user tests

Statistics Norway has previously asked the NFEREASON question as a show card question in CAPI with unclear instructions for how to handle it in CATI interviews. In 2016, it was asked as a series of yes/no questions, omitting the ranking task in both CATI and CAWI. With our test-retest design we wanted to try out the approaches with the smallest response burden while keeping the ranking task: an open question in CATI, and the "check all that apply" format in CAWI, and look at 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 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 10 – with the same options as in figure 8.

Figure 10. AES NFEREASON check all that apply question



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

□ Gjøre det bedre i jobben

☑ Forbedre karriereutsikter

□ Minske faren for å miste jobben

☑ Øke muligheten for å få jobb, skifte jobb eller yrke

☑ Starte egen bedrift

□ Organisatoriske/teknologiske endringer på jobben

□ Lovpålagt av arbeidsgiver å delta

□ For økt allmendannelse eller personlig utvikling

□ For å få kunnskaper/ferdigheter innen et emne som interesserer meg

□ For å få en sertifisering/et sertifikat

□ For å møte nye mennesker og ha det gøy

□ Helsemessige årsaker

□ For å utføre frivillige aktiviteter bedre

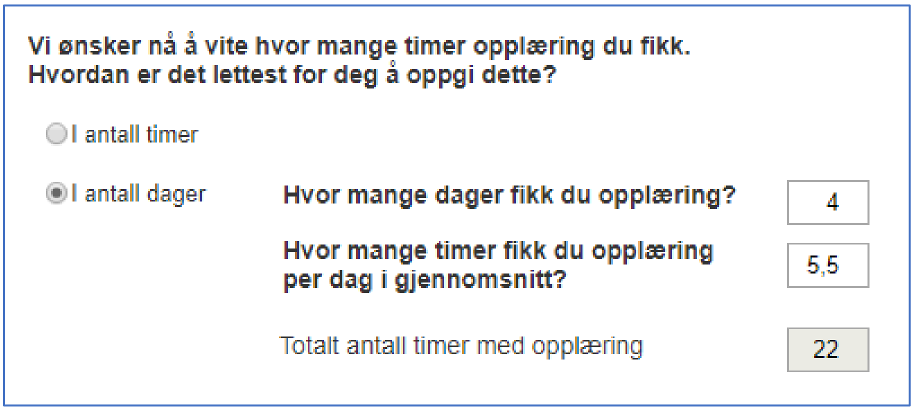
□ Andre grunner

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.

The NFENBHOURS question was also tested by Statistics Norway, but due to a programming error only the mode-specific CAWI version was tested. Here, we tried out a visual calculator for respondents who chose to report the number of days. Instead of a control question, the sum of hours was presented in a sum field as seen bottom right in figure 11.

Figure 11. 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?	4
Hvor mange timer fikk du opplæring per dag i gjennomsnitt?	5,5
Totalt antall timer med opplæring	22

Three of the test persons chose to report hours, and two chose to report the number of days. 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.

4.4.5. Key and typical questions – recommendations and suggestions for further testing

On questions like NFEREASON, the response process will be different in different modes and with different mode choices like open questions, yes/no questions or one list of options. The test situation described above 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 both the CATI and CAWI versions, 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. This would also make it and be better suited for mobile web, and remove the need for mode- or device-specific question design. The different approaches that have been used and tested by Statistics Norway illustrate the range of options and variation that can be expected when mode differences are no considered and discussed in survey documentation.

The tests of the hours calculator for NFENBHOURS demonstrate 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. A unimode approach with a control question may be a better solution.

4.5. LFS

4.5.1. *Description of legal basis/guidelines/model questionnaire*

The LFS does not have a model questionnaire, but there are minimum requirements in terms of variables that are described in a document containing explanatory notes (http://ec.europa.eu/eurostat/documents/1978984/6037342/EU-LFS-explanatory-notes-from-2017-onwards.pdf). This document also contains some recommendations on data collection procedures, and it is here that the only reference to data collection modes appears:

> *Questions have to be adapted to the method of data collection (face-to-face interviews or telephone interviews). The essential difference is that a show-card can be used to assist answering in a face-to-face interview, whereas this possibility does not exist in a telephone interview.*

Thus, CAWI mode is not mentioned at all, an indication that this part of the explanatory notes should be reviewed. As shown by the MIMOD survey and discussed by WP4 deliverable 1, all four main modes of data collection are used, though there still is a clear preference for interviewer administered modes.

4.5.2. *Mixed-mode related challenges*

The LFS was the case study for the DCSS project, the predecessor of MIMOD. The project identified several topics and questions that web respondents found difficult, chiefly problems that had to do with questions requiring clarification. Further, people who had a marginal connection to the labour market had greater difficulty responding than people in stable jobs. Recall issues regarding contractual and actual working hours were also among the key questions that were tested (see below).

Using the Campanelli criteria, the main problems that can be identified in the LFS questionnaire are related to:

A. Inherently difficult questions
B. Open questions with interviewer coding
C. Use of instructions

These findings are reflected in the MIMOD survey. Here, 20 of 31 participating NSIs found the 1st wave of the LFS not suitable for CAWI, stating that interviewers are needed to handle difficult questions, clarification and recruitment. Another questionnaire related issue where interviewers were considered paramount is protocols for identifying household members. This is not directly covered by the Campanelli criteria, although it could be considered a subset of difficult questions, it is also related to sample management and survey administration.

In the MIMOD survey, 26 NSIs nevertheless found the 2nd and later waves of the LFS to be suitable for CAWI. It must be assumed that once initial household identification and clarification issues have been handled, many of these problems can be solved. (For details, see WP4 deliverable 1.)

LFS questions on actual and contractual working hours per week have several issues. For respondents who are substitutes, do odd jobs or even work do black labour, it can be difficult to assess whether they actually have a work contract. If contractual and/or actual working hours vary from week to week, recall issues will increase the longer after the reference week the interview takes place.

4.5.3. *National implementations*

In Norway the LFS (2018) is carried out as a CATI survey. Although it has previously had CAPI and a paper questionnaire options, budget cutbacks and respondent's preferences had reduced CAPI administration drastically during the 1990s, leaving CATI as the single mode used. In the wake of DCSS user tests, a CAWI questionnaire has been developed and fielded in a mixed-mode pilot starting in June 2018. In the 1st wave, a sequential CATI > CAWI design was used for optimal recruitment. From the 2nd wave, CAWI was offered as the start mode for respondents with steady jobs, with follow-up in CATI. For the unemployed and respondents with temporary jobs, CATI was retained as the start mode, with follow-up in CAWI. This way we try to solve the problem of respondents with a marginal connection to the labour market needing more clarification and interviewer assistance, while also trying to maintain high response rates and achieving cost cuts. In later pilot waves, CAWI as the start mode will be offered to more groups of respondents.

4.5.3.1. Contractual and actual working hours – a follow-up from the DCSS project

In recent years, Statistics Finland as well as the UK's ONS have tried to use the visual and technical strengths of web mode by implementing day-by-day calendars for actual and contractual working hours.

In the DCSS project, Statistics Finland presented results from an experiment where a day-to-day calendar for actual working hours (HWACTUAL) in the reference week was presented in CAWI mode. (Pohjanpää 2014). In CATI mode the single question in "How many hours did you work during the reference week" was retained. A CAWI control group was also asked the single question. Although not statistically significant, the CAWI calendar question estimates were closer to the CATI estimates. than the CAWI single question. For the CATI question, it is possible that the interviewers probe and engage in a conversation trying to work out the actual working hours, but the eventual extent and nature of such practices is unknown.

The ONS has user tested tried a similar approach for the HWUSUAL as well as HWACTUAL variables, but with less positive results. (Nolan 2018) The test persons found the task of providing hours worked unnecessary, especially as most of them worked standard hours or were aware of the sum total of hours worked after each week. The ONS therefore decided to ask only for the number of hours per week.

In the all CATI Norwegian LFS, the current approach is to first ask about contractual (usual) working hours for the reference week. This is followed by questions on absence during the reference week, and then (unless the respondent was absent all week) by questions on extra hours. From this information, tentative actual working hours are calculated, but not presented to the respondent. The respondent is then asked, "how many hours did you work during the reference week?". If the response to this question is more than xx hours, a consistency check is triggered, and the CATI interviewer is instructed to review the previous answers. As the previous responses are overwritten and no paradata (process data from the questionnaire completion) is kept, we do not the percentage of respondents that are exposed to the consistency check. We do however know the number of cases where the interviewer and respondent were unable to work out the errors, as it is possible to continue without the numbers adding up. For the 1st quarter of 2018, this applied to 98 cases, or 0.5% of the net sample.

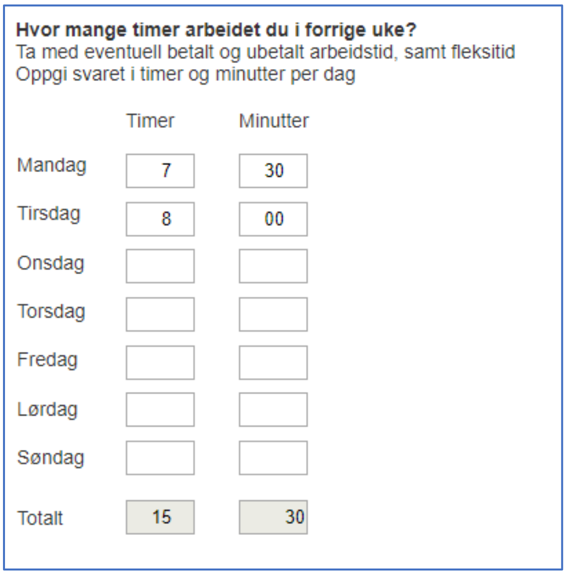
The original idea behind this approach was to aid the respondent's memory of the week and prevent satisficing. In a web questionnaire, it is less feasible to add a consistency check that requires the respondent to go back and respond to questions once more. For the LFS pilot currently running, we therefore have substituted this with a question on which number of hours is more correct: the calculated hours, the hours from the single question, or another number of hours.

4.5.4. MIMOD user tests

Because of the different results from the research done by Statistics Finland and the ONS with a day-by-day hours calculator for the HWACTUAL variable, it was decided to do qualitative tests of this approach in a Norwegian contest as well, as country-specific

The CATI version was a single question: “How many hours did you work during last week?” The CAWI version is shown in figure 12. 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 12. 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 fleksitid
Oppgi svaret i timer og minutter per dag

	Timer	Minutter
Mandag	7	30
Tirsdag	8	00
Onsdag		
Torsdag		
Fredag		
Lørdag		
Søndag		
Totalt	15	30

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.

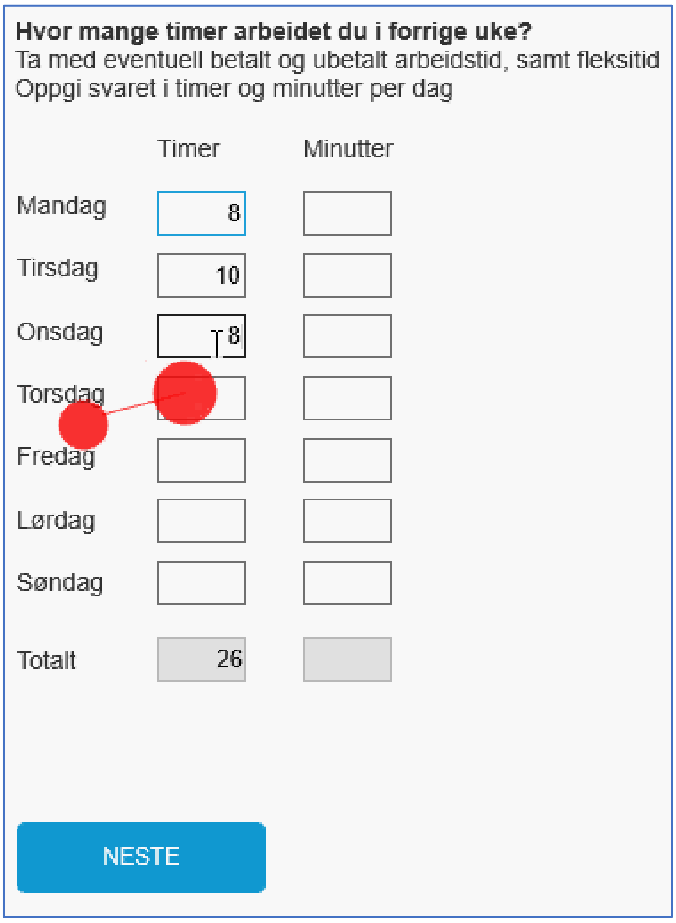
In the CAWI retest, the two test persons who had responded “37 and a half hour” in CATI both started filling in 7 hours and 30 minutes for each day of the week. One of them did so for Saturday and Sunday as well, discovered in the sum field that the total was 52 hours and 30 minutes, and removed the 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 13).

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



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	8	
Tirsdag	10	
Onsdag	8	
Torsdag		
Fredag		
Lørdag		
Søndag		
Totalt	26	

4.5.5. *Key and typical questions – recommendations and suggestions for further testing*

The MIMOD user test results are more in line with the findings at the ONS than Statistics Finland, and also in line with the results from the mortgage calculator tests for the EU-SILC questionnaire: using the expected advantages of a visual mode in this way required more of the respondent, and presupposes a familiarity with calculation setups.

The recommendations on other potentially problematic LFS variables in CAWI that were summed up by the DCSS project generally are still valid: the lack of interviewer support for explanation and identification of labour market status is still a main issue, something which is reflected in the modes used by ESS countries.

However, as a multi-mode panel survey with interviewer-administered modes used as main modes in the first wave, CAWI may have its place in later waves – for either the whole sample, or for respondents who are less likely to experience problems or cause measurement errors in CAWI. As the MIMOD survey indicates, this is a data collection strategy that can be expected to become more widespread in not too many years. Pilots like the one currently conducted by Statistics Norway will hopefully shed more light on this issue, and it should be addressed in the relevant revised LFS regulations that are expected.

5. Discussion and conclusion

As discussed in this and previous WP4 deliverables, the modes that Eurostat recommends for the different ESS surveys is often in conflict with the content of the survey, but also with the way they are actually conducted in the various ESS countries. Some of the recommendations will therefore conflict with Eurostat's pre-existing recommendations and regulation contents. Moreover, an individual survey can also contain some questions that will work better in one modes, and others that will work better in another mode.

The main recommendations should therefore perhaps be directed at Eurostat rather than the individual countries: changes should be made to model questionnaires, guidelines and other documentation to better facilitate mixed-mode questionnaire development and data collection. This could include shortening questionnaires by modularizing them, by limiting the number of items in grids questions, removing all non-essential questions, and considering how each question will work in different modes using the Campanelli or other sets of criteria.

The ONS in the UK has recently redesigned their LFS survey with the aim of making it smartphone friendly. Rather than trying to adapt their existing patchwork-like LFS to make it better suited, they saw it as an opportunity to start from scratch and Eurostat's output requirements, rebuilding questionnaire flow, content and wording, as well as visual design (Nolan 2018). This process should ideally be conducted at both Eurostat and national levels. To paraphrase the quote from Don Dillman on the first page of this report, *All ESS survey model questionnaires should be designed with mixed-mode in mind*.

The results from Statistics Norway's user tests of unimode and mode-specific approaches point in the direction of a unimode approach being safer than a mode-specific approach for the questions that were tested. Especially the more ambitious attempts at designing mode specific for PC web using calculators and an accounting style setup had issues. The PC versus mobile issue is another factor that limits the recommendations in this deliverable. Most of the tests of mode-specific CAWI solutions would only be feasible in on a PC, and not on a mobile phone screen.

The findings and recommendations presented in this deliverable may have a short expiry date. Many initiatives are running in parallel in the European Statistical System, and we have not been able to include as much of it as we would have wanted. Web data collection best practices are constantly developing. Innovative uses of smartphone features for survey communication may also mean new opportunities, and experiments with using chatbots in CAWI questionnaires are e.g. currently being planned at Statistics Norway. It is also not written in stone that we will be locked to the mobile format forever.

A good and innovative way to use, expand upon and revise the results could be the establishment of a wiki-type web resource dedicated both to general topics of mixed mode and web data collection for official statistics, but also for the specific surveys. With a user-generated repository of examples, test results and discussions, the hope is that it would better enable and encourage contributions from all the NSIs in the ESS in a way that a traditional ESSnet grant cannot.

In addition to this comes the different conditions and resources that influence how data collection is conducted in practice in each country, as demonstrated in previous WP4 deliverables. This can limit the applicability and usefulness of the recommendations presented in this report, but an online forum such needs can be voiced and hopefully addressed by other NSIs who are or have been in similar situations.

The establishment of such a forum is therefore the last suggestion and recommendation to be formulated by MIMOD WP4.

References and literature

2016 AES manual – Version 3, European Commission/Eurostat Unit F3 (February 2017)

Blanke, Karen (ed.) (2015), Data collection for social surveys using multiple modes final report. Revised version January 29 2015. European Commission/Eurostat.

Campanelli, P. et al. (2013), A Classification of Question Characteristics Relevant to Measurement Error and Consequently Important for Mixed Mode Questionnaire Design, Presented 2011 at the Royal Statistics Society, London, UK

Community Survey on ICT Usage in Households and by Individuals – 2018 Model Questionnaire version 1.4 (2018)

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008), International handbook
of survey methodology. NY: Lawrence Erlbaum Associates.

De Leeuw, E. D., Hox, J. J., De Leeuw and Hox (2015), Survey Mode and Mode Effects, in Engel, U. (2015) pp. 22-34.

Dillman, D. A., J. D. Smyth and L. M. Christian (2009): Internet, Phone, Mail and Mixed-Mode Surveys – The Tailored Design Method, John Wiley & Sons, New York.

EU Labour for survey – Explanatory notes (to be applied from 2017Q1 onwards) European Commission/Eurostat Unit F3 (2018)

European Health Interview Survey (EHIS wave 3) Methodological manual (2018 edition), Eurostat manuals and guidelines, Luxembourg.

Engel, U. et al. (eds.) (2015), Improving Survey Methods – Lessons from Recent Research, Taylor & Francis, London.

Finger, J. et al. (2011), Improvement of the European Health Interview Survey (EHIS) modules on alcohol consumption, physical activity and mental health: Final report, European Commission, Berlin.

Gravem, Dag F. (2018), Adapting ESS survey questionnaires to mixed-mode data collection, paper presented at the Q2018 conference, Krakow, Poland.

Gravem, Dag F. (2016), Measuring political affiliation on an 11-point metaphor, paper presented at the QDET2 conference, Miami, Florida.

Körner, T. (2014), Report on the definition, identification and analysis of mode effects. Deliverable for Work Package III of the ESSnet on Data Collection for Social Surveys Using Multiple Modes (DCSS), Federal Statistical Office Germany, Wiesbaden.

Luzi, O. and B. Buelens (2018), Dealing with mode effects, paper presented at the Q2018 conference, Krakow, Poland.

Methodological guidelines and descriptions of EU-SILC target variables – 2018 operation (Version August 2017), European Commission.

Nolan, Alex (2018), Respondent Centred Survey Design at ONS, presentation at the 2018 Quest workshop, Wiesbaden, Germany.

Pohjanpää, Kirsti (2014), Does mixed-mode data have influence to the quality of data of LFS? Paper presented at the Q2014 conference, Vienna, Austria

Appendix A. Campanelli typology and recommended modes

SC=Self-completion, regardless of paper or web

Question content		CAPI	CATI	SC
1	Sensitive questions			X
2	Factual, non-sensitive	X	X	X
3	Subjective, non-sensitive	X	(X)	X
4	Subjective, non-sensitive scalar question	X		X
5	Inherent difficulty due to concepts, comprehension and recall issues	X	(X)	(X)
Type of task – open questions		CAPI	CATI	SC
6	Unconstrained open question	X		X
7	Open question requiring a number	X	X	X*
8	Open question requiring a date	X	X	X*
9	Open question	X	X	X*
10	Open question with interviewer coding	X	X	
Type of task – closed questions		CAPI	CATI	SC
11	Agree-disagree scales	-	-	-
12	Unipolar and bipolar rating scales	X	(X)	X
13	Numeric bands	X	X	X
14	Mark all that apply	(X)		(X)
15	Yes/No for each	X	X	X
16	Ranking	(X)		(X)
17	Battery of ranking questions	X	X	X
18	Visual analogue scale			X**
Characteristic of task		CAPI	CATI	SC
19	Use of middle categories	X	(X)	
20	Number of response categories	X	(X)	X
21	End-labelled scalar questions	X	(X)	X
22	Branching	X	X	X
Implementation of task		CAPI	CATI	SC
23	Use of instructions, probes, clarification etc.	X	X	X
24	Edit checks	X	X	X***
25	Spontaneous "Don't know"	X	X	(X)
26	Size of answer box	X	X	X*
27	Formatting of response box	X	X	X*
28	Formatting of response list for closed questions	X	X	X
29	Showcards for long lists of response options	X		X

X = Recommended

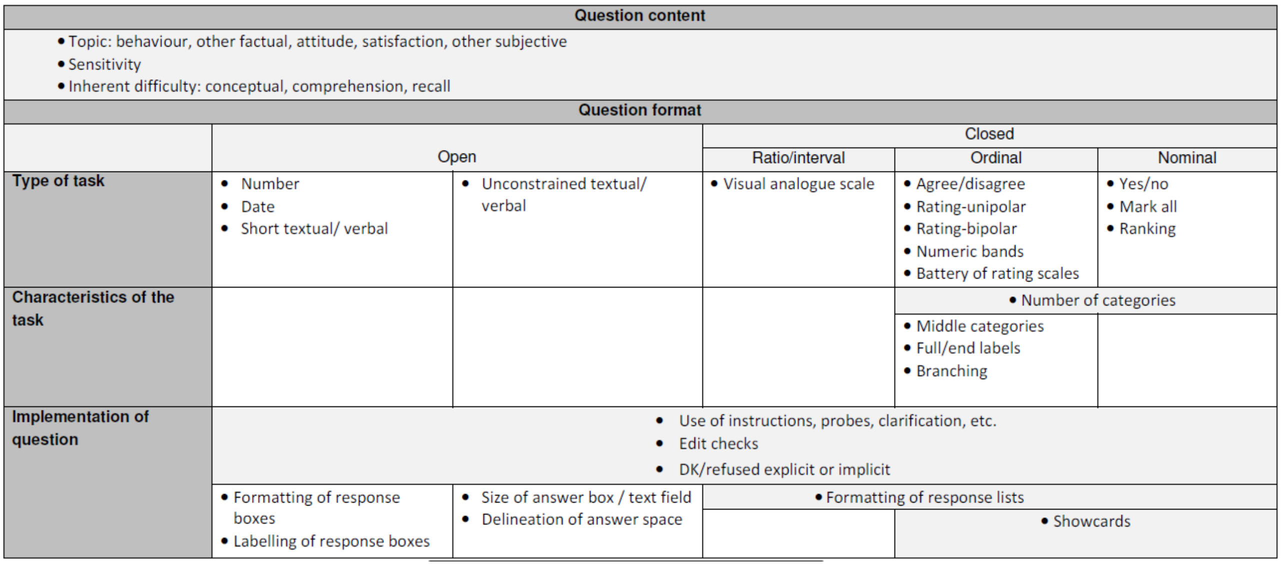
(X) = Possibly recommended

* "Be careful of visual layout"

** "Avoid web"

*** Excluding paper self-completion

Campanelli characteristics broken down on content formats, tasks, and implementation



Question content					
• Topic: behaviour, other factual, attitude, satisfaction, other subjective • Sensitivity • Inherent difficulty: conceptual, comprehension, recall					
Question format					
	Open		Closed		
			Ratio/interval	Ordinal	Nominal
Type of task	• Number • Date • Short textual/ verbal	• Unconstrained textual/ verbal	• Visual analogue scale	• Agree/disagree • Rating-unipolar • Rating-bipolar • Numeric bands • Battery of rating scales	• Yes/no • Mark all • Ranking
Characteristics of the task				• Number of categories	
				• Middle categories • Full/end labels • Branching	
Implementation of question	• Use of instructions, probes, clarification, etc. • Edit checks • DK/refused explicit or implicit				
	• Formatting of response boxes • Labelling of response boxes	• Size of answer box / text field • Delineation of answer space	• Formatting of response lists		
				• Showcards	

Full document with detailed explanations (2011 version) is publicly available here:

http://www.websm.org/uploadi/editor/1364221156Campanelli_et_al_2011_A_classification_of_question_characteristics_relevant_to_measurement_error.pdf

(Appendix B is available as a separate document)