WP 3 Summary Report

Annemieke Luiten, Trond Båshus, Bart Buelens, Dag F. Gravem, Thomas Körner, Bengt Lagerstrøm, Salah Merad, Kirsti Pohjanpää, and Barry Schouten

Deliverable for the ESSnet Project “Data Collection for Social Surveys using Multiple Modes” (DCSS)

- revised version 29 January 2015 -

With cooperation of Dirkjan Beukenhorst, Jan van den Brakel, Laura Clarke, Frank Engelen, Jan van der Laan, Anja Liersch, Krister Näsén, Vivian Meertens and Rita Rinaldi.
The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.
# Index

## Executive summary

1. **The organisation of mixed mode data collection**
   1.1 Mode strategies
   1.2 Reference week and mixed mode design
   1.3 Web response
   1.4 Who responds in web
   1.5 Costs
   1.6 Case management systems

2. **Mode effects**
   2.1 The definition, identification and analysis of mode effects
   2.2 Case study on mode effects in the German Labour Force Survey
   2.3 Mode effect decomposition for the Dutch Labour Force Survey
   2.4 Web pilot study of the Finnish Labour Force Survey
   2.5 The impact of survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey.
   2.6 Overall conclusions mode effects

3. **Estimation under multiple mode data collection**

4. **References**
   4.1 References based on ESSnet-DCSS deliverables
   4.2 General references
Executive summary

Work package III of the ESSnet-DCSS concentrated on mixed mode data collection, specifically with a view on the combination of web (CAWI) with other modes (CAPI, CATI, and / or paper). The focus was on mixed mode designs within one fieldwork period and one wave, although attention was also given to the use of web as a mode for the second wave of the LFS. There were three major topics in this work package: the organisation of mixed mode data collection, mode effects, and estimation and adjustment. The findings described here are based on pilots and experiments within the ESSnet consortium, and on the experiences of other NSIs. The latter were queried in an early stage of the ESSnet on their experience with web and mixed mode data collection. Findings of this query are described extensively in Blanke and Luiten, 2014.

1. The organisation of mixed mode data collection including web

This topic describes different strategies for mixing modes: which modes to use and in which sequence. Attention is given to the response rates in web, and the distribution of that response: who responds in web, and who does not. Various measures to heighten web response are described. Attention is also given to the important subject of case management systems: software systems that are able to support all modes and allow flexible transitions from one mode to another.

Which mixed mode design

From the query we learnt that a plethora of designs is used in the ESS, with all kinds of mixes, in all kinds of sequences. Also, within NSIs modes and mode mixes do differ for various surveys. An important lesson of this ESSnet was that there is no such thing as the best mixed mode design. The design should depend on the purpose of introducing mixed mode designs. If the purpose is to decrease costs, a consecutive design where the cheapest mode is offered first is the obvious choice. On the other hand, if the purpose is to decrease coverage error or non-response error, a concurrent design where the respondent is given a choice of modes, or a consecutive design where the most expensive mode is offered first, should be considered.

We recommend that design decisions need to take into account different aspects of data quality, response burden and costs. As the methodological, practical, technological, legal and other framework conditions will differ from country to country, it is difficult to give very detailed advice.

Mixed mode designs and response rates

A number of NSIs mentioned that they expected that response rates would rise as a result of implementing mixed mode data collection. However, both the query and the pilots in the ESSnet show that it is certainly not a given that response rates will go up, or will even be maintained in a mixed mode design. The response rate may go up if one starts with the most expensive mode, does all one’s best to attain the highest possible response in that mode, and subsequently offers other modes to non-respondents. On the other hand, response rates will probably go down somewhat if one starts with the cheapest mode and then offers other modes to non-respondents. The reason why this happens is not yet clear, but the phenomenon is found repeatedly.
Web response

A clear conclusion from the ESSnet and other experiences is that web cannot (yet) be the only mode in official statistics. Generally, response rates are low and biased. The rates that are attained depend on the design and are a function of the timing of the web mode (as first mode, as choice mode, as mode in a second wave, etc.), of the legal status of the survey (mandatory of voluntary) and of the topic of the survey (within one NSI, rates may vary from 10% to almost 50%, depending on the topic). In the pilots of this ESSnet, the web response of the LFS, when offered as the first mode in a voluntary survey, varied in the 25-30% region.

NSIs that consider web data collection, hope that offering web will attract respondents that are under-represented in the main stream modes. We found that this is not the case. The same respondents who respond in CATI and/or CAPI are also the ones who respond in web. The same groups that are underrepresented in CAPI and CAPI are underrepresented in web: the young, the ethnic minorities, and people with low education and/or low income. The highest web response in a web-first design was in various cases found in the 55-65 year old group, with the runner up in the group of 66 to 75 year old.

Because of these low rates, the bias, and the lack of complete internet coverage in most countries, web cannot, at least for the foreseeable future, be the only mode. It must be complemented with other modes in a mixed mode design. The ESSnet, as well as experiences in countries that already employ mixed mode data collection with web, showed that these mixed mode designs can show response rates and representativeness that are similar to CAPI.

The report describes various recommendations to attain a high web response rate, among those the advance letter, the number of reminders, the timing of the advance letter and reminders, incentives, look and feel of the questionnaire and the login procedure.

Costs

An important consideration for many countries in introducing mixed mode data collection is the costs of fieldwork. Most countries indicate that fieldwork costs decreased after the introduction of mixed mode data collection, sometimes by significant amounts. Others however indicated that costs became higher or remained the same. The costs development is essentially a factor of the initial design.

Although fieldwork costs for web data collection are relatively low, introducing web into the design may have negative cost implications for other aspects of fieldwork. For example, because the remaining persons may need more effort to trace and persuade, because interviewers may expect higher remuneration for the more difficult work, or because travel distances are longer.

Fieldwork costs are but one of the components of the costs of running a survey. Introduction of mixed mode data collection can have cost implications for all other stages, like sampling, data processing, case management, and adjustment. Especially if IT systems need to be bought or developed, that could mean a major investment. Also, the greater complexity of the other stages may potentially reduce the costs savings that can be attained. For example, multiple modes may mean multiple questionnaires, and more testing. NSIs indicated that some costs went up while others went down. Some indicate that, although other costs may have gone up, this is easily compensated
by the decrease in costs of data collection. Most countries foresee that the balance of costs and profits of mixing modes will become more profitable in the coming years.

## Case management systems

Case management is a critical function for efficient operation of an interviewing system. Case management systems provide the interviewer with sample cases, feed the appropriate information to the interviewing system, and accept the completed survey information, together with relevant paradata, back. For CATI systems, functionalities include sample management, call scheduling, case management, online interviewing, online monitoring, keeping track of call histories, keeping track of call outcome data, management of interviewing staff, etc. CAPI systems may in addition need functionality to record interviewer time and expenses, information about respondents, and possibilities for random sampling within a subset of household members. Designing an efficient case management system for one mode is complicated in itself but getting the systems to communicate, which is necessary in mixed mode designs, adds substantially to the complications. A number of NSIs are presently in the process of developing new case management systems for mixed mode data collection (e.g., Norway, Austria, the Netherlands, the UK). Some develop the system themselves; others try to adapt OTS software to their purpose.

The partners in this ESSnet have each provided lists of requirements within their organisation for case management systems. Since there are considerable differences in how the LFS data collection is organized, it is difficult to specify more than a very general list of requirements for a case management system. The report therefore provides a short list of minimum requirements for a multi-mode case management system.

When planning a case management system one has to decide whether to create the system in-house or to try and find an existing software package that meets the necessary requirements. There are several examples of differing solutions of current and planned systems: several NSIs are building their own system (e.g., Sweden and Austria), Statistics Netherlands is investigating a Customer-of-the-shelf solution. Most other countries investigate or build hybrid systems, with a combination of COTS and in-house systems.
2. Mode effects

The study of measurement effects specific to the mode of data collection (mode effects) is one of the major objectives of work package III. Within this ESSnet, several analyses on the prevalence and magnitude of mode effects in the Labour Force Survey were undertaken. One of the deliverables in the ESSnet was an overview of the definition, identification and analysis of mode effects. This report gives a short summary of its findings. Measuring mode effects is complex. The deliverable describes four quantitative approaches and a qualitative one that are used most frequently to determine mode effects. Each of these has its advantages, but all of them also have limitations, so that there is no 'best' way to determine mode effects. Subsequently, three experiments are described.

The German experiment compared web to CATI, CAPI and paper-and-pencil in a random allocation experiment. The conclusion was that mode-specific measurement effects are no major impediment to using multiple data collection modes in the LFS. The use of computer-assisted web interviewing as an additional data collection mode is feasible. For most of the variables under consideration, there were only minor differences between the data collected via interviewer-assisted data collection modes and CAWI. In contrast, numerous differences were found comparing the data collected via self-administered paper-and-pencil questionnaire (PAP) and the other three modes under consideration, indicating that the PAP data were clearly inferior in terms of data quality.

The Dutch findings were based on comparison of parallel runs of a mixed mode (web, CATI, CAPI) LFS with the traditional CAPI design. In addition, an experiment compared LFS and other variables in a random allocation of web, CATI, CAPI and paper-and-pencil. The first round of random allocation was followed by a re-interview round in (mostly) CAPI. The experiment was set up to be able to decompose mode effects into three components: mode-specific coverage bias, mode-specific nonresponse bias and mode-specific measurement bias. Coverage bias arises because of differences in access to survey modes, i.e., in access to Web and in registration of a telephone number. Nonresponse bias is the result of differences in contactability and response propensity over survey modes. Measurement bias is the consequence of different answers to different survey modes from the same respondents.

It was found that employment and unemployment estimates for CATI and Web differ significantly from those in CAPI. However, selection effects could be explained to a large extent by standard registry variables of Statistics Netherlands. Measurement effects for both CATI and web were not significantly different from zero. However, the estimates for the survey questions underlying employment status (having a job, wanting a job, seeking a job and being available) do show measurement effects. The questionnaires should be evaluated and possibly adapted in order to reduce these differences between modes.

Finland compared CAWI to the traditional CATI questionnaire of the LFS. They found that there were no significant mode effects on employment status between CATI and CAWI. CAWI respondents reported a lower number of hours worked in the reference week. The difference could be diminished by adapting the question in CAWI, and thus essentially offering a different question in the two modes.
Conclusions on mode effects in the LFS

There is evidence from this ESSnet and other findings that measurement errors are an important source of differences between modes in some but not all surveys. For the LFS the differences between modes (mail, web, CATI, CAPI) can mostly be explained with common weighting variables. This finding is replicated in the Netherlands, Germany and Finland. However, this is not always the case: large mode effects in the Dutch Safety Monitor have led to a restriction in the modes to web and paper, that is, non-interviewer modes. Likewise, in the Finnish Consumer Sentiments Survey large mode differences led to the decision not to introduce web data collection in this survey. Research on the British Opinions Survey showed that mode effects can be explained, but in order to do that, additional auxiliary variables are necessary, more than are standardly available for weighting.

Experiments and case studies show that mode effects should be taken seriously, however, it is hardly possible to determine for each survey and each mode combination whether mode effects are of importance. We recommend to develop rules of thumb for choices in the survey design: is every sample unit subjected to the same modes, will we be able to adjust afterwards, can we stabilize findings. These choices depend on the relative amount of selection and measurement effects, and on the choice of benchmark (e.g., CAPI, or a mix of modes).

3. Estimation in mixed mode data collection

When the composition of the response obtained in a mixed-mode data collection strategy can fluctuate – between subpopulation or between time periods – comparability of the resulting statistics may be compromised. If the modes exhibit relative measurement errors for a survey variable, then these measurement errors constitute part of the final estimate for that variable, as variations seen in the mode composition will manifest themselves as fluctuations in resulting LFS estimates, a situation that is to be avoided. The deliverables written for the third topic in the ESSnet investigate the effect on the estimates of stabilizing the mode composition.

Although a large body of literature is available on quantifying and explaining mode effects, not many approaches have been developed so far to correct for mode-dependent measurement errors. In the chapter on estimation, three methods are tried: the methods of Suzer-Gurtekin et al. (2012) of Buelens and van den Brakel (2014), that address comparability and stability of estimates in scenarios where the response mode composition fluctuates. The objective of this approach is to stabilise the potential mode-specific measurement error bias in the estimates over time and across subpopulation groups. Another approach, developed in Kim (2013), attempts to adjust returned values at unit level for continuous variables, assuming that the measurement error results in no bias but increases the variance. This approach attempts to remove any potential measurement error bias in the estimates of all subpopulation groups.

The method of Buelens and van den Brakel (2014) adjusts the GREG weights in such a way that they simultaneously correct for selection effects and obtain a specific, pre-set
response mode composition. The method is applied once and is valid for all survey variables. The method of Suzer-Gurtekin et al., (2012) leaves the GREG weights unaltered, but adjusts the survey answers of the respondents. This method is specific to the survey variable in question. Both methods assume a particular measurement error model and assume that the covariates explain the selection effect.

It appeared that the two methods give very similar results for the estimated number of unemployed based on the first wave of monthly LFS data. Furthermore, the two adjusted series are comparable to the original, regular series in use for official publication purposes.

The third method, by Kim (2013), was applied to employment data from a split face-to-face and web sample design on British Opinion Survey data, and used a number of socio-demographic variables as model covariates. In the resulting solution, the estimated measurement error bias was thought to be rather too large, indicating that the set of available covariates used in the models could not account for the effects of mode selection.

This is an area where research has only just started, and further work needs to be done. We have seen that methods for adjusting for measurement errors can be developed but they have limitations and rely on assumptions that are difficult to verify. In all methods, auxiliary data need to be available and of good quality. These can be registry data or paradata that resemble registry information. The method in Buelens and van den Brakel (2014) shows promise and is relatively easy to implement, but it only addresses the problem partially since the bias is not removed. Even though the measurement bias in the survey estimates will not be removed, applying these kind of adjustment methods is recommended to keep the measurement bias under control.

For the time being, it is only through careful questionnaire design and pre-testing of questionnaires in mixed mode settings that relative measurement effects can be avoided or reduced.
Introduction

Work package III of the ESSnet-DCSS concentrated on mixed mode data collection, specifically with a view on the combination of web (CAWI) with other modes (CAPI, CATI, and/or paper). The focus was on mixed mode designs within one fieldwork period and one wave, although attention was also given to the use of web as a mode for the second wave of the LFS. There are three major topics in this ESSnet. This is a summary of the findings. Chapter 2, on the organisation of mixed mode data collection, describes mode strategies: which modes in which sequence, the timing of mode switches in view of the reference week, response rates and the measures to heighten web response, and who are the web respondents. Attention is also given to the important subject of case management systems: software systems that are able to support all modes and allow flexible transitions from one mode to another. Chapter 3 is on mode effects: the phenomenon that substantive findings may be different, as a result of differences in mode specific coverage, response and measurement errors. Chapter 3 starts with a short summary of the essence of mode effects, and how they can be detected. Subsequently, three studies are described on mode effects in mixed mode LFS designs. Chapter 4 discusses estimation and adjustment: given that there are mode effects, can we adjust for them, and how should that be performed.

In the query among European and other NSIs, a large amount of questions concerned aspects that are part of Work Package III. This includes the timing of mode switches, response rates, organisation of mixed mode, mode effects and adjustment for mode effects. The results of the query will not be summarised separately, but findings will be incorporated in the chapters below.

Where possible, we report findings on the LFS, but findings with other surveys are mentioned as well.
1. The organisation of mixed mode data collection

One of the goals of work package III of the ESSnet-DCSS is to compile the available knowledge regarding the organization of multi-mode data collection, including aspects like fieldwork, case management, non-response, panel attrition, and timeliness. In discussion with the partners, the most relevant topics were ascertained, i.e.,

- Mode strategies
  - How to mix modes within waves
  - How to mix modes between waves
  - Can we use different modes within one household
  - When should you switch from one mode to another?

- How to achieve the highest possible web response rates
  - How to formulate the advance letter
  - How to remind (which mode, how many times)
  - Who uses the web

- Case management:
  - Which software to use
  - Requirements for case management systems

- Costs
  - How to balance costs, quality and user friendliness
  - Reducing travel costs for CAPI interviewers

- Timeliness

All partners contributed to these subjects, and most subjects were covered by more than one partner, so that findings could be compared. A summary of the findings is written by Gravem, Båshus, and Lagerstrøm (2014). More extensive findings can be found on the cross portal http://www.cros-portal.eu/content/data-collection.

In the following paragraphs we discuss mode strategies: how to incorporate web in a mixed mode strategy (2.1), how to reconcile mixing modes with a fixed reference week (2.2), web response rates (2.3) web respondents (2.4), and costs (2.5). A well-functioning multi-mode case management system is important in order to keep costs and man-hours down, and to be in line with timeliness demands. Many NSIs lack integrated case management systems, necessitating the transfer of information e.g. from a CAWI system to a CATI system or vice versa, before the data collection can move from one mode to the next. We summarize requirements for case management systems in 2.6, where we also discuss the issue of building or buying a system.
1.1 Mode strategies

As to the use of web in a mixed mode design, a number of general designs can be envisaged:

1. Web is offered as the first mode in a sequential design, i.e., a design where first one and then one or more other modes is offered. Non-respondents after the first data collection phase in web are followed up in other modes. This design offers opportunities for costs savings, depending on response rates, sampling frame, data collection process, see chapter 2.5. It runs the risk, however, of lower response rates and may be difficult to use in a survey with a fixed reference period, like the LFS. This design is used by the Netherlands (both for the LFS and other social surveys), Bulgaria (for the census), France (for the ICT sample without known telephone), Spain (for the EHSIS), and the US Census Bureau (for the American Community Survey).

2. Web is offered concurrently, alongside other options, as a choice for respondents. This is more often used when the other option is a paper questionnaire. The design will limit potential costs savings, as respondents tend to favour the paper questionnaire. This is the most commonly used mixed mode design, used by Denmark (for the 2nd wave of the LFS; the concurrent mode is CATI), Belgium (EU-SILC), Italy, Hungary and the UK (for the census), Sweden (for the Health Survey), Norway (for the Rental Survey; the concurrent mode is CATI), Switzerland (for the Structural Survey). In countries where the concurrent mode is not mentioned explicitly here, the concurrent mode is a paper questionnaire. In some cases, modes are offered concurrently, but NSIs put differential stress on the different modes, in order to move respondents to the preferred mode.

3. Web is offered to non-respondents of earlier CATI or CAPI trials. This is the design of choice when the aim of augmenting response rates is the driver of design change. This design is used by France in the first and later waves of the LFS, albeit that a smaller version of the original questionnaire is offered to the non-respondents of the regular mode.

4. Web can be offered in subsequent waves, after a first wave in a different mode. This is done by Denmark (for the LFS household sub sample and EU-SILC), Canada (LFS) and Norway (for the Rental Survey). Germany will use this design for the LFS.

5. What mode or what sequence of modes is offered to which sample unit does not need to be an all or one decision: in adaptive or responsive survey designs, designs are adapted to characteristics of the sample unit. A simple version would be to not offer web to the (very) elderly, or not to call households with more than three adults. More advanced possibilities are to use insights into measurement errors per group per mode to fine-tune mode allocation. This is demonstrated by Calinescu and Schouten (2013) for the Dutch LFS, who propose mode switches dependent on a given sample and budget. In their model, response probabilities of nine different subsamples based on demographic data are calculated, while also taking effect on data quality and costs into account. The model predicts that households with unemployed people, non-working young people, and non-western people should be approached with CAPI, while working western people should be approached
with web followed by CATI. The remaining groups may as well not be approached, as response propensity is very low and the impact on the estimates would be negligible.

### Recommendations

- **Which design is chosen, is closely linked with the purpose of the mixed mode design and the organisational, legal, institutional and cultural context.** If the purpose is to decrease coverage or non-response error, then either a concurrent design where respondents are given a choice of modes or a consecutive design where the most expensive mode is presented first are the options. If however the purpose is to reduce costs, then a consecutive design with cheapest mode first is the best option.

- **Design decisions need to be evidence-based as far as possible, taking into account different aspects of data quality, response burden and costs.** As the survey methodological, practical, technological, legal and other framework conditions will differ from country to country and NSI to NSI, it is difficult to give very detailed advice.

- **Where appropriate, adaptive or responsive designs should be considered, so that modes are allocated to some (groups of) people to minimise measurement bias, response, and / or costs.** Further research on this topic is necessary.

### 1.2 Reference week and mixed mode design

The fixed reference week in the LFS puts constraints on the mixed mode design, especially if a sequential design is envisaged. Three ESS countries (Germany, the Netherlands and Slovenia) circumvent this problem by using a rolling reference week where the reference week is the week prior to the interview or the week of the interview. The other countries manage the constraint through rigorous adherence to time frames. Denmark mentions that experiments have shown that after five days no further responses are to be expected in CAWI, which is why they choose that date as the point to switch from CAWI to other modes. Likewise, Finland has introducing the second mode after three days. Statistics Canada offers the web option for four days, during which two reminders are sent. On the other hand, in the Netherlands the CAWI period is extended to one month in order to profit longer from the cheaper mode. By sending reminders, substantial additional response is gained, but the fieldwork period in the Netherlands is too long for a fixed reference week.

### Recommendations

The further away from the reference week the interview is held, the higher the chance that measurement error is introduced. A (consecutive) mixed mode design will need to be performed with this time constraint in mind. Flexible transition from one mode to another is a prerequisite, and with that IT systems that support that flexibility. We will come back to this in chapter 2.6.
### 1.3 Web response

<table>
<thead>
<tr>
<th>Response rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the query that was sent out to the ESS, we asked for the reasons for NSIs to adopt mixed mode data collection. Increasing response rates was an answer that was often given. However, experience, both from the query and the pilots in the ESSnet, shows that it is certainly not a given that response rates will go up, or will even be maintained in a mixed mode design. Out of 14 comparisons of multi-mode designs with prior single modes for voluntary surveys in the ESS, four response rates went up after redesign to mixed mode, five response rates went down, two remained the same and three comparisons could not be made. Out of 12 comparisons for mandatory surveys, two response rates went up, one went down, three remained the same, and six comparisons could not be made.</td>
</tr>
</tbody>
</table>

We tried to ascertain under which conditions response rates go up or down after introducing mixed mode data collection. Conclusions are hard to draw from the non-experimental data set of the query and the many different designs, but it appears that in the countries where the response rates went up, the mixed mode design was mostly concurrent, while in countries where the response rates went down the mixed mode design was more often sequential, with the cheaper modes as the first option offered. The direction of these findings concurs with those of Lynn (2013), who stated that response rates generally go up when going from a single mode to a mixed mode design, but only if one starts with the most expensive mode, does all one’s best to attain the highest possible response in that mode, and subsequently offers other modes to non-respondents. When, on the other hand, one starts with the cheapest mode and subsequently offers other modes to non-respondents, response rates are liable to drop some percentage points. |

The response rates that can be attained with web data collection are again dependent on the design: if web is offered first as the only mode, higher web response rates will be attained than if web is offered as a choice alongside other modes, or as a final offer after other modes. If web is offered as a choice, while, for example, simultaneously offering a paper questionnaire, the overwhelming majority of respondents will choose the paper questionnaire. |

Response rates are dependent on a range of other design options as well: the length of the fieldwork period, the number of reminders, and whether the survey is voluntary or mandatory. The subject of the survey is of importance as well: web response for CBS statistics (voluntary surveys were web is the first mode) range from 15% for the Household Budget Survey to 36% for the Health Survey. The web response for the Dutch LFS is 27% (fluctuating over months between 24% and 29%). This is comparable to the Finish web response rate of 30% in the LFS pilot (range 26% to 34%). Other web response rates in the ESS are 23% for the voluntary Norwegian Rental Survey, 33% for the Italian Census, 15% for the British Census, 27% for the Swiss Census, 23% for the Spanish EHSIS, 35% for the Dutch General Health Survey. |

<table>
<thead>
<tr>
<th>Advance letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>In order to attain high web response, several measures can be taken. Via the query, information became available on the advance letters and the number of reminders. Advance letters for web surveys need specific attention. In addition to the regular information on the purpose of the survey and why the respondent should participate,</td>
</tr>
</tbody>
</table>
information on how to log on to the web, what passwords and other security measures are necessary, need to be described. This may result in a long and complicated letter. We have seen various examples of advance letters that use different approaches to handle this complication, like breaking up the information in various communications, or structuring the letter with clear headings and short descriptions. See also Luiten (2014) for an experiment with advance letters for web surveys.

Statistics Netherlands and Statistics UK time the arrival of advance letters and reminders just before the weekend, but the finding that this secures a speedy response may be culturally dependent, and needs replication.

---

**Reminders**

Most countries send two reminders, even when the period in which web data collection is foreseen is very short. For example, Statistics Canada sends two reminders in four days. The US Census Bureau likewise sends the first reminder after two days. Each reminder results in a clear increase in response.

Various channels may be envisaged for the reminders: letters, cards, emails, text messages on cell phones or telephone calls. The most commonly sent reminder is a letter, although the timing of the letter varies widely; it can be sent either after one, two or three weeks for the first reminder and one, two or three weeks after the first reminder. Australia embraces modern communication means for the reminders by sending e-mails and text messages. The Netherlands mention that research has shown that additional reminders above two would still bring in additional web response. However, sending more than two reminders has a bad influence on subsequent CATI and CAPI response because sample units start feeling harassed.

Statistics Norway has experimented with dedicated paper and sms reminders for specific groups of respondents, based on paradata. They used Cialdini’s principles of compliance to determine which group of respondents received which message. Statistics Netherlands has experimented with the form of the reminder (visually salient card or letter). The card appeared not to work as well as the letter, because the login information could not be printed on a card for all the world to see. Statistics Finland experimented with motivation calls versus text messages. Both showed a clear increase in response rates. Finland suggests further research into these methods. Norway warns not to send text invitations on cell phones, unless the questionnaire can be filled in on smartphone.

When a consecutive design is used, the number of reminders must be handled with care, in order not to unduly harass the sample units who will still need to be approached with other modes.

---

**Web access**

Not all persons have access to the web. That does not need to imply that high web response rates are not possible. For example, Statistics Finland has shown that the web access in the groups with the highest web response rates (the persons of 55 years and older) is actually lower than in the younger groups with the lower response rates. This indicates that the lower possibility to access the web is more than compensated by a higher willingness.

Even if people can access the web, make sure that respondents understand the instructions on how to access the web questionnaire. Qualitative research by Statistics...
Canada showed that a substantial part of the respondents were not able to follow the instructions in the initial letter. Very clear instructions appeared to be necessary. See also the WP2 contribution on login procedures.

When designing CAWI questionnaires, take into consideration that a large part of the population will have very low computer skills, as confirmed by the PIAAC study of the OECD (OECD 2013). The accessibility and usability of CAW questionnaires should perhaps accommodate persons with computer skills corresponding to PIAAC’s level 1.

**Incentives**

Another measure that can be taken to increase (web) response rates is giving or promising an incentive. The benefits of cash incentives for improving mail response rates are clearly established, and research shows that advance cash incentives are more powerful for improving web response than mail survey response (Messer and Dillman, forthcoming). The measure is not commonly taken by NSIs, although some do, and some do it in times of expected or manifest low response rates (Blanke and Luiten, 2014). Experimental findings in the Netherlands showed that unconditional incentives in a web survey may increase response rates by over 15 percentage points. In the Community Life Survey study conducted by the Cabinet Office in the UK, incentives increased response rate from 16% (with no incentive) to 19% with a conditional £5 incentive, 22% with a conditional £10 incentive, and 25% with an unconditional £5 incentive. However, the web incentives did not achieve the response rate of face-to-face interviews with a conditional cash incentive of £5 (60%).

Of course, using incentives has implications for the costs of the survey. Depending on the design, however, even an unconditional incentive of €5,- to all sample units may be cost effective if the alternative mode is CAPI, but probably less so if the alternative is CATI.

**Recommendations concerning web response**

Although more knowledge on this subject is necessary, from the experience thus far, we recommend:

- Do not make advance letters too ‘commercial-looking’.
- Warn respondents in the advance letter and reminder(s) that they will be contacted by an interviewer if they do not respond in web. The finding that this increases web response has been replicated by several countries.
- Study the effect of the arrival time of the letter. Statistics Netherlands makes sure that the advance letters arrive on a Friday, making it possible for respondents to complete the survey online during the weekend. This finding is replicated in the UK, but should be replicated further to see if this can be generalised,
- Formulate instructions concerning web access very clearly and perform usability tests.
- The accessibility and usability of CAW questionnaires should accommodate persons with low computer skills.
- Send reminders. Even within very short fieldwork periods. More research should be performed as to the timing of reminders.
- Study if sending reminders has negative consequences for later wave participation.
- Always include login information with CAWI reminder notifications.
- If notifications are sent via text messaging (SMS), the questionnaire should be developed using a “mobile first” approach.
- Apart from the aspects concerning communication with the respondent (letters, reminders, incentives), the look and feel of the questionnaire (the subject of WPII) needs to be inviting.
- Finally, the technical aspect of web servers is of particular importance: servers need to be secure and available at all times.
- The effectiveness of incentives should be carefully evaluated in each situation.

1.4 Who responds in web

It is often hoped that the introduction of web data collection will reach groups that are traditionally under-represented in the response, especially younger people. The experience in this ESSnet shows that this is not the case: the groups that tend to respond well in other modes are also the ones with the highest web response. Both Finland and the Netherlands show that the highest web response is attained in the group of 56 to 65 year old people, with the runner up in the group of 66 to 75. In a different setting (a follow up study) Norway finds the highest response in the age group of 30 to 44. The lowest response rates are for the youngest age groups in all three countries. Web response is higher the higher the education (FI and NO), and the higher the income (NL). CBS does not find gender differences, but Norway and Finland find higher web response for females.

The web use reported here is found in situations where the web is the only mode offered. If web is offered concurrently, the distributions will be different.

<table>
<thead>
<tr>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>For the time being, web cannot be the only mode, as the response rates are low and biased. Mixed mode designs on the other hand show response rates and representativeness that are similar to CAPI. An important finding is that web shows unstable response rates over months of data collection. That is particularly problematic if there are measurement differences between the modes. If shares of modes of response vary over time, substantive findings may vary too, as a result of these fluctuations. It may therefore be desirable to stabilize relative mode use, both by fieldwork and by calibration. We come back to this issue in chapters 3 and 4.</td>
</tr>
</tbody>
</table>

1.5 Costs

An important consideration for many countries in introducing mixed mode data collection is the costs of fieldwork. The information on this subject derives from the query. NSIs indicated how costs of fieldwork developed after the introduction of a mixed mode design. Most countries indicated that fieldwork costs decreased after the introduction of mixed mode data collection, sometimes by significant amounts. For example, fieldwork costs in the Netherlands diminished with more than 50% when a CAPI design was changed for a consecutive web - cati - capi design. Others however indicated that costs became higher or remained the same. The costs development is essentially a factor of the initial design. For example, in Italy, where a paper questionnaire was replaced by a CATI/CAPI mixed mode design, costs increased.
Although fieldwork costs for web data collection are relatively low, introducing web into the design may have negative costs implications for other aspects of fieldwork. Examples are:

- The sample units responding in web are the ‘easiest’ respondents. The remaining persons may need more effort to trace and persuade.
- Because the work is harder, interviewers may expect higher remuneration.
- If web is introduced as the first mode, less work is available for the CAPI field interviewers, and interviewers have to travel longer to reach their addresses. This may have substantial influence on the costs of fieldwork, thereby diminishing the potential costs saving by web data collection. The following list from Statistics Netherlands summarizes a number of alternatives that can be chosen to alleviate the situation:
  - Train CAPI interviewers to do CATI work as well,
  - Cluster sample units, with the risk of diminished precision,
  - Limit the number of visits per address, with the risk of introducing bias. Precision may remain intact if the sample is enlarged,
  - (Regionally) subsample CAPI addresses, with the risk of introducing bias. Precision may remain intact if the sample is enlarged.

Fieldwork costs are but one of the components of the costs of running a survey. Introduction of mixed mode data collection can have cost implications for all other stages, like sampling, data processing, case management, and adjustment. Especially if IT systems need to be bought or developed, that could mean a major investment. Also, the greater complexity of the other stages may potentially reduce the costs savings that can be attained. For example, multiple modes may mean multiple questionnaires, and more testing. NSIs indicated that some costs went up while others went down. Some indicate that, although other costs may have gone up, this is easily compensated by the decrease in costs of data collection. A substantial number of NSIs state that costs did not change as a result of mixed mode data collection, although they were not referring to the introduction of web. Most countries foresee that the balance of costs and profits of mixing modes will become more profitable in the coming years by the introduction or the increased use of web surveys.

1.6 Case management systems

Case management is a critical function for efficient operation of an interviewing system. Case management systems can be seen as ‘the glue between the system user and the system software (Connett, 1998). Case management systems provide the interviewer with sample cases, feed the appropriate information to the interviewing system, and accepts the completed survey information, together with relevant paradata, back. For CATI systems, functionalities include sample management, call scheduling, case management, online interviewing, online monitoring, keeping track of call histories, keeping track of call outcome data, management of interviewing staff, etc. CAPI systems may in addition need functionality to record interviewer time and expenses, information about respondents, and possibilities for random sampling within a subset of household members. Designing an efficient case management system for one mode is complicated in itself, but getting the systems to communicate, which is necessary in a mixed mode designs adds substantially to the complications. A number
of NSIs is presently in the process of developing new case management systems for mixed mode data collection (e.g., Norway, Austria, the Netherlands, the UK). Some develop the system themselves, others try to adapt OTS software to their purpose.

The partners in this ESSnet have each provided lists of requirements within their organisation for case management systems. Since there are considerable differences in how the LFS data collection is organized, it is difficult to specify more than a very general list of requirements for a case management system. We therefore provide this short list of minimum requirements for a multi-mode case management system:

- The system should support multi and mixed mode surveys in all the relevant modes.
- A modular approach should be considered in order to easily support modes as needed, where modules/components can be added or replaced rather than making modifications to a monolithic system.
- A modular case management system should accommodate alternative collection tools, it may for example not be avoidable to use different tools for CAPI and CAWI.
- Real-time updating of case information from electronic mode is a prerequisite for monitoring, and also for conducting concurrent mixed mode data collection. For example, interviewers need to know real time if their cases have filled in a web questionnaire, preferably before they contact them.
- This requirement implies that the system must be able to notify mobile devices.
- The system should be able to handle a variety of sample types, such as:
  - Individual samples
  - Household/family samples
  - Address samples
  - Business samples
- Interviewer administration.
  - Registration of working hours.
  - Payment
  - Reports on performance, working hours, etc
- Monitoring and progress reports.
- Real-time reports necessary for informed decision making and responsive data collection.
- Built-in flexibility for future requirements.
- The system should be designed in modular fashion, so that it can be easily expanded to cover future functionality.

**Buy or develop**

When planning a case management system one has to decide whether to create the system in-house or to try and find an existing software package that meets the
necessary requirements. There are several examples of differing solutions of current and planned systems:

‘In-house’ systems:
- Sweden built a complete survey system, which is currently being expanded to better be able to support multi-mode.
- Statistics Austria is in the process of developing their own mixed-mode case administration system handling all planning, conducting and monitoring of the fieldwork. According to Statistics Austria, this system could potentially be turned into an off-the-shelf product (Plate, 2014).

(Customer) off-the-shelf (COTS):
- CBS is currently investigating a COTS solution for a new survey system integrating all modes

Hybrids:
- Common in many NSIs including Norway. Often they use a combination of Blaise for call scheduling, with an ‘in-house’ developed case management system built on top.

The advantage of an in-house product is that tailor-made solutions are possible. On the other hand, building in-house asks much from the organisation in terms of planning and project management, and it can be costly if mistakes are made in either. OTS products are usually developed for the requirements of market research, and it is likely that functionality needed by NSIs is lacking and needs to be developed additionally. Linking the OTS product to existing functionality or to newly developed solutions may be difficult.

Despite the differences between NSIs, there are bound to be similarities too. The solutions found by one NSI may be inspiring for others. Especially in view of the costs of the development of case management systems, this could be a fruitful topic for collaboration between NSIs.

### In-house support for multi-mode data collection

The following is a list of support and collaboration issues between the data collection department and other actors in an NSI based on feedback from all five DCSS participants. In practice, several of these services can and will be covered fully or partly by the data collection unit of a given NSI, but which ones and to what extent is likely to vary.

**Subject matter department**
- Questionnaire content, concepts and requirements. Participation in development and adaptations for different (new) modes

**Methodologists**
- Questionnaire design and pretesting for multiple modes
- Estimation and weighting design for multi-mode

**Respondent helpdesk**
- Extended opening hours to accommodate respondents’ needs and preferences
- E-mail service
- Real-time interaction with users, screen mirroring or interview initiation
IT department
- Secure infrastructure
- Continuous server checks, instant reaction to server failure
- Survey application development
- Database development
- Technical support for interviewers, field staff and survey managers, also outside of regular working hours

Legal department
- Confidentiality
- Respondents’ rights and obligations
2. Mode effects

The study of measurement effects specific to the mode of data collection (mode effects) is one of the major objectives of work package III. Mode effects can form an important issue in the decision whether an additional data collection mode should be introduced. Within this ESSnet, several analyses on the prevalence and magnitude of mode effects in the Labour Force Survey were undertaken. These are described in paragraphs 2.2 to 2.4. Paragraph 2.5 describes work by CBS on the impact of survey item characteristics on mode specific measurement bias. We start however with a summary of the deliverable written by Körner (2014) on the nature of mode effects, how they can be identified and analysed.

Even though there exists a plethora of findings in literature on mode effects, the lessons to be learnt for the LFS are limited: most empirical studies concern questions relating to attitudes and behaviour, and not on factual questions, the predominant question type in the LFS. Another category of questions frequently asked in the LFS, questions referring to difficult concepts that require additional explanations by interviewers and are cognitively demanding (e.g., employment status, working time), have not been a major topic in survey research. The existing studies that do exist on the LFS are often not based on experimental designs. In this respect, the studies performed for this ESSnet form an indispensable additional source of information on the existence of mode effects in the LFS.

2.1 The definition, identification and analysis of mode effects

Measurement effects induced by the use of different data collection modes are often presented as one of the most serious impediments to the use of multiple modes in social surveys. However, the literature contains many different definitions of what is referred to as a mode effect. Correspondingly, there are various approaches used for the study of mode effects. As a basis for the identification and the assessment of mode effects, this report defines what should be referred to as a mode effect in the context of official statistics, makes an inventory of possible methodological sources of mode effects, and gives recommendations for the identification and analysis of mode effects. The report is based on the relevant literature that has been identified in the context of a literature search. It tries to apply the concepts and recommendations given in the literature to the specific context of the Labour Force Survey (LFS) of the European Union. Wherever possible it gives examples taken from the LFS or other social surveys of official statistics.

Defining mode effects

Körner gives an extensive overview of several definitions of ‘mode’, and the influence of the communication channel, the contacting process, the degree of involvement of the interviewer and the respondents, and the degree of computer assistance. When comparing the system of data collection, mode differences can be caused by differences in coverage between modes, mode specific non-response, as well as measurement differences. The latter are the so called ‘pure’ mode effect. The remainder of the report by Körner focusses on these mode specific measurement effects.
First, the various sources of mode effects are distinguished, within the four cognitive stages of the response process (comprehension, retrieval, judgment and response (Tourangeau, Rips, and Rasinski, 2000)). These may interact with task difficulty, respondent motivation and ability to lead to ‘satisfying’, that is, not giving the best answer, but rather a satisfactory one. Examples are selecting the first plausible answer, or rather the last one that was read, acquiescence, non-differentiation, random answering, or answering ‘don’t know’. Table 3.1 gives an overview of the key factors that are distinguished in the paper: type of social interaction, type of communication, questionnaire design, and the extent of computer assistance. The chapter on sources of mode effects specifies what aspect within these key factors potentially lead to measurement effects, what these measurement effects are, and which question types are specifically susceptible to measurement effects.

Table 3.1: Key factors for differential measurement in different data collection modes, possible measurement effects, and types of questions likely to be affected

<table>
<thead>
<tr>
<th>Key factor</th>
<th>Aspects leading to measurement effects</th>
<th>Possible measurement effect</th>
<th>Examples for relevant question types</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Type of social interaction</td>
<td>Involvement of an interviewer</td>
<td>Social desirability bias</td>
<td>Sensitive questions</td>
</tr>
<tr>
<td></td>
<td>Social norms governing interaction</td>
<td>Satisficing, social desirability bias</td>
<td>Any; sensitive questions</td>
</tr>
<tr>
<td></td>
<td>Perceived confidentiality</td>
<td>Social desirability bias</td>
<td>Sensitive questions</td>
</tr>
<tr>
<td></td>
<td>Respondent’s control over questionnaire</td>
<td>Question order effects, satisfying</td>
<td>Any; difficult questions, questions with long lists of response items</td>
</tr>
<tr>
<td>(2) Type of communication</td>
<td>Personal vs. impersonal communication</td>
<td>Social desirability bias, satisfying</td>
<td>Any; sensitive questions</td>
</tr>
<tr>
<td></td>
<td>Verbal vs. non-verbal communication</td>
<td>Social desirability bias, satisfying</td>
<td>Any; sensitive questions</td>
</tr>
<tr>
<td></td>
<td>Computer-mediated vs. not computer-mediated communication</td>
<td>Social de-contextualisation</td>
<td>Sensitive questions</td>
</tr>
<tr>
<td>(3) Questionnaire design options</td>
<td>Visual vs. aural stimuli</td>
<td>Recency and primacy effects</td>
<td>Long questions Questions with long lists of response items</td>
</tr>
<tr>
<td></td>
<td>Mode-specific differences in questions</td>
<td>Measurement of deviating concepts</td>
<td>Any</td>
</tr>
<tr>
<td>(4) Computer assistance</td>
<td>Automated routing</td>
<td>Routing errors; item nonresponse</td>
<td>Any; questionnaires with complex skip instructions</td>
</tr>
<tr>
<td></td>
<td>Range and consistency checks</td>
<td>Errors in completion</td>
<td>Any, difficult questions</td>
</tr>
<tr>
<td></td>
<td>Dependent interviewing</td>
<td>Errors in completion, satisfying</td>
<td>Any</td>
</tr>
<tr>
<td></td>
<td>Supervision and paradata monitoring</td>
<td>Deviation from interview protocol</td>
<td>Any</td>
</tr>
</tbody>
</table>

The study of mode effects

Measuring mode effects is complex. Körner describes four quantitative approaches and a qualitative one that are used most frequently to determine mode effects. Each of these has its advantages, but all of them also have limitations, so that there is no ‘best’ way to determine mode effects.

Körner distinguishes experimental and non-experimental approaches. Experimental approaches rely on a design that enables the isolation of one element (in our case, the data collection mode) while the other can be kept constant. Experiments can either
rely on split samples (that are randomly assigned to different data collection modes) or on repeated measurements of the same sample (making use of different data collection modes, but keeping other aspects constant).

Experiments have the drawback that they are resource-consuming as well as difficult to design and to implement, so that in practice analyses are often based on existing data. Examples for approaches using existing data are reweighting approaches (in which a reweighting tries to achieve structurally equivalent groups for each of the data collection modes under consideration) as well as the record linkage approaches (comparing the survey data at micro level with a reference statistics, e.g. a register).

A role somewhat apart is assumed by qualitative research. Here, quantification of mode effects the object (nor is not possible because of the limited number of respondents). Still, qualitative research is one of the few ways of shading light on the cognitive processes that lead to mode effects. While the quantitative approaches heavily rely on the interpretative skills (and sometime the imagination) of the researcher, qualitative research can be a powerful tool to learn more about why mode effects actually occur and how they can be prevented. In order to study mode effects effectively, a combination of approaches needs to be applied. Table 3.2 shows a summary of the approaches described in Körner’s paper.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Objectives</th>
<th>Limitations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split sample experiments</td>
<td>Quantifying the magnitude of mode-specific measurement effects</td>
<td>Practical obstacles to randomised sample allocation; limited sample size</td>
<td>Körner/Liersch 2014; Jäckle/Roberts/Lynn 2010; Schouten/van der Laan 2014; Larja/Taskinen 2014</td>
</tr>
<tr>
<td>Re-interview studies</td>
<td>Quantifying the magnitude of mode-specific measurement effects</td>
<td>Memory effects (independence of interview and re-interview); limited sample size</td>
<td>Schouten/van der Laan 2014</td>
</tr>
<tr>
<td>Record-linkage approaches</td>
<td>Assessing mode-specific differences in comparison with a reference data source</td>
<td>Conceptual and methodological deviations from reference data source</td>
<td>Kreuter/Presser/Tourangeau 2008; Villund 2009, 2010</td>
</tr>
<tr>
<td>Re-weighting approaches</td>
<td>Reweighting the sub-populations interviewed using different modes in regular survey (with non-random allocation of respondents to modes)</td>
<td>Limited choice of calibration variables</td>
<td>Statistisches Bundesamt 2010; Schouten/van der Laan 2014; Vannieuwenhuyze/Loosveldt/Molenberghs 2014</td>
</tr>
<tr>
<td>Cognitive interviewing</td>
<td>Analysing cognitive processes leading to mode effects</td>
<td>Quantification of mode effects hardly possible</td>
<td>Nicolaas et al. 2011</td>
</tr>
</tbody>
</table>

In the next paragraphs of this chapter, three case studies of mode effects are describes, all three deliverables of the ESSnet-DCSS. All three cases are examples of split sample experiments; in all three re-weighting was also applied. One of the three Dutch studies described also used re-interview techniques. Cognitive interview approaches used by the partners of the ESSnet are described in the chapters on work package II. A combination of qualitative and quantitative research is recommended to understand the differences between modes.
2.2 Case study on mode effects in the German Labour Force Survey

Method

The results presented here are from a secondary analysis of an extensive prior experiment with random allocation to four modes: CATI, CAPI, CAWI and self-administered paper-and-pencil questionnaire (PAP). The analysis was targeted on the differences between CAWI and the other modes and focusses on employment status, working hours and variables with numerous answering categories.

2233 households were recruited for this experiment from among respondents to a prior survey. Their ability to answer in the four modes was ascertained. Participation was voluntary. 471, 507, 512 and 503 households participated in CAWI, CAPI, CATI and PAP, respectively.

Conclusions

Mode-specific measurement effects are no major impediment to using multiple data collection modes in the LFS. The use of computer-assisted web interviewing as an additional data collection mode appears to be feasible. For most of the variables under consideration, there were only minor differences between the data collected via interviewer-assisted data collection modes and CAWI. In contrast, numerous differences were found comparing the data collected via self-administered paper-and-pencil questionnaire (PAP) and the other three modes under consideration, indicating that the PAP data were clearly inferior in terms of data quality.

Findings

1. Employment status according to ILO

   Mode specific differences were limited regarding the employment status. One important exception is the PAP sub-sample: Here, the employment rate was significantly lower compared to CAWI as well as CAPI and CATI. CAWI has the highest proportion of employed persons, but is closer to the distributions of CAPI and CATI than PAP.

2. Hours worked

   For both the hours usually and the hours actually worked significant differences between the data collection modes were identified, with PAP respondents indicating a higher number of working hours as in the other modes. The distribution of CAWI is closer to the one found for CAPI than for the one found in PAP.

3. Variables with numerous response categories

   A number of mode-specific differences were identified for variables with numerous response categories. Highly relevant results were found for the highest level of educational attainment and the status in employment. Both seem to indicate that the possibility to present visual stimuli has an important impact on the responses received for questions with numerous response categories. As a general trend, one may conclude that in modes disposing of aural stimuli only, respondents tend to select more general response categories, possibly ignoring the more specific ones equally available on the list.
(e.g., ‘university degree’ instead of ‘doctorate’). Significant differences of this type were found between CAWI and PAP on the one hand and CAPI and CATI on the other.

Also in the question on the main reason for fewer working hours in the reference week, respondents in self-administered modes more often mention less frequent reasons for working fewer hours.

4. Socially desirable answering

No clear evidence was found for the existence of socially desirable answering in the LFS, possibly because LFS variables are generally not considered as sensitive and prone to social desirability bias.

### Cautions

The conclusions above are tentative, as there are a number of circumstances that call for caution in interpreting the results:

1. The assignment of the respondents in the CAWI sub-sample was not fully at random. It is unclear to what extent weighting compensated this bias.

2. Many variables of interest were not surveyed for the entire target population of the LFS, but only for sub-populations identified through LFS variables that were themselves mode sensitive.

3. For many interesting variables, the sample size of the experiment LFS was insufficient.

4. The CAWI questionnaire developed for the experiment was implemented as a single person questionnaire. It is unclear whether an online questionnaire covering the entire household would lead to the same results.

5. The inferiority of PAP cannot be easily generalised as the household PAP questionnaire used in the German LFS is cognitively extremely demanding and therefore inherently error-prone.

### Recommendations

1. The use of self-administered PAP questionnaires in the LFS should be restricted as much as possible. The data quality obtained with the complex format of a household questionnaire used in Germany has been confirmed to be largely inferior to the other three modes under consideration.

2. The implementation of questions with numerous response categories is problematic in interviewer-administered data collection modes (CATI en CAPI). Possible alternative approaches may include branching of the question or the use of field coding instead of presenting a list. A possible follow-up project of the present ESSnet should focus on this issue.

3. Additional empirical research is needed, ideally increasing the sample size available for each mode. Further studies should take special care to achieve better randomisation of the allocation of the sub-samples regarding CAWI and probably use an additional screening for questions regarding the main job.

4. More attention is necessary for other questions (e.g., search for work).
5. Larger experiments are necessary to detect differences in subtle phenomena, or differences between small groups.

2.3 Mode effect decomposition for the Dutch Labour Force Survey

**Method**

Two large scale parallel runs of the Labour Force Survey were conducted in 2010 and 2012, employing two different mixed-mode designs, CATI + CAPI and Web → CATI + CAPI. Furthermore, an extensive mixed-mode experiment was designed and conducted in 2011 within the Crime Victimisation Survey (CVS) in which the key LFS questionnaire module on employment status was included. For this deliverable, a secondary analysis was performed, targeting the LFS variables. Aim was a decomposition of mode effects relative to CAPI and methods to reduce or minimize them.

For the mixed mode experiment, 8800 sample units were randomly assigned to one of four survey modes, CAPI (Computer Assisted Personal Interviewing), CATI (Computer Assisted Telephone Interviewing), web or mail. The full sample, excluding administrative errors and some exceptional nonresponse types like language problems, were approached once more in a second wave, where 70% was allocated to CAPI and 30% to CATI.

The first wave of the CVS experiment was the regular CVS with two modifications. Part of the modules at the end of the survey questionnaire were replaced by the LFS module for employment status and by two sets of four questions from the European Social Survey (ESS). The second wave of the experiment employed a new questionnaire, consisting of: a repetition of the key statistics from the CVS, various attitudinal questions about the survey topic and surveys in general, various evaluation questions about wave 1 and questions about access to web and mode preferences. The additional questions about access to web were necessary to identify wave 1 nonrespondents to web who were not able to respond because of undercoverage.

Thus, mode effects could be decomposed into three components: mode-specific coverage bias, mode-specific nonresponse bias and mode-specific measurement bias. Coverage bias arises because of differences in access to survey modes, i.e., in access to Web and in registration of a telephone number. Nonresponse bias is the result of differences in contactability and response propensity over survey modes. Measurement bias is the consequence of different answers to different survey modes from the same respondents.

**Findings**

- Employment and unemployment estimates for CATI and Web differ significantly from those in CAPI. There is no dominant component; coverage, nonresponse and measurement effects all play a role. The exceptions are the web coverage and nonresponse effects for employment; Web attracts significantly more employed respondents.

- Mode effects for educational level indicate that some categories are difficult to classify by respondents and/or interviewers. This links to findings in WP II that for educational attainment questions the visual design and the presence of more than
just one question on the screen appeared to be essential to give the respondent
the right context to answer in web.

- Selection effects with respect to CAPI can be explained to a large extent by
standard registry variables of Statistics Netherlands. Measurement effects for both
CATI and web are not significantly different from zero at the 5% level but predict a
downward adjusted mode effect for unemployment. This prediction is confirmed
by an analysis of LFS parallel runs.
- For some sub-groups within the population, the differences are quite large
- There are strong indications that it may become problematic to compare groups.
- There is no clear indication for social desirable answering in interviewer-
administered surveys; the proportion of respondents looking for a job is higher in
Web. The proportion of respondents that is available for a job is, however, lower in
the Web. This may also be the result of interviewer assistance in determining
availability, or of visual answering options of this question in web. See for example
the section on critical variables in WPII.

### Cautions

- The LFS mode effect estimates are based on an experiment linked to the CVS. It
may be debated whether mode effects on LFS variables are the same whenever
they are asked in the CVS as opposed to the LFS itself.
- The mode effect estimates apply to the Dutch context with a high proportion of
households that have access to web and a relatively low percentage of households
with a registered phone number. These two limitations imply that results should
be translated with care to other countries.
- It is extremely costly to experimentally determine selection and measurement
effects on unemployment, because of the precision that is necessary. Even the
large scale experiment fielded by Statistics Netherlands lacked the power for
determining measurement effects in relatively rare phenomena.
- The registry information that corrects selection effect in the Dutch situation may
not be generally available to other NSIs.

### Recommendations

- The estimates for the survey questions underlying employment status (having a
job, wanting a job, seeking a job and being available) show measurement effects. The
questionnaires should be evaluated and possibly adapted in order to reduce
these differences between modes.
- Parallel runs are needed to assess mode effects that still exist after nonresponse
adjustment. Findings should be used as input to an optimization of the LFS design.
- Adaptive survey designs show promise in minimizing measurement effects under
realistic, practical constraints on costs and comparability, especially when they
differentiate the survey mode.
- The LFS standard adjustment for coverage and nonresponse is effective in
removing selection effects between modes in the Dutch situation.
- There is no indication that measurement bias is strongly linked to contact effort.
That means that variation in time is likely to lead only to a varying mix of
measurement biases per mode. This mix can be stabilized to a large extent by
calibration to a fixed distribution of modes in the response.
The last two recommendations need a replication of findings in other EU countries in order to generalize them.

### 2.4 Web pilot study of the Finnish Labour Force Survey

#### Method

In an experimental design, CAWI was compared to the traditional CATI questionnaire of the Finnish LFS. 2345 CAWI responses were available for comparison with 2570 CATI responses.

#### Conclusions

There were no significant mode effects on employment status between CATI and CAWI. The results are encouraging for further development of mixed mode LFS.

#### Findings

- CAWI produces higher estimates for the number of employed persons, although these are not statistically different
- The estimate for unemployment also is higher in CAWI, but again not statistically different
- Not-employed people in CAWI answer more often that they have been looking for a job. Whether these differences are caused by selection bias or by measurement bias remains to be ascertained.
- CAWI respondents reported a lower number of hours worked in the reference week. The difference could be diminished by adapting the question in CAWI, and thus essentially offering a different question in the two modes.

#### Cautions

It is unclear whether the lack of significant differences is a result of actual lack of differences, or a lack of statistical power. More experiments are clearly needed.

#### Recommendations

In order to secure equivalent data quality across modes, it may be necessary to sometimes use different questions in different modes, that is, not use a unimode approach, but a functional equivalence approach.

### 2.5 The impact of survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey.

The final paragraph in this chapter on mode effects describes the deliverable by Beukenhorst, Buelens, Engelen, van der Laan, Meertens and Schouten (2014), on the interaction between survey item characteristics and survey mode. The goal of this
deliverable is to better understand the relation between type of person, nature of the survey question, mode dependent nonresponse and answering behaviour.

Mode-specific measurement bias can be a major component to mode effects; for the Crime Victimisation (CVS) survey items it turned out to be the dominant component (in contrast to the LFS, where selection and non-response effects were dominant, see paragraph 2.3). In order to achieve comparability over time and between population subgroups that respond to different modes, it is imperative that mode effects can be kept at the smallest possible level. The key to minimal measurement bias lies in questionnaire design for multiple modes. However, even with careful questionnaire design, mode-specific measurement biases occur. In order to be better able to predict the risk of such biases in advance and to be better able to perform questionnaire design to multiple modes, it would be very helpful to know how such biases relate to the type of person and type of survey item. If such dependencies can be detected and explained, then future questionnaire design may profit from this knowledge and be more effective in ensuring comparability.

**Method**

CVS questions were coded with a classifying system with a strong resemblance to Saris and Gallhofer (2007) and Campanelli et al (2011). Respondent answers were coded for response styles like don’t know (DK) answers, primacy, recency, extreme response style and straightlining. The dependencies between measurement bias, type of person and type of survey item are modelled by multilevel models. The survey items form a level within individuals, so that it is possible to investigate between survey item variation and between individual variation.

**Findings**

- It appeared that don’t know (DK) answers are a prominent source of differences between survey modes. In the CVS, DK answers were not explicitly offered in the interviewer-assisted modes. In the non-interviewer modes they were explicitly offered for around 80% of the items that are posed to all respondents. DK answers indeed make up a considerable part of the total difference between modes.

- The item level analysis did not reveal strong predictors for the occurrence of DK answers from the typology due to the limited number of survey items.

- The category level analysis showed large differences between questionnaire sections and a relatively large between-individual variation. The latter implies that DK answers cluster within individuals.

- The analysis of the other response styles, primacy, recency, extreme response style and straightlining, only partially confirmed hypotheses in the literature: Respondents do not show primacy effects for the non-interviewer modes, but recency effects occur more often for interviewer modes.

- Respondents also show more extreme response style in the interviewer modes. Hence, the interviewer modes show more differentiation of respondent answers over item categories. This finding does not clearly confirm hypotheses in the literature.
• Straightlining occurs more often in telephone interviews and increases in frequency for non-interviewer modes during the questionnaire. Taken together with the DK answers, the various response styles support the observed large differences between the modes.

<table>
<thead>
<tr>
<th>Cautions</th>
</tr>
</thead>
<tbody>
<tr>
<td>When generalizing results to other surveys, care is needed. First, the set of survey items in the CVS is limited, but more importantly is not a random, arbitrary set of items. The results in this paper cannot be applied to surveys with completely different profiles of survey item characteristics. Second, simultaneously occurring response styles may mask each other; A respondent that provides a lot of DK answers and that otherwise chooses the first answer category may not seem to show a primacy effect. Still, it is believed that some of the findings translate to other, similar surveys.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presentation and analysis of DK answers needs careful consideration; it plays a dominant role in differences between modes. The results indicate that rating scale questions lead to a different mix of straightlining and non-differentiation in different survey modes, which affects especially the reliability of underlying scales and factors. These effects should also be accounted for in questionnaire design.</td>
</tr>
</tbody>
</table>

2.6 Overall conclusions mode effects

There is evidence from this ESSnet and other findings that measurement errors are an important source of differences between modes in some but not all surveys. For the LFS the differences between modes (mail, web, cati, capi) can mostly be explained with common weighting variables. This finding is replicated in the Netherlands, Germany and Finland. However, this is not always the case: large mode effects in the Dutch Safety Monitor have led to a restriction in the modes to web and paper, that is, non-interviewer modes (Schouten et al., 2013). Likewise, in the Finnish Consumer Sentiments Survey large mode differences led to the decision not to introduce web data collection in this survey (Pohjanpää, 2013). Research on the British Opinions Survey showed that mode effects can be explained, but in order to do that, additional auxiliary variables are necessary, more than are standardly available for weighting, see chapter 4.

<table>
<thead>
<tr>
<th>Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>These findings lead to the following recommendations: tribe</td>
</tr>
</tbody>
</table>
• Instead, we need to develop rules of thumb for choices in the survey design: is every sample unit subjected to the same modes, will we be able to adjust afterwards, can we stabilize findings. These choices depend on the relative amount of selection and measurement effects, and on the choice of benchmark (e.g., CAPI, or a mix of modes). These rules of thumb, in the form of a decision tree, need to be developed by methodologists and fieldwork experts.

• Whether or not adjustments can be made for mode effects is dependent on the availability of relevant auxiliary variables.
3. Estimation under multiple mode data collection

When the composition of the response obtained in a mixed-mode data collection strategy can fluctuate – between subpopulation or between time periods – comparability of the resulting statistics may be compromised. If the modes exhibit relative measurement errors for a survey variable, then these measurement errors constitute part of the final estimate for that variable.

Figure 4.1 below shows the relative contribution of modes in a 19 month period of the first wave of the LFS response in the mixed mode design of Statistics Netherlands. The modes that are used in the Dutch LFS are CAWI, CATI and CAPI. Since there are no procedures in place to enforce fixed shares of the modes in the final response, there is monthly variation in the obtained response mode composition. If measurement errors are at play, the variations in mode composition will lead to a variation in total bias in the estimates. Consequently, the variations seen in the mode composition will manifest themselves as fluctuations in resulting LFS estimates, a situation that is to be avoided. The deliverables on data processing and estimation, written by Salah Merad (UK) and Bart Buelens (NL) investigate the effect on the estimates of stabilizing the mode composition.

![Response mode composition in the monthly Dutch LFS samples (weighted).](image)

A number of methods have been developed to make adjustments through weighting. Examples are propensity score weighting and propensity matching (Schonlau et al., 2009) and calibration (Bethlehem, 2009; and Duffy et al., 2005). In both methods there is a need to have available appropriate auxiliary variables for weighting: the auxiliary variables need to be correlated with response propensity and the target variables of the survey. For more details, see section 3.4 in Betts and Lound (2010).

With regard to mode-dependent measurement errors, although a large body of literature is available on quantifying and explaining mode effects - see, for example, (Vannieuwenhuyze and Revilla, 2013; and Schouten et al., 2013) - not many
approaches to correct for these effects have emerged. Two approaches that have appeared recently are Suzer-Gurtekin et al. (2012) and Buelens and van den Brakel (2014). Both methods address issues related to comparability and stability of estimates based on mixed-mode surveys in scenarios where the response mode composition of the overall response may fluctuate, either between publication cells or between editions of a survey. Another approach, developed in Kim (2013), attempts to adjust returned values at unit level for continuous variables, assuming that the measurement error results in no bias but increases the variance. Kim’s approach is based on random effect statistical regression models and the evaluation of conditional probability distribution functions. He presented an explicit solution when the model error terms are Normal and an iterative numerical method, based on parametric fractional imputation (Kim, 2011), when the error terms are not Normal.

This report is focused on adjustment methods for mode-dependent measurement errors and contains two contributions: one by CBS, which applies the approaches presented in Buelens and van den Brakel (2014) and Suzer-Gurtekin et al. (2012) to Dutch LFS data, and another by ONS, which attempts to extend the approach in Kim (2013) to categorical variables where one of the modes suffers from measurement bias, and applies the method to data collected in a pilot where a web sample was run in parallel with the main face-to-face ONS Opinions Survey in 2010. In the CBS contribution the objective is to stabilise the potential mode-specific measurement error bias in the estimates over time and across subpopulation groups, whereas the ONS contribution attempts to remove any potential measurement error bias in the estimates of all subpopulation groups.

The CBS contribution is concerned with the situation when the composition of the response obtained in a mixed-mode data collection strategy can fluctuate – between subpopulations or between time periods. The comparability of the resulting statistics may be compromised if the modes exhibit relative measurement errors for a survey variable: these measurement errors constitute part of the final estimate for that variable. Variations in response mode composition will manifest themselves as undesired variations in the survey estimates. Since the first wave of the Dutch LFS is conducted using a sequential mixed mode design, the survey is exposed to the risk of undue variations in resulting survey statistics associated with variations in the composition of the realised response each month.

The CBS paper investigates the effect of varying response composition on the monthly unemployment estimates. Two adjustment methods are applied and compared. The first method involves the addition of calibration constraints: it adjusts the GREG weights in such a way that they simultaneously correct for selection effects and obtain a specific, pre-set response mode composition (Buelens and van den Brakel, 2014). This method is applied once and is valid for all survey variables; if the chosen levels deviate a lot from the actually obtained response mode composition, the method is inefficient and results in large variances. The second method leaves the GREG weights unaltered, but rather adjusts the survey answers of the respondents (Suzer-Gurtekin et al., 2012). Using a model, relative mode-dependent measurement errors are estimated, and used to predict respondents’ answers under alternative response modes. This method is
specific to the survey variable in question, but does not adversely affect the sampling variance. Additional variance may be introduced because of the uncertainty associated with the estimation of the measurement errors. A final estimate is obtained as a weighted average of the mode-specific predictions. Both methods assume a particular measurement error model and assume that the underlying variable to be estimated is independent of mode, conditional on the covariates in the model (i.e. that these covariates explain the selection effect). It is difficult to assess the validity of these assumptions without additional experiments.

While the two methods have different motivations, they give very similar results for the estimated number of unemployed based on the first wave of monthly LFS data when calibration proportions and mixing coefficients are equal. Furthermore, the two adjusted series are comparable to the original, regular series in use for official publication purposes.

In general, a study like that conducted in the CBS paper may reveal that survey statistics obtained through a mixed mode survey suffer from variability or instability due to variations in response mode composition. Even though the measurement bias in the survey estimates will not be removed, applying the adjustment methods studied in the CBS paper, or an alternative with similar effects, is recommended to keep the measurement bias under control.

The ONS paper describes an attempt to extend the methodology in Kim (2013) to categorical variables that are subject to a measurement error that can result in bias, with a specific focus on binary variables, in view of removing potential bias in estimates of all subpopulation groups. The proposed method involves replacing the observations in one mode by estimated values of what would be observed in a reference mode conditional on the observed values and other auxiliary information. A method for estimating the overall measurement bias is proposed; it is based on fitting logistic regression models and estimating counterfactuals. It assumes a reference mode that is not subject to measurement error and the availability of auxiliary variables on which mode selection is conditionally ignorable. The conditional distribution of the outcome under the reference mode, given the return under the alternative mode and auxiliary data, is defined in terms of two misclassification probabilities.

These probabilities could be estimated using an appropriate experimental design, although the estimation process could be complex. However, this can only be done once or at most from time to time, which could make the estimated probabilities out of date. In the ONS paper, an equation that relates the two unknown misclassification probabilities is derived. This equation is not sufficient by itself to obtain a unique solution, and so another criterion is needed. Variance minimisation is a possible additional criterion but it would only produce approximately unbiased estimates at all levels of aggregation if the optimisation is performed over a set of approximately unbiased solutions. A solution could be obtained from data collected in an appropriately designed experiment; when the main survey is run, other plausible neighbouring solutions could then be then constructed by fixing one probability and solving the proposed equation to compute the second probability.

In the paper, we applied the methodology to employment data from a split face-to-face and web sample design on OPN data, and used a number of socio-demographic
variables as model covariates. The estimated measurement error bias is thought to be rather too large, indicating that the set of covariates used in the models could not account for the effects of mode selection. This is a serious problem; therefore, we may need to adopt a design where mode selection is minimal by embedding an experimental design within the overall sample design. An example of such a design is the “dual frame” design that has been proposed in ONS for the LFS mixed mode pilot (see Fletcher et al., 2010): here a random sample of households are requested to register for LFS on the web, but a subsample of the sample are also selected for a face-to-face interview. This means that a subsample of web registrants will be approached for a face-to-face interview. This subsample (web registrants, face-to-face response) and the subsample of web registrants who are approached for a response on the web (web registrants, online response) can be used to estimate pure mode effects: there should only be selection effects due to non-response, which should be minimised with the use of auxiliary variables.

We have seen that methods for adjusting for measurement errors can be developed but they have limitations and rely on assumptions that are difficult to verify. The method in Buelens and van den Brakel (2014) shows promise and is relatively easy to implement, but it only addresses the problem partially since the bias is not removed, which may be a concern to producers of official statistics. It is only through careful questionnaire design and pre-testing of questionnaires in mixed mode settings that relative measurement effects can be avoided and reduced, and that NSIs may hope to produce estimates that are close to the truth.

### Recommendations

- This is an area where research has only just started, and further work needs to be done.
- Test the assumptions underlying the models.
- Try to prevent the necessity to adjust, by careful questionnaire design and pre-testing of questionnaires.
- Auxiliary data need to be available and of good quality. These can be registry data or paradata that resemble registry information.
- Be aware that time series will be compromised by the introduction of mixed mode data collection and accept that trend breaches will be a regular occurrence.
- Eurostat may play a role in making clear that breaches in time series are to be expected and not necessarily problematic.
4. References

4.1 References based on ESSnet-DCSS deliverables


4.2 General references


