European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

# Moving towards Web Data Collection in Household Surveys

Pertti Taskinen[1], Laura Hulkko[2]

[1] *Statistics Finland, Helsinki, Finland; pertti.taskinen@stat.fi*
[2] *Statistics Finland, Helsinki, Finland; laura.hulkko@stat.fi*

**Abstract**
Today, the data collection relies strongly on the interviewer. To be realistic, CATI mode will remain in household surveys but a partial transition to the electronic mode will be welcomed as a part of the digital development of the society. The EU-LFS (Labour Force Survey), a rather complicated survey, has been set by Statistics Finland as one of the first household survey carried also by CAWI mode. Before that is possible, many preparations has to be done, e.g. building infrastructure and software for a new questionnaire, testing the design and analyzing the test results. In this presentation the transition to mixed mode data collection in Statistics Finland is enlightened. As an example survey case the EU-LFS shows that the transition is challenging due the sensitivity of the results and the tight schedule of the data collection.

**Keywords:** EU-LFS, CAWI, mixed-mode data collection

## 1.1 Household surveys at Statistics Finland

Statistics Finland produces all the EU-regulated household surveys. Labour Force Survey and Consumer Survey are produced and published monthly. Annual surveys include SILC and the Survey on Use of Information and Communications Technology (ICT). The data collection in these surveys has so far been done mainly by telephone interviews (CATI) and supplemented with some personal interviews (CAPI). Time Use Survey, Household Budget Survey, Adult Education Survey and Leisure Survey are produced with longer intervals (every 5–10 years), and mainly using CAPI interviews.

All household surveys are based on the probability sample on individuals. They are all voluntary for the respondent, and generally no substitutes are allowed.

### 1.2 Including web as a data collection method in social surveys

The use of internet is already very common in Finland. According to the Survey on use of ICT in 2015, 93 percent of the population aged 16–74 have used the internet during past 3 months, and 85 percent use it daily. Internet use is obviously most common in younger age groups; practically all people under the age of 45 use it, and most of them daily. But in recent years internet use has also become more and more common in older age groups. When in 2010 only 43 percent of the age group 65 – 74 had used the internet during past 3 months, in 2015 it was 69 percent.

Not only do people use the internet, they also start to expect and demand internet based services. Considering this development, Statistics Finland has set a strategic goal to include web as a data collection method in all household surveys in 2018. At the same time, the pressure to reduce data collection costs is leading to considerable reduction of CAPI interviews. This means that the data collection process and methods as well as the questionnaires have to be revised in all social surveys. This is a challenging goal which requires considerable investments in developmental work during several years.

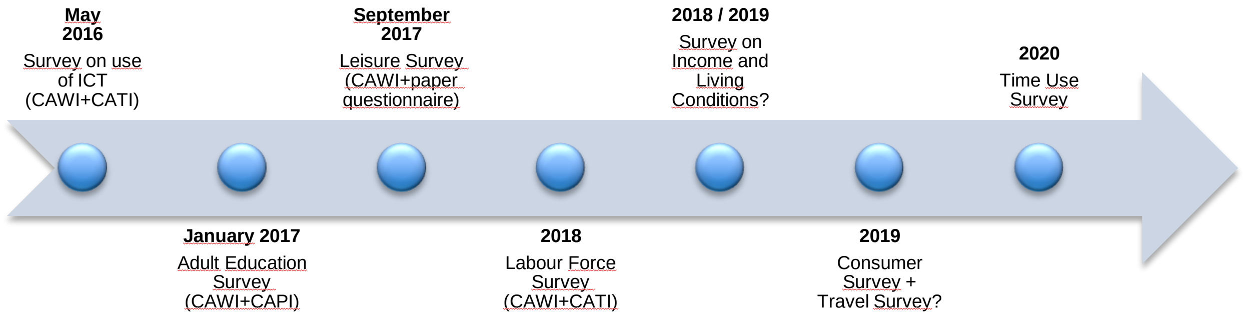
All surveys will in the following years be conducted using a mixed mode approach. The design of the data collection model and the modes depend on the survey. For some shorter surveys like the LFS and the Survey on the use of ICT combining web (CAWI) data collection with CATI interviews is the best choice, but for others it may be better to include other modes like paper questionnaires. For each survey we also need to consider whether web should be offered to all respondents equally or if some groups (e.g. elderly people) should rather be interviewed.

In order to be able to efficiently conduct all these surveys using multiple modes we first have to build a data collection management system which supports mixed mode surveys. This work has started in the end of 2014, and the new system will be piloted in September 2016. The challenge has been to build a system which is flexible enough to be able to manage all surveys with different schedules and modes of data collection as well as allow all data collection tools used now and in the future. The aim at Statistics Finland has been to use as much automatization as possible, but at the same time the system also has to allow manual operations. For example, when a new survey is established, the data collection model (e.g. modes and schedules) are defined and entered into the management system. After this, the system automatically makes the required transfers between different modes during the data collection period as well as manages the work division between interviewers. However, when changes occur or data collection does not progress according to plans, we need to be able to react quickly with manual transfers.

At the same time as the new data collection system has been built, a lot of work has been done with the questionnaires. Transforming an interview questionnaire into a user friendly web questionnaire without creating problems with comparability is not an easy task. All questions are not applicable as such; for example long lists of alternatives simply do not work in the web mode, and difficult concepts have to be simplified so that the respondent can understand them fully without the help of an interviewer. On the other hand, even small changes – like whether or not to offer a “cannot say” category – can create a problem with comparability between modes. All web questionnaires have been developed with the support of the Statistics Finland Survey Laboratory, and they are all pre-tested using cognitive interviews. The aim of the testing is to observe the answering process and find out possible problems with technical usability of the questionnaire as well as understanding the concepts and the questions. Careful questionnaire design and testing is considered to be a crucial step in the implementation of mixed mode at Statistics Finland – it is seen as a significant way to reduce mode effect.

The current schedule plans for implementing mixed mode at Statistics Finland are described in figure 1. The following section describes the situation as well as the main challenges in different surveys. Labour Force Survey will be covered more specifically in the next chapter.

**Figure 1. Planned implementation schedule of mixed mode data collection in household surveys at Statistics Finland**


May 2016
Survey on use of ICT (CAWI+CATI)
January 2017
Adult Education Survey (CAWI+CAPI)
September 2017
Leisure Survey (CAWI+paper questionnaire)
2018
Labour Force Survey (CAWI+CATI)
2018 / 2019
Survey on Income and Living Conditions?
2019
Consumer Survey + Travel Survey?
2020
Time Use Survey

The first survey to pilot CAWI data collection at Statistics Finland was the Consumer Survey. Two pilot tests have been done, the first preliminary one in 2011 and a mixed mode pilot survey (including Travel survey) in 2012. The results were somewhat conflicting. On the other hand the questionnaire was fairly easy to adapt into CAWI mode, but the risk of mode effect was found considerable. This is mostly due to the nature of the questions – the Consumer Survey inquires peoples' opinions rather than facts. The fact based travel survey (which was piloted in the same survey) showed practically no sign of mode effect. Implementing mixed mode data collection in the Consumer Survey could possibly mean a break in the time series, which needs to be taken into account. At the moment the plan is to take mixed mode into production after 2018.

The first household survey to use mixed mode in the actual production will be the Survey on Use of ICT. The data for year 2016 is collected using a CAWI-CATI approach in May and June 2016. According to a pilot test done in 2015 the biggest risks in this survey have to do with the difficult and technical concepts. In the CAWI mode it is vital to use concepts that are clear to the respondents, and extra instructions on the screen or behind a help button have been found less useful in cognitive testing. It is also somewhat problematic to collect data on the use of internet using a web-based tool. This creates a risk of bias, so it is necessary to supplement the CAWI mode with adequate telephone interviews as well as review the estimation process in order to reduce this bias.

The next mixed mode survey will be Adult Education Survey, which will use a combination of CAWI and CAPI modes in January 2017. This survey includes a part which is EU harmonised as well as questions for national use only. The national part makes the questionnaire considerably longer, which is a problem in the CAWI mode. This is why only the EU harmonised part of the questionnaire will included in the CAWI questionnaire, and the longer questionnaire is used in the CAPI data collection, which is made possible by some outside funding from data users.

In some surveys supplementing CAWI with telephone interviews is not the ideal option. One example of these is the Leisure Survey, where the questionnaire is simply too long to go through over the telephone (CAPI interviews have previously lasted up to 90 minutes). However, the questions – which mostly have to do with hobbies and participation in cultural activities – are fairly easy for the respondent and quick to answer, so supplementing CAWI with a self-filled paper questionnaire is a possible option. The work on these questionnaires is at the moment ongoing and the approach will be tested in a pilot study in the autumn of 2016.

The Labour Force Survey is the biggest household survey due to its large annual sample. This is why developing a mixed mode approach for the LFS has been defined as a priority at Statistics Finland. In this survey the questions are based on facts and the questionnaire is short, so the biggest challenges have to do with the extremely tight schedule and the panel design. These are described more closely in the following chapter.

## 2. Developing the Finnish LFS

### *2.1. LFS as a short telephone survey*

At the moment, the vast majority of the LFS interviews in Finland are collected by phone when only residuals are concluded by CAPI (1.2 % in 2015). However, personal interview has held its position at the first wave as an option, or more precisely, as a “deterrent”. It means that an interviewer may send a message to the respondent that if you are not able to answer by telephone the interviewer will visit you.

The LFS telephone interview in Finland appears to be one of the shortest among peers: variables are collected average in 7 minutes per interview. However, one hour interviewer working time produces only 1.2 interviews due to complexity of contacting, or “hunting” people. So, despite the efficiency of the time use in an interview, data collection by the telephone is not so cheap at all because the actual interviewing time is only a small part of the whole work of interviewers.

Looking at the cost of the massive data collection - over 100,000 interviews per year in the LFS - someone could hope that data will be received from already existing data repositories. However, in many statistics, data will only be obtained only by asking some questions from a sampled person. The LFS is based on a specific reference week, or two weeks, or four weeks, which are targeted for each person in a sample for every week of the year.

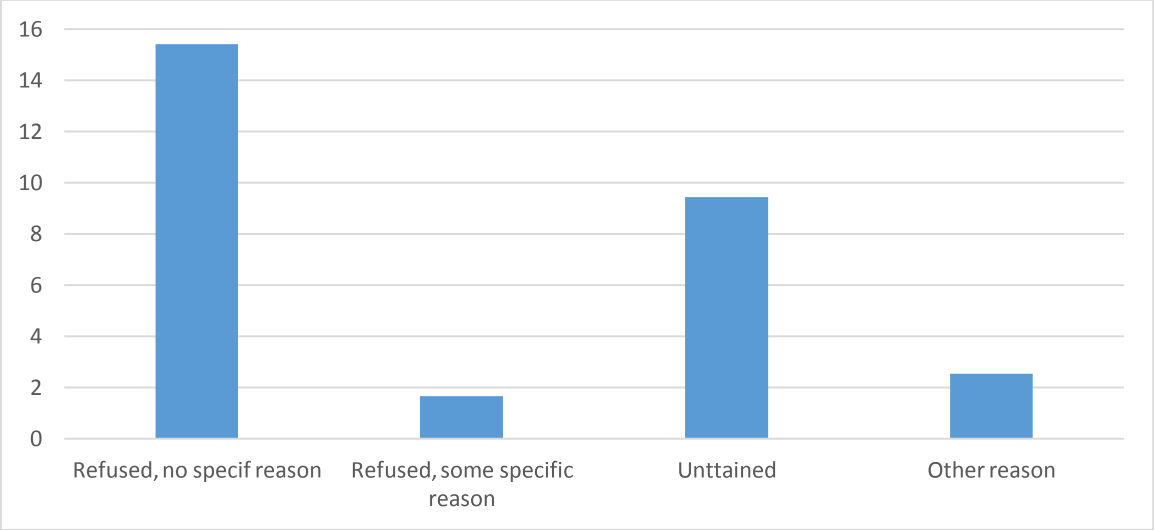
Despite the apparent easy of the telephone interview surveys are today less popular than at the time when CATI system became common. If we look at the LFS in Finland, the response rate has declined steadily (table 1).

**Table 1. The response rate in the Finnish LFS, biannually 1997-2015, from net sample.**

| 1997 | 1999 | 2001 | 2003 | 2005 | 2007 | 2009 | 2011 | 2013 | 2015 |
|---|---|---|---|---|---|---|---|---|---|
| 89,8 | 86,4 | 86,5 | 85,4 | 83,8 | 80,8 | 79,9 | 76,1 | 72,9 | 70,8 |

It is plausible that CAWI would produce low response rate despite the importance of the topic (employment and unemployment for instance). CAWI as a method has a lack of the important motivational aspect which is an interviewer. In 2013, the LFS web pilot ended up in 30 percent response rate even though it benefitted a longer data collection period than normal and also so called motivational contact calls were made by interviewers. The latter may be as expensive as interviewing itself and thus not useful.

**Figure 2. Reason for non response, share from the net sample, Finnish LFS 2015.**



In 2015, the reason not to take part in the LFS was mainly refusal, as seen in a figure 2. In most cases any particular reason for refusal was not given by respondents: according to that, the rejection wouldn't be changed if another mean (e.g. CAWI) would have been introduced for them. Eventually, nearly 10 percent remained unattainable – for those the CAWI mode informed by a letter would be a possible solution. Anyway, it is not appropriate to build a mixed mode infrastructure for a tiny part of the sample.

### *2.2. Planning data collection design*

Each of the Finnish LFS sample unit undergoes five waves during 15 months. This model has not been seen possible to test as such. The final test for the mixed mode collection in the LFS will be done as a three round pilot. Three test rounds will obviously give fairly good information on the process and the decent data as well.

The aim is that for every sample unit the web option is offered first. The time frame is tight: the plan is that an advance letter will be delivered on Friday and the web opens up on the following Monday. Respondents have got only four days to use web; on Friday, the 5th day of the data collection, interviewers will start their work and the web questionnaire will remain closed since then. Respondents are informed utmost clear on this arrangement in the letter and also with text messages if the number is known.

This type of mixed mode system is called sequential design, which has disadvantages (e.g. short period for web mode) but comparing to the concurrent design it is clear. In the LFS, it is simply not possible to give respondent longer period to choose the mode because "skirting" between modes is an inevitable thread.

In the end of the first interview round (CAWI and CATI), email address is asked and also the phone number checked. At the next rounds, the advance letter is not send but text messages and emails are delivered to announce the opening of the web questionnaire. The same sequential mode is used again at the later rounds.

The process is not entirely digital. Paper post is still needed and in fact an advance letter is the way to get a person initially informed that he is a sampled person and secondly, to enter to the web questionnaire.

### *2.3. Building mixed mode questionnaire*

Not only the infrastructure and the collection design but also the questionnaire plays an important role in combined collection. The CATI questionnaire in the Finnish LFS has been used from 1997 and a unimode web questionnaire was piloted at first in 2013. In a mixed mode collection two sides of the questionnaire – web and telephone - should be seen as a unity. Therefore, a new project was needed to transfer both sides together to the newer software platform as well as to design the questionnaire and to test it, too.

The combined questionnaire was created first in Excel form where web and telephone questions (as like categories and instructions) are next to each other to catch a view on every detail on the both sides of the questionnaire. Many questions are exactly the same in CAWI and CATI but taken account that web and telephone are based on different senses (sight and hearing) there are some light applications to them at the questionnaire.

The transfer to the newer software mode has been done thoroughly. The new LFS web questionnaire was tested twice in the cognitive laboratory. Two main issues raised. First, the majority of respondents didn’t notice the help button which contain some useful instructions for many items. Second, the panel feature of the LFS with dependent interview was tested and it resulted that technical reliability and the appropriate use of the previous round data was found as a big challenge among the questionnaire developers.

It is important to notice that devices for the web use are many kinds. The questionnaire project underlined not to restrict the use of mobile devices when filling the web questionnaire. The new software and its applications should be in service with mobile phones and other portable devices.

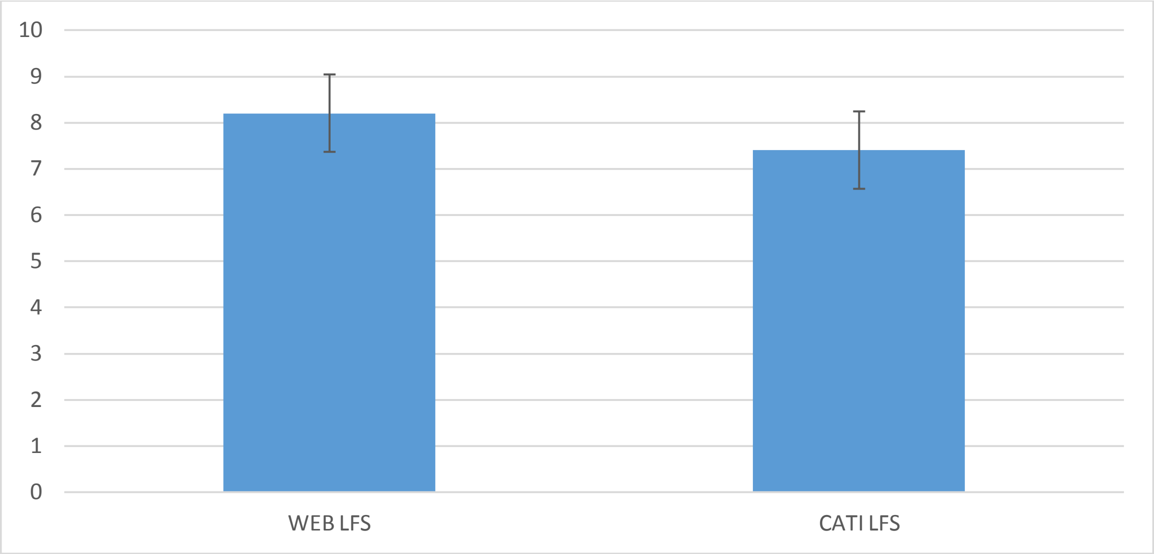
## 3. Challenges with results and respondents will continue

### *3.1. Unemployment as a target variable*

In this chapter, a very short insight to the challenges of the mixed mode data collection is given, which is, possible effects on the results of the additional mode.

The Finnish LFS web pilot in 2013 results were compared with the LFS results based on the concurrent “official” LFS data. The estimates of the unemployment rate in October 2013 can be seen in figure X. The figures are not seasonally adjusted. According to the LFS data the unemployment rate was 7.4 and corresponding web pilot data provided 8.2. The difference is not statistically significant.

Figure X. The unemployment rate estimate and its 95 % confidence interval according the LFS web pilot and the LFS in October 2013.



The web data is based on 2,366 respondents and CATI-data on 3,299, respectively. As seen in the figure, the width of the confidence interval in unemployment rate underlines the complication of measurement of the unemployment rate: if the sample size is not big enough results can be more or less vague due to the confidence interval. The number of unemployed against the population is often small and in addition to that, the labour market data is needed every month; so a huge yearly sample size is needed and that cannot be imitated in practice in any pilot. On the other hand, when disseminating the unemployment rate, already a little change in estimate is remarkable.

The definition of unemployment consist on many questions. One of them is that did a person seek a job during the previous four week. If the answer is yes, it must be clarified that was a person available to work. We found significant difference (95 % level) between modes in this question though the amount of respondents was restricted. In the web pilot, the estimate of the variable “seeking work but not available” was 75,000 and the corresponding estimate by CATI mode was 21,000. The wording was exactly the same on the both questionnaire. It is unclear what factor leads to the different results and thus we need further testing.

We can be assumed that the mode-specific non response produces slightly different results by different mode. That is why the LFS mixed-mode pilot data which will be achieved 2016-2017 by combined telephone and web collection will be analyzed as total and will be compared with the data which will be obtained at the same time with unimode CATI.

### *3.2. Various means for various people*

The household survey data collection in the official will not be the same in future than it is at present. The LFS data will be obtained only if means are suitable for people. For many of them, filling the questionnaire will happen by punching a mobile device in a café or in a public transport or sometimes at home when watching television.

In the outlook, many of us are alarmed the quality of the web data when an interviewer is not contacting the interview. The situation strongly underlines the necessity of simple questionnaire, so to say, the importance of the questionnaire design. However, comparing the future web mode to the age of the telephone interviewing, which we are still living, the difference may not be so huge. Also today, people are hasty and are not often carefully listening to all what interviewers are saying.

The upcoming CAWI is useful in the sense that it creates a new feedback platform and is a thus link between respondents and experts in the statistical office. In the pilot in 2013 we got interesting clues which subject at the questionnaire are easy and on the opposite, which are difficult for different type of respondents.

It is obvious that some people still like more a personal contact and that will be arranged by telephone interviewing. Also many or even more people will not like to take part at all in the survey. For those a telephone contact is continuously necessary for the purpose of the persuasion, at least as long as telephone numbers are used and number files are available. However, the telephone or smart phone will not be an everlasting device and the forthcoming communication form between an informant and statistical office is unknown; what is coming next is beyond the paper.

## References

Oinonen S. (2014), Liite 5. Työvoimatutkimuksen web-tiedonkeruun vastauasteista ja vastaajista. Unpublished paper.