Survey Research Methods (2010)
Vol.4, No.3, pp. 151-164
ISSN 1864-3361
http://www.surveymethods.org

Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach

Melanie Revilla
Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra, Spain

So far, most surveys used face-to-face or telephone questionnaires in order to collect data. But the costs of achieving a survey using these traditional modes increase. At the same time, the response rates decrease, making the idea of switching mode very attractive. Because each mode has its own weaknesses and strengths, the idea of mixing modes of data collection is becoming more and more popular. Nevertheless, combining different modes of data collection may be problematic if people answer differently depending on the mode. Also, a switch from a unimode to a mixed-mode design may threaten the comparability of the data across time. This paper focuses first on the selection effect and shows that different kinds of respondents answer in different modes: therefore, mixing modes might make sense since it may improve the representativeness of the sample keeping the costs low. It is still necessary however to guarantee that mixing modes would not threaten the comparability. Then, the paper therefore compares the quality of questions asked in a unimode and two mixed-mode surveys. Using data of the European Social Survey (ESS) in the Netherlands, and following a multitrait-multimethod approach (MTMM), few differences are found between the unimode and mixed-mode designs in terms of quality. Looking at the differences across modes lead to slightly less similarities, but overall the quality does not change much.

Keywords: Modes of data collection, concurrent or sequential designs, selection bias

1 Choosing a data collection approach

Each researcher designing a survey makes, consciously or not, a lot of decisions, about the formulation of the questions (e.g. introduction, exact wording) and their scales (e.g. number and order of response categories, middle point, labels, dont know option), but also about the sampling procedure (e.g. frame, population to be sampled, selection of the sampling units), and so on. All these decisions may impact the results and conclusions reached. One of these important decisions concerns the mode(s) of data collection. For a long time, few modes were available: surveys were done mainly by mail, face-to-face and later telephone interviews (de Heer, de Leeuw, van der Zouwen, 1999). In the last decades however, these modes of data collection have shown important limits: their costs increase whereas the associated response rates tend to decrease. Households with two working adults are becoming more and more frequent, such that it is harder and harder to get in touch with them. Besides, the development of entry codes and answering machines make it even more difficult to establish the contact with the sampling units, as well as the decrease of fixed-line telephone accompanying the increase in mobile phones, for which no sampling frames are usually available.

In parallel, the development of new technologies lets appear the possibility of using different modes of data collection, such as Web surveys. The Internet is more and more used by European citizens and offers an attractive alternative to the established modes of data collection: it may reduce the costs, shorten the fieldwork period, and offer more flexibility to the respondents, who complete the survey when and where they want.

But introducing new modes of data collection (for example Web) may threaten the comparability (across time, across groups) of the data, since the specific characteristics of each mode can both influence the choice of sampling units to participate and the way respondents answer the questions. Specific non-response and measurement errors may therefore be expected. Coverage and sampling errors may also vary depending for instance on the available sampling frames.

Concerning the decision of participation, one element to take into account is the respondents' access to each mode: not all sampling units have a telephone or Web access allowing them to complete a survey in that mode. A low coverage of the population of interest in one mode can be a barrier to the participation of some subpopulations. Besides, even if all units have access to each mode (e.g. the researcher provide them with an access in case they do not have it), still the willingness to participate of the different units may be influenced by the mode proposed, since depending on the mode and on how comfortable the units feel with using it, the amount of efforts needed to answer the survey changes. Hence, it is often argued that a "digital divide" exists (see e.g. Rhodes, Bowie, Hergenrather, 2003) and that new modes of data collection such as the Web incite more young people and more men to participate, and on the contrary discourage older people and women.

Contact information: Melanie Revilla, Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra, Spain, e-mail: melanie.revilla@hotmail.fr

Concerning the way respondents answer the questions, Tourangeau, Rips, and Rasinski, (2000) decompose the process of answering questions in four components: comprehension of the item, retrieval of relevant information, use of that information to make required judgments, and selection and reporting of an answer. All these components might be affected by the characteristics of the mode of data collection.

First, some of the modes are visual (e.g. mail, Web), others are oral (e.g. telephone), still others are simultaneously oral and visual (e.g. face-to-face using show cards). The comprehension of the item, most of all if the item is quite complicated, can be easier in a visual mode than in an oral one. On the contrary, if the reading skills of the respondents are limited, an oral mode may be more appropriate. Even if the direction of the effect is not obvious, at least the fact that the characteristics of the mode can impact the process to answer questions is clear. Also, to select and report an answer, respondents need to remember the possible response categories. When these categories are proposed visually to the respondents, memory is not an issue. But when the categories are proposed orally, a memory effect can be expected: mainly for long and complex scales, it is assumed that oral modes convey more recency effects whereas visual modes convey more primacy effects (Krosnick, Alwin, 1987).

Second, some of the modes require the presence of an interviewer (face-to-face, telephone) whereas others (mail, Web) are self-administered. Consequently, more social desirability bias (Krosnick, 1991, 1999) is expected in some data collection modes, due to this presence of an interviewer. Self-completed modes give also more freedom to the respondents (e.g. to choose the moment of the completion of the survey, to choose the space, to do several activities at the same time). As a result, Krosnick (1991, 1999) shows that distinct modes of data collection elicit varying levels of satisficing bias. The presence of the interviewer may also affect the comprehension of the questions: depending on the intonation used, on the words that are emphasized by the interviewer, a different understanding of the question is possible compared to the case where the respondent is let to itself. For complex questions, the interviewer can also provide clarifications or explanations that facilitate the understanding of the questions. In self-completed modes, such help is not so easy to implement, even if a question desk can for instance be organized such that respondents can call and ask questions.

Because the advantages and drawbacks of the different modes of data collection seem at least partly complementary, the idea of combining several modes is particularly attractive. In that way, the drawbacks of one mode could be compensated by the advantages of another. In particular, the coverage and non-response problems could be partially solved by mixing modes of data collection. For instance, the population with Internet access could be surveyed online and the population without Internet access by face-to-face. By adding a second mode of data collection, the costs would be reduced (compared to only face-to-face), and the response rates might increase (compared to only Web).

The mixed-mode literature is articulated around two main questions:

(1) "To mix or not to mix modes of data collection?" (de Leeuw, 2005)

(2) If we mix, how? Is there a more efficient way of mixing modes?

Concerning the first issue, Voogt and Saris (2005:385) advice to mix modes: they conclude that "a mixed mode design is an efficient way of fighting bias in survey research" since even if using different modes brought some response bias, the total bias stays lower than in a uni-mode design. On the contrary, Dillman et al. (2009) are more reluctant about mixing modes since they find that switching to a second mode of data collection is "not an effective means of reducing nonresponse errors based on demographics". Other authors do not answer either yes or no. They argue that mixing modes of data collection can reduce the costs, increase the response rates and even tackle specific sources of errors, but that at the same time it introduces other forms of errors (Roberts, 2007; Kreuter, Presser and Tourangeau, 2009). Therefore, "in mixed-mode designs there is an explicit trade-off between costs and errors" (de Leeuw, 2005:235) but also between different kinds of errors.

Concerning the second issue, there are many different ways to combine modes of data collection: "a distinction can be made between *multi-mode* and *mixed-mode* approaches. The former are where different modes are used for different sets of survey items, but each survey item is collected by the same mode for all sample members. The latter are where the same item might be collected by different modes for different sample members" (Lynn et al., 2006:8). So it is possible to use different modes of data collection at different stages of the data collection procedure, for instance sending first an advance letter, then making a phone call to recruit the respondents, and finally making an appointment with them in order to go to their house to do a face-to-face interview. This is a *multi-mode* design. It differs from what is called *mixed-mode* designs, i.e. designs where different modes are used at the same stage. It also differs from mixed-mode *panel* designs, where one mode is used at one point in time and another is used latter on (Dillman, Smyth, Christian, 2008). This paper focuses on mixed-mode designs and how to mix modes at the specific stage where respondents are effectively answering the questions. Usually, the mixed-mode approach is divided into two main designs: a concurrent (people are offered a set of modes and can choose the one they prefer) and a sequential one (people are first proposed to answer in one specific mode, if they refuse or do not answer, they are offered another mode, etc).

Previous research has compared sequential and concurrent designs both together and with a unimode design (e.g. Brambilla and McKinlay, 1987; Dillman, Clark and West, 1995; Shettle and Mooney, 1999; de Leeuw, 2005; Dillman et al., 2009). Nevertheless, most of the research has focused on a comparison of costs and of simplistic indicators of quality (response rates, variable distributions, social desirability and satisficing bias). But low response rates are only "a warning of potential trouble" (Couper, Miller, 2009:833) and higher response rates does not necessarily imply higher representativeness (Krosnick, 1999). Therefore, studying response rates is not enough to evaluate the quality. Similarly, measuring the quality by assessing the level of social desirability bias and satisficing (Dillman et al., 2008; Heerwegh, Loosveldt, 2009) is too restrictive since mainly adapted to some particular topics (e.g. sensitive topics as drug use). But little has been done yet on unimode and mixed-mode designs comparing other (more elaborated) indicators of the quality (Roberts, 2008).

Our study aims to address this gap, by comparing two mixed-mode designs with a unimode survey in terms of the quality of measurement, when the quality is defined as the strength of the relationship between the observed variable and the variable of interest, and can be computed as the product of the reliability and validity (Saris and Andrews, 1991). Defining the quality in that way presents the advantage that it allows to differentiate between random and systematic errors (sometimes referred to as “correlated errors”) and to correct for measurement errors (Saris and Gallhofer, 2007). The paper also has a second goal: determining if different kinds of respondents are reached when different modes and designs are used. If not, mixing-mode would indeed have little sense. This is therefore a preliminary condition to have an incentive to implement a mixed-mode survey.

It is important to notice finally that one cannot speak about “face-to-face surveys”, or “Web surveys”, as one unit. The term of “Web surveys” for example is too broad (Couper, Miller, 2009): two Web surveys can be as different as one Web and one mail survey, depending on several choices made (e.g. number of items by page, possibility to come back to previous questions, “don’t know” option proposed). The same is true for “sequential” and “concurrent” designs: depending on the particular procedure (e.g. number of modes, order in which they are offered, access provided when not present) two sequential (or concurrent) designs might differ a lot. Therefore, even if these general terms are used for the sake of simplicity, it is important to remember that what we are dealing with is one specific unimode face-to-face design, one specific concurrent design and one specific sequential design.

The exact design of the surveys playing a central role, section 2 gives more details about the data used in this study: the European Social Survey (ESS) round 4 (2008/2009) and the mixed-mode experiment implemented by the ESS (2008/2009). Then, section 3 conducts a preliminary exploratory analysis of these data, with the main objective of detecting whether different kinds of respondents are participating using different modes of data collection. If not, there is indeed no argument to mix modes of data collection; using only the cheapest mode is sufficient. Once established that it might make sense to use a mixed-mode design, section 4 refocuses the interest on the quality and presents the multitrait-multimethod approach used to get the reliability and validity estimates. The quality is obtained by taking the product squared of these reliability and validity coefficients. The results obtained by applying this method to the ESS data are exposed in section 5. Finally, section 6 discusses some limits and proposes ideas for further research.

2 The European Social Survey (ESS)

ESS round 4

The ESS is a biannual cross-national project designed to measure changing social attitudes and values in Europe.[1] An important effort is made to ensure the best possible quality of the data collected. Particular attention is given to the sampling procedure in each country in order to guarantee the “full coverage of the eligible residential populations aged 15+” (Lynn et al., 2007).

The ESS round 4 took place in around 30 countries between September 2008 and June 2009. We focus on one country, the Netherlands, because the mixed-mode experiment has been implemented there. The data of round 4 has been collected by face-to-face in the Netherlands: the interviewers went to the respondent’s home to administer a computer-assisted personal interviewing (CAPI). An important specificity of the ESS is the use of show cards providing visual help for the majority of the questions.

In average one interview takes around one hour. It contains a main questionnaire, administered to all the participants, and a supplementary questionnaire, composed of questions already asked in the main questionnaire but formulated in another way, i.e. using another method: for instance first a 6-point scale is offered and latter an 11-point scale. Theses repetitions are used to evaluate the quality associated to the different methods.

ESS mixed-mode experiment

Because of the increasing costs and difficulties to reach people using face-to-face data collection, the option of allowing some countries to switch in a near future to another mode or combination of modes of data collection is tempting. But if different modes of data collection lead to different answers, the comparability would be threatened. Therefore, studying first the different modes of data collection is necessary, which pushed the ESS to launch a series of research on mixed-modes, which is considered as the most realistic alternative to the traditional face-to-face design.

In parallel to the ESS round 4’s fieldwork a mixed-mode experiment was implemented in the Netherlands from November 2008 to July 2009. The country has been chosen because it is a good candidate for a switch in the data-collection approach. Indeed, on the one hand, the traditional data collection is becoming more and more problematic, as the response rates show: 67.9% in the first round and only 52.0% in the fourth. Even if the fieldwork period has been extended in the forth round, the response rate is almost 20% lower than the ESS objective. The ESS response rates however are still higher than the average response rate of surveys in the Netherlands, which is around 40% nowadays. Even if low response rates are not always an issue, such a decrease in response rates incites researchers to question the well-functioning of the current data collection approach. On

[1] Cf. http://www.europeansocialsurvey.org/

the other hand, the Netherlands beneficiate from a large Internet coverage (around 85%): introducing Web as a complement of the traditional face-to-face in that country could really make sense. Other countries of the ESS have similar profiles as the Netherlands, in particular the Nordic countries (Sweden, Denmark and Finland). They could also have been chosen for the experiment, whereas other countries on the contrary have much lower Internet coverage (30% to 45% for Greece, Bulgaria, Romania, Portugal, Lituania)[2] and seem less likely to switch data collection approach in the next years.

Telephone could also be introduced in complement or replacement of face-to-face, even if it may be more difficult to implement for such a long survey as the ESS. In particular, Nordic countries' high fixed-line telephone coverage, together with their experience of telephone survey, could be candidate for a switch to telephone interviews. The mixed-mode experiment considers therefore the three modes and compares a concurrent with a sequential design. In order to reduce the burden of the telephone interviews, respondents were able to do them in two parts (two interviews of around 1/2 hour).

As Figure 1 shows, the general design of the experiment is however more complex, since a separation is done between people with and without known phone number. This is because of the nature of the sampling frame and mode of contact used. The sampling frame consisted of postal addresses, but the contact was done when possible by telephone. So first the fieldwork agency matched as many addresses as possible to phone numbers: this corresponded to only 70%. These 70% were randomly divided in two groups: the first group was assigned to a sequential design (Web offered first, then phone, then face-to-face), whereas the second group was assigned to a concurrent design (choice between face-to-face, phone and Web). For the remaining 30% without known phone number, the contact was made face-to-face, and therefore, respondents were first proposed to do a face-to-face interview. If they refused, they were then offered sequentially Web and finally telephone.

The face-to-face (CAPI) version of the main questionnaire was the same in the mixed-mode experiment as in the ESS round 4. For the telephone (CATI = computer-assisted telephone interviewing) and Web-based (CAWI = computer-assisted web interviewing) versions, some changes were necessary in order to adapt the questionnaire to another mode.

Topics and methods analyzed

In order to compare unimode and mixed-mode designs, only the questions and methods shared by both surveys are used: usually, each experiment contains three common traits measured with the two same methods, except for the experiment about social trust. In that case, we have a smaller model with two traits (instead of three) and two methods. Table 1 summarizes the experiments analysed.

3 A preliminary observatory analysis of selection effects

The first goal of this paper is to compare unimode and mixed-mode designs in terms of quality. Nevertheless, the paper has a second goal: looking whether it makes sense to mix modes of data collection. The introduction showed that the literature often underlines the trade-offs that should lead the decision process. The paper does not aim to give a general answer to the complex question of whether one should or should not mix modes of data collection. However, this section's goal is to explore with the ESS data one important point to consider when deciding to mix or not to mix: does one gain something by adding modes?

Mixing modes is a quite complex approach that requires more work to prepare the survey (adapting the questionnaires, training interviewers to different modes), sometimes to implement it (following of the respondents' decisions across different modes), and finally to analyse it (harmonisation of the data). So it is necessary that a mixed-mode approach also allows gaining something, otherwise it does not make sense to implement it: the extra difficulties due to mixing modes need to be balanced by extra opportunities. The attractiveness of mixed-mode approaches is principally based on the idea that in such approaches the drawbacks of one mode can be compensated by the advantages of another. In particular, it is often argued that Web surveys have lower costs but are less representative of the general population, whereas face-to-face surveys are more expansive but lead to more representative samples. By mixing modes, a better representativeness can therefore be achieved at lower costs, if different kinds of respondents are reached by different modes. If this is not the case, the cheapest mode could as well be used alone, since the sample would be as representative with reduced costs.

These preliminary analyses of the data focus on this last point and try to look at potential differences in respondents' profile and how this can affect respondents' choice of participation (participate or not) and mode of participation (if participate, in which mode).

Differential preference and tolerance of modes due to gender and age

The first question is: are the different modes chosen? If respondents all choose the same mode, then it is not useful to propose additional modes. To answer that question, Table 2 gives the repartitions of CAPI, CATI and CAWI interviews for the following groups: concurrent, sequential, unknown phone number, all respondents from the mixed-mode experiment and ESS round 4.

It shows that the total number of respondents is very similar in the concurrent and sequential designs. Knowing that the initial sample sizes were identical, it means that the response rates in these two groups are very similar: 45.0% for the sequential group and 45.9% for the concurrent one. Moreover, the table shows that people with known telephone

[2] See for instance Eurobarometer 71.2 2009

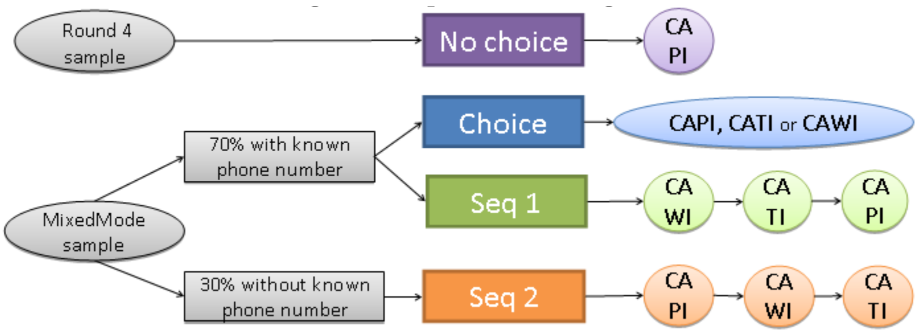

Figure 1. Experimental design

Table 1: Questions and methods analyzed

Experiments	Wording of the questions	M_1	M_2
Media	On an average weekday, how much time, in total: – do you spend watching television? – do you spend listening to the radio? – do you spend reading the newspapers?	8 points	Hours and min
Satisfaction	How satisfied are you with: – the present state of the economy in NL? – the way the government is doing its job? – the way democracy works?	11 points (extreme)[a]	11 points (very)[b]
Social trust	– Generally speaking would you say that most people can be trusted or that you can't be too careful in dealing with people? – Do you think that most people would try to take advantage of you if they got the chance or would they try to be fair?	11 points	6 points
Political trust	How much do you personally trust each of the institutions: – Dutch parliament – The legal system – The police	11 points	6 points

[a] "extreme"= extreme used in the labels of the end points
[b] "very"= very used in the labels of the end points

number choose in majority Web interviews, then, face-to-face, and finally telephone. There are only a few more Web interviews in the sequential design compared to the concurrent design. There are no real differences in terms of modes repartition between these two designs: this is probably linked to their quite similar implementation in practice. All three modes are chosen by a significant number of respondents. The group with unknown phone number on the contrary is very different, with mainly face-to-face interviews. This is linked to the facts that it is a group with different characteristics, that the mode of contact changes (face-to-face contact) and that the sequence in which the modes are proposed also varies. Only two respondents in that last group did a telephone interview: proposing telephone to that group is not useful.

Table 3 goes a step further and considers the question: are different modes chosen by different respondents? The table gives the distributions in terms of gender and age of the respondents that did a CAPI, CATI or CAWI interview, in the two principal groups (concurrent and sequential).

It seems that depending on their gender and age, respondents are more willing to participate in one or another mode. Looking at the concurrent group gives us an idea about the preferences of respondents to the different modes, assuming that they usually choose the mode they prefer as mode of participation. Thus, 52.8% of the male respondents decide to participate in a CAWI interview, whereas only 38.0% of the women do so. On the contrary more female respondents choose CAPI and CATI. A Kolmogorov Smirnov test indicates that there are significant differences in the distributions of modes by gender for this concurrent group. Looking at age, between 55 and 60% of the respondents aged from 20 to 64 take CAWI, against 16.9% of the 65-79 and 0% of the 80 and more. This last group chooses principally CAPI (61.9%), whereas the 65-79 choose more CATI (45.8%). Again, testing for significance in difference in distributions for the mode variable by age groups leads in most of the cases to rejecting that the distributions are equal. So, different age groups have

Table 2: Number (and percentages) of observations by mode and design

	Concurrent	Sequential	Unknown phone no.	Total mixed-mode	ESS round 4
CAPI	114 (31.1%)	103 (28.4%)	226 (85.0%)	443 (44.5%)	1778 (100%)
CATI	90 (24.5%)	88 (24.2%)	2 (0.7%)	180 (18.1%)	0 (0%)
CAWI	163 (44.4%)	172 (47.4%)	38 (14.3%)	373 (37.4%)	0 (0%)
Total	367 (100%)	363 (100%)	266 (100%)	996 (100%)	1778 (100%)

Table 3: Repartition of the respondents in the concurrent and sequential designs by gender and age categories (rows and columns percentages)

		Concurrent group only				Sequential group only			
		CAPI	CATI	CAWI	Total	CAPI	CATI	CAWI	Total
Gender	Male	27.0	20.1	52.8	100	24.7	24.1	51.3	100
		37.7	35.6	51.5	43.3	36.9	42.1	45.9	42.4
	Female	34.1	27.9	38.0	100	31.1	24.4	44.5	100
		62.3	64.4	48.5	56.7	63.1	57.9	54.1	57.7
	Total	31.1	24.5	44.4	100	28.4	24.2	47.4	100
		100	100	100	100	100	100	100	100
Age	16-19	60.0	10.0	30.0	100	23.1	7.7	69.2	100
		5.3	1.1	1.9	2.7	2.9	1.1	5.2	3.6
	20-39	23.6	20.8	55.6	100	28.9	8.4	62.7	100
		15.0	16.7	24.6	19.7	23.3	8.0	30.2	22.9
	40-64	25.6	15.6	58.9	100	24.7	25.3	50.0	100
		40.7	31.1	65.0	48.2	40.8	48.9	49.4	46.7
	65-79	37.3	45.8	16.9	100	36.5	28.4	35.1	100
		27.4	42.2	8.6	22.7	26.2	23.9	15.1	20.4
	>80	61.9	38.1	0	100	30.4	69.6	0	100
		11.5	8.9	0	5.7	6.8	18.2	0	6.4
	Total	30.9	24.6	44.5	100	28.4	24.2	47.4	100
		100	100	100	100	100	100	100	100

different preferences in terms of modes. In particular, there is a clear distinction between the elder and the rest. A unimode survey proposing only CAWI would therefore probably underrepresent elder respondents, which can bias the results if they have different opinions or attitudes than younger respondents. This is all as expected. As expected also, when looking at the sequential design, where Web is offered first, the percentages of people doing a Web interview is almost always higher. In that case, the percentages, more than a preference for a specific mode, can be seen as a tolerance to a certain mode: if respondents "tolerate" a Web interview, then they will accept it, even if this is not their preferred mode.

Two figures are however surprising. First, if the tolerance of the 16-19 for CAWI is very high (69.2%), their preference for that mode is quite low (30.0%). But this may be due to the very small sample size of this group. Second, in the group of the 40-64, the percentage of respondents doing CAWI is almost 9% higher in the concurrent than in the sequential group. This may be partially due to random errors (the concurrent and sequential groups can be different just by chance), but the difference is quite high to just result from hazard.

Differential access to modes

A crucial element ignored so far is that all people do not have access to all modes. Assuming that people choose in a concurrent design the mode they prefer is therefore too simplistic: they choose the mode they prefer *given* the list of modes they have access to. The choice is conditional on having access to the modes. Even in the Netherlands, still 15% of the population does not have a Web access. The telephone access also, even if very high, is not complete. Some surveys are correcting for these potential coverage biases by providing the respondents willing to participate with an access to the mode chosen or assigned[3], but this is not the case in the ESS mixed-mode experiment. It is therefore interesting to have a look at the telephone and Internet coverage in our data. Table 4 gives this information both when dividing respondents by modes and by designs.

There is a relatively large percentage of respondents interviewed by face-to-face (20 to 28%) that do not have either a fixed-line telephone and/or Internet access. Concerning the group interviewed by CATI, obviously few do not have a fixed-line telephone but more than 23% do not have Internet access. One would expect even nobody in this group not to have fixed-line telephone since all did a telephone interview,

[3] See for instance the LISS panel: www.centerdata.nl/en/MESS

Table 4: Percentages of respondents without fixed-line phone or Internet access

	CAPI	CATI	CAWI	Concurrent	Sequential	Unknown phone no.	Total MM	ESS round 4
No fixed-line telephone	27.6	3.3	4.3	5.5	5.8	38.7	14.5	15.0
No Internet access	20.3	23.3	0	14.2	13.8	11.3	13.3	13.7

but some people may have used a mobile phone to answer the interview. Besides, the ESS question used to obtain the numbers in Table 4 asks about having access to a "fixed-line telephone in the accommodation". Some people therefore may have a fixed-line telephone access somewhere out of the accommodation. On the contrary, people that did CAWI interviews have usually both Internet and telephone access. Looking at the designs, sequential and concurrent groups are very similar, with around 5% of their respondents that do not have a fixed-line telephone access in their accommodation and around 14% that do not have Internet access. This similarity is not surprising since the groups were randomly drawn, but different selection biases could have produced differences. The total mixed-mode data shows a similar pattern as the ESS round 4.

In brief, one could say that while people completing a CAWI interview could have done a CAPI or CATI interview as well, more than 20% of the respondents who did a CAPI or CATI interview could not have done a CAWI one. Within the group of respondents for which we said before that they "prefer" CAPI, one part had in fact no other choice: a bit less than 5% of the CAPI respondents do not have access to both telephone and Internet. For these 5%, more than a preference, doing CAPI indicates the absence of choice. But most of the respondents have some choice, even if the options may be reduced to two instead of three. Table 4 shows also that the coverage in fixed-line telephone and Internet is overall quite high, offering real alternatives for the traditional face-to-face, at least in the Netherlands.

One can also look at the telephone and Internet coverage by gender and age. The idea is to see for instance if the higher number of men in CAWI is related more to higher Internet coverage in this group than to a higher preference of men for answering a survey in this mode.

Table 5 shows that if the repartition of men and women *not* having telephone and Internet access is very similar, the repartition by age categories is changing: 31% of the 20-39 years old do not have a fixed-line telephone, against only 4% of the >80 years old. On the contrary, almost all young people have Internet access (except 3 or 4%) whereas a lot of the older respondents do not (almost 30% of the 65-79 years old and 70% of the >80). Therefore, variations in terms of age repartition depending on the mode of data collection as observed in Table 3 are probably influenced by the variations in telephone and Internet coverage of the different age groups.

What determines the mode of interview?

The analyses presented so far explore the idea that respondents' choices of participation in one mode depend on their gender, age, and their access to the different modes. The design (concurrent or sequential) may also play a role. To conclude with these preliminary analyses, a multinomial logistic regression with the mode of interview as dependent variable and the list of variables just mentioned as independent variables is run. Our dependent variable takes three values: CAPI, CATI and CAWI. CAPI is used as base outcome: since it is the established mode, it seems reasonable to take it as the reference with which the two others are compared. The independent variables are all dummy variables (with value 1 if the respondent is a woman, has access to a fixed-line telephone, has access to Internet, and is in the sequential group) except age that is continuous. The regression does not include the unknown telephone number group. Table 6 gives the coefficients of this regression: basic coefficients and coefficients expressed on the odd ratios scale.

Table 6 shows that the probability of choosing CATI versus CAPI increases with age, access to a fixed-line telephone in the accommodation and Internet access at home or at work. The gender and the design on the contrary do not significantly change the probability of participating by telephone instead of face-to-face. Looking at CAWI participation, the design again is not significant, which is as already mentioned probably at least partially due to the way the designs were implemented: in practice it seems they were not as different as they were supposed to be in theory. The probability of choosing a Web interview instead of a face-to-face one decreases significantly for women and older respondents but increases for respondents with fixed-line telephone and Internet access.

The size of the effects is higher for the access variables than for the personal characteristics of the respondents. For instance, having a fixed-line telephone versus not having it multiplies by 6.07 the odd ratio of choosing CATI instead of CAPI, and having Internet access by 1.93, whereas the odd of choosing CATI compared to CAPI increases by a factor of only 1.02 for each year age increases, controlling for other variables in the model. However, this difference has to be put in perspective. A one year change in age may not be the most pertinent change to consider: a 10-year might already be more interesting. Being 10-year older multiplies the odd by 1.22. Being 20-year older multiplies it by 1.49; being 30-years older by 1.81; and being 40-years older by 2.21. Therefore a 40-years change has a bigger impact on the odd ratio of choosing CATI and not CAPI than having Internet access. The importance of age should not be underestimated because the odd ratio is very close to 1. In the CAWI versus CAPI comparison nevertheless the access variables are really much more important than the personal characteristics variables.

To summarize, the probabilities of participating in different modes vary with the gender and age of the respondents, but also their access to telephone and Internet. So, differ-

Table 5: Non coverage by gender and by age for the mixed-mode experiment respondents (in percentages)

	Gender		Age				
	Men	Women	16-19	20-39	40-64	65-79	>80
No fixed-line telephone	16.9	12.6	6.7	31.4	9.5	5.1	4.3
No Internet access	13.0	13.4	3.3	3.6	7.8	29.2	68.1

Table 6: Multinomial logistic regression of the mode of interview

	Mode	Coefficient	Odd
CATI (versus CAPI)	Woman	-.05	.95
	Age	.02*	1.02*
	Access_tel	1.80*	6.07*
	Access_int	.66*	1.93*
	Sequential	.10	1.11
	Constant	-3.56*	
CAWI (versus CAPI)	Woman	-.64*	.53*
	Age	-.02*	.98*
	Access_tel	2.67*	14.52*
	Access_int	21.11*	1.47e+09*
	Sequential	.19	1.21
	Constant	-21.59	

Number of observations = 730

Pseudo R^2 = .15

p<.05 indicated by a star

ent modes of data collection allow getting somehow different kinds of respondents: one of the main arguments in favour of mixing modes seems to be verified in our data, at least for the few variables that have been considered. We focused on gender and age as two important determinants of mode's choices but more background variables could be analysed.

This section tried to provide some evidence that mixing modes of data collection may present some advantages, and therefore that it may constitute an attractive alternative to a unimode design. But showing that mixed-mode is an attractive approach is not enough to make the decision of using such a design. If the data has been collected using a unimode design in the past, as it is the case for the ESS, another important issue is to determine if switching from a unimode to a mixed-mode design will not threaten the comparability of the data across time. If the switch is implemented in some of the countries but not all of them (for instance in countries with high Internet coverage only), cross-national comparisons may also be threaten by a change in the data collection approach. The next sections focus on that question of comparability, and assess for one specific indicator, the quality of the questions, if there are significant differences between unimode and mixed-mode designs.

4 Estimation of the quality

How should we combine the groups?

In the mixed-mode experiment design, the group without known phone number, which represents 30% of the total sample, is treated separately. So we cannot compare directly the concurrent and sequential groups to the ESS round 4. What we are really interested in is to compare what can be called the "complete designs": designs that consider the total population. So, the 30% group of sampling units without known phone number should be combined to the 35% of the concurrent and the 35% of the sequential designs.

This combination can be done in several ways. Lynn, Revilla and Vannieuwenhuyze (forthcoming) choose to add the whole group of respondents without phone number to each of the other two groups but using weights of 1/2 in order to avoid a too important overrepresentation of this group. We follow another approach in this study because the big overlap between groups created by adding the whole group of unknown phone number to the concurrent and sequential groups may generate more similarities than one would have if really collecting the data using a complete concurrent or sequential design. So we create a dummy variable (*"randomsplit"*) which takes the value one if the respondent is in the concurrent group, zero if he/she is in the sequential group.[4] Then, we randomly split the group of unknown telephone number into two halves: the first half gets a value of one for the variable "randomsplit", whereas the other half gets a value of zero. Finally we compare three groups: the "concurrent" group (which corresponds in fact to randomsplit = 1), the "sequential" group (which corresponds in fact to randomsplit = 0) and the ESS round 4 (unimode face-to-face).

[4] Before going on, a check for outliers was done. In the media experiment, respondents have to give in hours and minutes (M_2) the time spent on three media. If the sum of the three activities' time is superior to 24 hours or if the time of one activity is higher than 20 hours, we consider the observation as an outlier. Because few outliers (four) were detected, we dropped from the dataset these four outliers.

Analytic method: the multitrait-multimethod (MTMM) approach

The quality is computed as the product of reliability and validity: ${q_{ij}}^2 = {r_{ij}}^2 * {v_{ij}}^2$. In order to get the reliability and validity coefficients (i.e. r_{ij} and v_{ij}), the data is analysed using an MTMM approach, which consists in repeating questions (called "traits") in several ways (i.e. with several "methods"). Proposed first by Campbell and Fiske (1959), the approach has been used later with structural equation models (Werts and Linn, 1970; Jöreskog, 1970; Alwin, 1974) and applied to single questions (Andrews, 1984). Three is usually the minimum number of methods needed in order to avoid identification issues. In our case, we have only two methods for each of the traits. However, doing a multi-group analysis with constraints of invariance of the parameters across groups allows identifying the model.

Each experiment is studied separately. Figure 2 shows the model used for six variables. It contains three correlated traits (F_1, F_2 and F_3), each measured with two methods (M_1 and M_2). It is assumed that the methods are not correlated with each other, nor with the traits, and that the effects of the methods on the different traits are the same ($m_{11} = m_{12} = m_{13}$ and $m_{21} = m_{22} = m_{23}$). This leads to six true scores T_{ij} ($i = 1, 2, 3$ and $j = 1, 2$). The true scores correspond to the systematic components of the observed variables Y_{ij}, i.e. once random errors e_{ij} have been corrected. The random errors are not correlated with each other, neither with the traits. The strength of the relationship between the true scores T_{ij} and the observed variables Y_{ij} is the reliability. The strength of the relationship between these true scores T_{ij} and the variables of interest F_i is the validity. Only the first observed variable is represented in Figure 2 for clarity purpose but there is in fact for each true score a corresponding observed variable.

More formally, the model, called True Score model, can be described by the following system of equations (Saris and Andrews, 1991; Saris and Gallhofer, 2007):

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad \text{for all} \quad i, j \tag{1}$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad \text{for all} \quad i, j \tag{2}$$

Where, for the i^{th} trait and the j^{th} method: Y_{ij} refers to the observed variable, r_{ij} to the reliability coefficient, T_{ij} to the true score, e_{ij} to the random error component associated with the measurement of Y_{ij}, v_{ij} to the validity coefficient, F_i to the trait, M_j to the variation in scores due to the method, and m_{ij} to the method effect coefficient.

The estimates of reliability and validity coefficients are obtained from the Lisrel output analysing the covariance matrices by Maximum Likelihood estimation in a multiple group context.[5] We have three different groups which correspond to the different designs. We test the null hypothesis that there are no significant differences in terms of reliability and validity across groups. In order to do so, the parameters are first specified as invariant across groups. The fit of the model is then tested using the procedure proposed by Saris, Satorra and Van der Veld (2009), which has the double advantage to take into account the power and so type I and II errors, and to provide a test at the parameter level (by opposition to chi-square for example that tests the complete model). Therefore, using the JRule software based on this procedure (Van der Veld, Saris, Satorra, 2009) information about potential misspecification of each parameter is obtained: this provides guidelines on how to correct the initial model when necessary. Corrections are introduced step by step till an acceptable model is obtained.[6]

5 Main findings

Comparison of the quality estimates by designs

Table 7 gives in each experiment the reliability and validity coefficients and the quality for each trait and method for the different designs: unimode face-to-face, concurrent ("*randomsplit*=1") and sequential ("*randomsplit*=0") mixed-mode. When different groups have equal estimates, they are grouped in a same row. The last column gives the mean quality of the three (or two for social trust) traits.

In the social trust experiment, the quality is the same for all three designs. In the experiments about media and political trust, concurrent and sequential designs lead to the same coefficients. The difference is between the unimode face-to-face design on the one hand and the mixed-mode designs on the other hand. Nevertheless, the variations between unimode and mixed-mode designs are quite small. If we consider the average quality of the three traits, the highest difference is 0.07 in the political trust experiment (cf. Table 8 for a clearer picture). In the experiment about satisfaction, not only the unimode is different of the mixed-mode designs, but also the concurrent and sequential approaches have different quality estimates. The difference is mainly coming from variations in validities, even if some reliability estimates do vary too. Since validity ${v_{ij}}^2 = 1 - {m_{ij}}^2$ (where ${m_{ij}}^2$ is the method effect), the lower quality for example of the second method in the satisfaction experiment seems to result from higher method effects. One interpretation of this would be that the impact of using "very" in labelling the end points of the scale on the respondents' answer is more important in telephone and/or Web than in face-to-face: it leads to more systematic errors. But once again, the overall differences in quality are small (maximum 0.06).

In fact, much more differences are found between the quality estimates of different methods: comparing the mean quality, a difference of more than 0.2 points separate methods one and two in the media as well as in the political trust experiment. This is much higher than the differences across

[5] An example of Lisrel input is available online: http://docs.google.com/Doc?docid=0AbQWMcvxT-2KZGQ3Mm10MzRfMTY1ZGN0YjZtY3Q&hl=en

[6] A list of the adaptations of the initial model done for each experiment can be found online: http://docs.google.com/Doc?docid=0AbQWMcvxT-2KZGQ3Mm10MzRfMTY2YzZncjdzZmY&hl=en

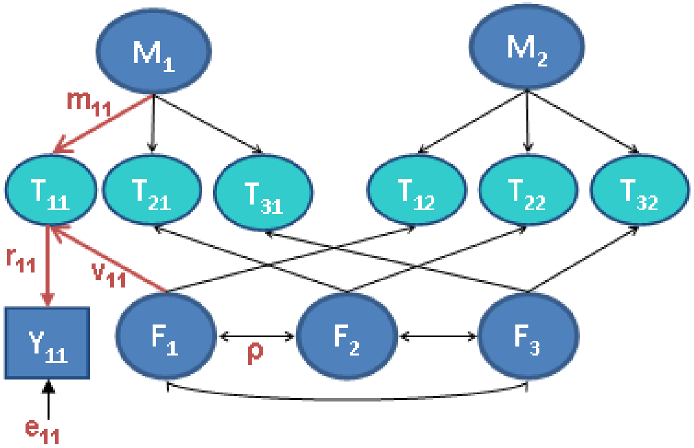

Figure 2. The MTMM model for 3 traits and their repetitions

Table 7: Estimates in the different designs[a]

Experiments	Group	Method	r_{1j}	r_{2j}	r_{3j}	v_{1j}	v_{2j}	v_{3j}	$q^2{}_{1j}$	$q^2{}_{2j}$	$q^2{}_{3j}$	$q^2{}_{mean}$
Media	ESS 4	8 points	1.00	.79	.91	.98	.97	.98	.96	.59	.80	.78
		h-min	.68	1.00	.62	.96	.98	.95	.43	.96	.35	.58
	concurrent + sequential	8 points	1.00	.82	1.00	.96	.94	.96	.92	.59	.92	.81
		h-min	.73	1.00	.71	.91	.96	.90	.44	.92	.41	.59
Satisfaction	ESS 4	11 extreme	.83	.96	.91	.93	.95	1.00	.60	.83	.83	.75
		11 very	.92	.94	.91	.90	.90	.87	.69	.72	.63	.68
	concurrent	11 extreme	.83	.96	.91	.94	.96	1.00	.61	.85	.83	.76
		11 very	.93	.94	.91	.86	.86	.83	.64	.65	.57	.62
	sequential	11 extreme	.82	.95	.91	.97	.98	1.00	.63	.87	.83	.78
		11 very	.92	.94	.91	.89	.89	.86	.67	.70	.61	.66
Social trust	ESS 4 + concurrent + sequential	11 points	.86	.83	na	1.00	1.00	na	.74	.69	na	.71
		6 points	.91	.84	na	.89	.87	na	.66	.53	na	.60
Political trust	ESS 4	11 points	.84	.91	1.00	.98	.98	.99	.68	.80	.98	.82
		6 points	.90	.92	.84	.88	.88	.85	.63	.66	.51	.60
	concurrent + sequential	11 points	.87	.96	1.00	.92	.89	.94	.64	.73	.88	.75
		6 points	.88	.92	.86	.83	.84	.80	.53	.60	.47	.53

[a] "h-min"=time asked in hours and minutes, "extreme"= extreme used in the labels of the end points, "very"= very used in the labels of the end points, "na"= not applicable (no third trait in that experiment)

designs. The same is true when comparing the quality estimates of the different traits: again for media a difference of around 0.2 is found between radio and newspapers for method 1. The difference between these two traits goes even till 0.6 for method 2. These huge differences appear in the different modes in a similar way.

In conclusion, the average quality is similar for the different experiments when the approach of collecting data changes. Using the sequential or concurrent design does not have any impact on the quality of the questions. Potential differences in quality could be expected if the composition of the samples was different in the sequential and the concurrent designs with respect to variables that are influencing the quality of answers, and/or if the proportions of interviews done in the different modes varied a lot and that modes directly impact quality. The preliminary analyses have shown that the proportions of interviews done in the different modes are approximately the same in the two designs: if the modes are combined in the same proportion, then in average the quality of the design should not change if the sample composition is not too different, which seems to be the case. A more central result is that between the unimode and the mixed-mode approaches also few differences are found.

Comparison of the quality estimates by modes

Nevertheless, it is still possible that different subgroups in one sample have different qualities, depending in particular on the mode of data collection they receive. In order to see if this is the case, one can focus on the data from the mixed-mode experiment and analyse it in a different way: instead of dividing the data between groups assigned to different designs, we divide the data between groups interviewed in different modes: CAPI, CATI and CAWI.

The main limit in doing so is that there is a potential selection bias when comparing modes, so if differences are found, we do not know if they are coming from the fact

Table 8: Differences in mean quality between designs for each experiment and method

Experiments	Method	ESS4-Concurrent	ESS4-Sequential	Concurrent-Sequential
Media	8 points	-.03	-.03	.00
	h-min	-.01	-.01	.00
Satisfaction	11 extreme	-.01	-.03	-.02
	11 very	.06	.02	-.04
Social trust	11 points	.00	.00	.00
	6 points	.00	.00	.00
Political trust	11 points	.07	.07	.00
	6 points	.07	.07	.00

that different populations are answering in different modes, or from the fact that answering in another mode change the way of answering of a respondent. In order to test that, we would need respondents to be randomly assigned to the modes. This is not the case in this experiment, since they are randomly assigned to designs, *not* to modes. If differences between modes are due to different populations answering in different modes, this is fine, or even desirable: indeed, if we get the same kind of respondents with the different modes, why should we use several modes? Using only one would be sufficient, and inferences could as well be drawn from the respondents answering in one mode than from the respondents answering in several modes. The interest of adding and mixing modes would therefore be null. On the contrary, if differences are due to a change in the way of answering due to the mode used, then there is a mode effect threatening the comparability of the data across groups of respondents getting different modes. This is what we would like to detect and isolate. But the design of this experiment does not allow directly doing so.

Still, it is interesting to compare the quality in different modes, even if we cannot be sure if differences are found of where they come from, for three reasons. First, previous analyses (e.g. Revilla and Saris, 2010) suggest that differences in sample composition with respect to variables like age or gender or even education do not change much the correlations between other variables of interest as political or social trust. Since the estimation of the quality is based on correlations, we can assume that the impact of having different samples in the different modes does not matter too much. Second, if we do not find differences, even if it is still possible to argue that the two kinds of errors go in opposite directions and cancel each other, we think that this is very unlikely. Third, if without random assignment of respondents to different modes, comparing modes does not allow separating selection from pure mode effects, on the other hand, it provides information on selection biases and therefore is to some extent more realistic.

Table 9 provides the same estimates as Table 7 but focusing on the mixed-mode data and differentiating the groups of people answering by CAPI, CATI and CAWI. Table 10 gives the differences in mean quality between modes.

The mean quality over the three traits is really similar in the three modes for the media experiment. It is the only experiment asking about concrete behaviors, by contrast with the other experiments asking about opinions or attitudes, so this might be a reason why the media experiment leads to more similarities. The similarity of the mean quality however hides some differences: for instance, the first method (8 points) has in fact a .08 higher reliability for radio in CAWI than in CATI and CAPI, but slightly lower validities for TV, radio, newspapers. Therefore, CAWI leads in that case to less random errors than CATI and CAPI, but to more systematic errors.

For the other experiments, there are slight differences even in the mean quality, in particular between CATI and the two other modes. The highest difference is 0.15 in the political trust experiment between CATI and CAWI. This difference comes both from the reliability and validity, which vary for all three traits. In the social trust experiment, no significant differences are found between CAPI and CAWI but a difference of .12 separates the quality in these two modes from the one in CATI when an 11-point scale is used. The lower quality in CATI results both from lower reliability and validity. In the satisfaction experiment, again the biggest differences concern CATI. Besides, even when the mean quality of CATI is almost identical to the one of another mode this may hide differences in reliabilities and validities: considering the difference between CATI and CAPI in the satisfaction experiment for the second method (“11 very”) the mean quality difference is only 0.01. Nevertheless, for the first trait, there is a 0.12 absolute difference in reliability between CAPI and CATI and a 0.07 absolute difference in validity.

6 Discussion – Limits

Comparing one unimode and two mixed-mode designs, little differences are found between these designs in terms of quality. Moving to a comparison of the quality in different modes shows slightly more differences, but principally when comparing CATI with the two other modes.

Finding more differences between CATI and the two other modes can easily be interpreted in terms of differences in measurement’s properties of this mode: indeed, CATI is the only mode purely oral (show cards in CAPI). This could explain the often lower quality (comprehension, memory issues). Nevertheless, CAWI is the only self-completed mode, so one could also have expected more differences between CAWI and the others. The results do not support this idea. It seems instead that the distinction between oral and visual plays a more important role than the presence of the inter-

Table 9: Estimates in the different modes[a]

Experiments	Group	Method	r_{1j}	r_{2j}	r_{3j}	v_{1j}	v_{2j}	v_{3j}	q^2_{1j}	q^2_{2j}	q^2_{3j}	q^2_{mean}
Media	CAPI	8 points	1.00	.79	1.00	.96	.94	.96	.92	.55	.92	.80
		h-min	.72	1.00	.77	.89	.94	.90	.41	.88	.48	.59
	CATI	8 points	1.00	.80	1.00	.96	.94	.96	.92	.57	.92	.80
		h-min	.75	1.00	.70	.89	.94	.90	.45	.88	.40	.58
	CAWI	8 points	1.00	.88	1.00	.93	.90	.93	.86	.63	.86	.79
		h-min	.74	1.00	.73	.89	.94	.90	.43	.88	.43	.58
Satisfaction	CAPI	11 extreme	.85	.96	.93	.96	.97	.97	.67	.87	.81	.78
		11 very	.92	.94	.89	.85	.85	.83	.61	.64	.55	.60
	CATI	11 extreme	.83	.95	.84	.96	.97	.97	.63	.85	.66	.72
		11 very	.80	.90	.82	.92	.92	.91	.54	.69	.56	.59
	CAWI	11 extreme	.84	.96	.93	.99	.99	.99	.69	.90	.85	.81
		11 very	.95	.95	.91	.85	.85	.83	.65	.65	.57	.62
Social trust	CAPI+CAWI	11 points	.89	.86	na	1.00	1.00	na	.79	.74	na	.77
		6 points	.91	.82	na	.91	.89	na	.69	.53	na	.61
	CATI	11 points	.81	.87	na	.96	.96	na	.60	.70	na	.65
		6 points	.92	.84	na	.85	.82	na	.61	.47	na	.54
Political trust	CAPI	11 points	.84	.94	.96	.93	.86	.94	.61	.65	.81	.69
		6 points	.90	.90	.83	.92	.93	.91	.69	.70	.57	.65
	CATI	11 points	.95	.91	.95	.77	.83	.94	.54	.57	.80	.63
		6 points	.81	.87	.87	.89	.89	.87	.52	.60	.57	.56
	CAWI	11 points	.87	.98	1.00	.93	.86	.94	.65	.71	.88	.75
		6 points	.88	.92	.82	.97	.97	.96	.73	.80	.62	.71

[a] "h-min"=time asked in hours and minutes, "extreme"= extreme used in the labels of the end points, "very"= very used in the labels of the end points, "na"= not applicable (no third trait in that experiment)

Table 10: Differences in mean quality between the modes for each experiment and method

Experiments	Method	CAPI-CATI	CAPI-CAWI	CATI-CAWI
Media	8 points	.00	.01	.01
	h-min	.01	.01	.00
Satisfaction	11 extreme	.06	-.03	-.09
	11 very	.01	-.02	-.03
Social trust	11 points	.12	.00	-.12
	6 points	.07	.00	-.07
Political trust	11 points	.06	-.06	-.12
	6 points	.09	-.06	-.15

viewer. One can notice that the similarity of the visual stimulus could even be higher than it was in this experiment, since the show cards could be made with the exact same layout as the screens of the Web survey, or vice-versa. It is clear also that the difference between CATI and CAWI is larger than the one between CATI and CAPI, suggesting, not surprisingly, that when modes differ at the two levels (e.g. interviewer and oral versus self-completed and visual), the quality varies more than when modes differ only at one level. It is important to remark that the findings may depend a lot on the topics and the complexity of the questions analyzed. In this study, the questions are not very complex. Even if more social desirability bias might be expected when an interviewer is present, the topics studied are also not very sensitive. It may be more social desirable to report less television watching and more newspapers reading. Kalfs (1993) for instance observes that respondents report more television watching in Web surveys. Social desirability associated to media use may have changed since 1993, but in any case watching television is still a much less sensitive topic than drug use for example. More work would be useful for really sensitive and complex questions, since more differences could appear between modes.

However, if using CATI instead of CAPI or CAWI conveys differential measurement bias, then, how can we account for the fact that little differences have been found previously when comparing designs? The two mixed-mode designs, according to Table 2, have almost identical proportions of interviews done in the three different modes. This equal repartition of interviews in the different modes in the sequential and concurrent designs may explain that few differences are found between designs even if differences are found between modes. If the number of respondents answering in different modes would have been more different between sequential and concurrent designs, more differences could have been found.

Moving to the comparison unimode versus mixed-mode designs, the argument of equal repartition of modes clearly does not hold. But in that case, the high similarity in quality estimates between on one hand the unimode design and on the other hand the mixed-mode designs may be related to the relatively low proportion of telephone interviews in the mixed-mode designs. Indeed, the comparison of modes suggests that CATI is the most different mode. Only around 18% of the interviews of the mixed-mode designs (once the unknown telephone number group has been added) are done by telephone. This could explain why the differences between designs are lower than differences between modes.

However, differences between modes might encompass both the effect of differential measurement *and* differential selection. An alternative way of looking at the difference between modes would therefore be to think in terms of selection: the lower quality observed in CATI in several cases can be due to the characteristics of the respondents choosing this mode. Table 3 showed differences in respondents in terms of gender and age depending on the mode of interview. If other variables related to the quality of the responses also differ across respondents answering in different modes, they can cause the observed variations in quality across modes. When combining the modes however, the complete sample becomes more similar to the one of the unimode survey, and therefore fewer differences are found when comparing designs. Nevertheless, if the differential selection is the explanation, it could be expected than CAWI would differ from CAPI more than observed in this paper.

Overall, it seems that a mixed-mode using only CAPI and CAWI should not be problematic in terms of quality comparisons. Adding CATI however may be an issue if the difference between CATI and the two other modes comes from differential measurement and not from differential selection. In this study it was not an issue because CATI was the less chosen mode, but one can probably expect more differences between unimode and mixed-mode designs if CATI interviews are more numerous and the difference in quality is due to varying measurement biases. But the study suggests that a mixed-mode approach does not necessarily threaten the comparability of the data, at least concerning the quality.

This result means that switching from a unimode to a mixed-mode data collection should not lead to differences in correlations between observed variables because of the introduction of additional modes. However, it should be clear that this does not mean that the different designs are comparable in terms of means or unstandardized relationships. Studying if means and unstandardized relationships are similar across modes requires different tests that could be the object of further research.

Besides this result about the quality, the ESS mixed-mode experiment is interesting to put in light the difficulties of implementing a mixed-mode design, beginning with the adaptation of the questionnaires from one mode to another (two-step procedures, treatment of the "don't know"), passing by the sampling (no frame of Internet addresses) and the fieldwork (reminders, follow-up) and going till the treatment of the data (standardization of the data, combination of groups). By experimenting them in practice on a relatively large scale, it should help to improve the implementation of such data collection approaches in the future. Because of all these difficulties however, there are several limits to this study.

The first one has already been discussed: it concerns the comparison across modes and the difficulty in differentiating selection and measurement bias. But in this study where the quality turned out to be rather similar this problem is less serious because it is unlikely that the selection bias has compensated exactly for the measurement bias.

The second has also been mentioned: it is the issue of generalizing from the specific unimode and mixed-mode surveys considered in this paper to unimode and mixed-mode surveys in general. We are only focusing on the face-to-face ESS questionnaire, compared with one sequential mixed-mode proposing first CAWI, then CATI and finally CAPI and with one concurrent design offering the same three modes. Many characteristics may vary in other surveys: nature, number and order of the modes proposed, contact procedure, use of incentives, length of the questionnaire, complexity of the questions, sensitivity of the topics, sampling procedure, etc. Moreover, the surveys are all implemented in the Netherlands. Other countries may also have distinct characteristics: differences in telephone and Internet coverage, in the practice of surveys, in the nature of available sampling frames, etc.

The third concerns the way the sequential design has been implemented. In theory, sampling units should have been asked first if they had access to Internet. If they had, they should have been asked to participate by Internet. If and only if they refused, they should have been proposed a second mode (telephone). If and only if they refused again, they should have been offered the third mode (face-to-face). In practice, some doubts exist that this procedure was fully respected. Sequential and concurrent approaches may have been more similar than they should have been. If this is so, it becomes not surprising that the results of the two mixed-mode designs are extremely similar, and it gives limited evidence on the better way to mix modes of data collection. However, it does not change the results concerning the main issue we wanted to study: what is the impact on the quality of switching from a unimode to a mixed-mode design? The study suggests that there is only a slightly impact.

In order to reduce the uncertainty of the results, further research tackling the different problems just mentioned is needed. The design of the study clearly had important limits, but we can learn from this experience and try to overcome these limits. The problem of inference will never be completely suppressed, but it could be limited a bit, by considering for instance different countries. Nordic countries with similar profile as the Netherlands could be used in order to see if the results can be replicated. A mixed-mode approach with face-to-face and Web only (i.e. excluding telephone) may be more appropriate. It would also be interesting to study countries with much lower Internet coverage (Greece, Bulgaria and Romania) in order to see how this affects the main findings of that paper. The repartition of respondents into the different modes would probably be quite different, and the expected reduction of costs would be lower, since fewer respondents would answer with the cheapest mode (Internet). However, the quality may still be quite similar. More analyses would be needed to confirm that. The problem of inference could also be limited by varying more the complexity and sensitivity of the topics.

Acknowledgements

I am very grateful to Willem Saris and Peter Lynn for their extremely helpful comments and support; and to two anonymous reviewers for their useful remarks on a previous draft of that paper.

References

Alwin, D. F. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.), *Sociological methodology 1973-74.* San Francisco: Jossey-Bass.

Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *46*, 409-442.

Brambilla, D. J., & McKinlay, S. M. (1987). A Comparison of Responses to Mailed Questionnaires and Telephone Interviews in a Mixed-mode Health Survey. *American Journal of Epidemiology*, *126*, 962-971.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *6*, 81-105.

Couper, M. P., & Miller, P. V. (2009). Introduction to the special issue. *Public Opinion Quarterly*, *72*(5), 831-835.

De Heer, W., de Leeuw, E. D., & van der Zouwen, J. (1999). Methodological Issues in Survey Research: a Historical Review. *BMS: Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *64*, 25-48.

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233-255.

Dillman, D. A., Clark, J. R., & West, K. K. (1995). Influence of an Invitation to Answer by Telephone on Response to Census Questionnaires. *Public Opinion Quarterly*, *51*, 201-219.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, interactive Voice Response and the Internet. *Social Science Research*, *38*(1), 1-18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3rd ed.). New York: John Wiley & Sons, Inc.

Heerwegh, D., & Loosveldt, G. (2009). Face-to-Face Versus Web Surveying in a High-Internet-Coverage Population. Differences in Response Quality. *Public Opinion Quarterly*, *72*(5), 836-846.

Jöreskog, K. G. (1970). A general method for the analysis of covariance structures. *Biometrika*, *57*, 239-251.

Kalfs, N. (1993). *Hour by hour: effects of the data collection mode in time use research.* Amsterdam: NIMMO.

Kreuter, F., Presser, S., & Tourangeau, R. (2009). Social Desirability Bias in CATI, IVR and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201-219.

Lynn, P., Häder, S., Gabler, S., & Laaksonen, S. (2007). Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience. *Journal of Official Statistics*, *23*(1), 107-140.

Lynn, P., Laurie, H., Jäckle, A., & Sala, E. (2006). *Sampling and Data Collection Strategies for the Proposed UK Longitudinal Household Survey.* Report to ESRC.

Lynn, P., Revilla, M., & Vannieuwenhuyze, J. (forthcoming). *A comparison of five practical mixed mode and unimode strategies in terms of costs, response rates and sample composition.*

Revilla, M., & Saris, W. E. (2010). *A comparison of surveys using different modes of data collection: ESS versus Liss panel.* RECSM working paper 13.

Rhodes, S. D., Bowie, D. A., & Hergenrather, K. C. (2003). Collecting Behavioural Data using the World Wide Web: Considerations for Researchers. *Journal of Epidemiology and Community Health*, *57*(1), 68.

Roberts, C. (2007). *Mixing modes of data collection in surveys: a methodological review.* ESRC National Centre for Research Methods. NCRM Methods Review Paper. NCRM/008.

Roberts, C. (2008). *Designing equivalent questionnaires for a mixed-mode European social survey: report on the findings of the ESS mode experiments.* European Social Survey JRA1-task3.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (p. 575-597). New York: John Wiley & Sons, Inc.

Saris, W. E., & Gallhofer, I. (2007). *Design, evaluation, and analysis of questionnaires for survey research.* New York: John Wiley & Sons, Inc.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural equation modeling: A multidisciplinary Journal*, *16*(4), 561-582.

Shettle, C., & Mooney, G. (1999). Monetary incentives in us government surveys. *Journal of Official Statistics*, *15*(2), 231-250.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2009). *Judgement Rule Aid software.*

Voogt, R. J. J., & Saris, W. E. (2005). Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, *21*, 367-387.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*, 194-212.