UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 21

2011

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici ESCI-Born - Despatxos 19.517-19.527
Passeig Pujades, 1 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Impact of the mode of data collection on the quality of survey questions depending on respondents' characteristics

Melanie Revilla

Research and Expertise Centre for Survey Methodology
Universitat Pompeu Fabra

Abstract

The Internet is more and more used to conduct surveys. However, moving from traditional modes of data collection to the Internet may threaten the comparability of the data if the mode has an impact on the way of answering of the respondents. In previous research, Revilla and Saris (2010) find similar average quality (defined as the product of reliability and validity) for several survey questions when asked in a face-to-face interview and when asked online. But does that mean that the mode of data collection does not have an impact on the quality? Or may it be that for some respondents the quality is higher in Web surveys whereas for others it is lower, such that in average the quality for the complete sample is similar? Comparing the quality for different groups of respondents of a face-to-face and a Web survey, no significant impact on the quality of the background characteristics, the mode and the interaction between them are found.

Keywords

Web and face-to-face surveys, quality of survey questions, MTMM experiments

Acknowledgements

I am really grateful to Willem Saris and Peter Lynn for all their very helpful remarks and suggestions on previous drafts of that paper.

1) Modes of data collection and quality

In the past decades, the number of surveys implemented around the world increased a lot. If surveys were for a long time the relatively closed domain of few scientists, nowadays, most people are able to launch their own survey.

This democratisation of the survey practice has been accompanied by an increasing concern about the representativeness and the quality of different surveys. If most people are able to conduct a survey, not all of them can do a "good" survey. Many online surveys are everything but representative. Therefore, it is necessary to be careful about some of the claimed results (Saris, 2008).

However, using the Internet to conduct surveys is attractive since in principle it can be both quicker and cheaper than more traditional modes, even if in practice it is not always the case. High quality surveys as the European Social Survey (ESS) have started to consider the possibility of switching from their current mode of data collection to Web surveys or to a mixed-mode approach including the Web. However, this may threaten the comparability of the data both across time and across groups (countries if not all the countries adopt the same mode, or subpopulations that answer in different modes if a mixed-mode approach is used in one country).

Because of both the attractiveness and the risks linked to Web surveys, an important literature started to compare Web to other modes of data collection. The comparisons focus mainly on the response rates and non response (Kaplowitz, Hadlock, Levine, 2004; Fricker et al., 2005) and on satisficing and social desirability (Heerwegh, 2009; Kreuter, Presser, Tourangeau, 2009) as indicators of the quality.

Nevertheless, low response rates are only a warning of potential troubles (Couper, Miller, 2009): they do not systematically correspond to a low quality. At the other side, higher response rates does not imply higher representativeness neither higher quality (Krosnick, 1999). The central question is whether higher response rates also mean less non response bias (Voogt, Saris, 2005). Satisficing and social desirability on the other hand are specific to a certain kind of questions and as such are not adapted to measure the quality for all topics.

On the contrary, following Saris and Andrews (1991), Scherpenzeel (1995, 2008) uses a measure of the quality (product of reliability and validity) that can work for all topics and, moreover, allows correcting for measurement errors. This is crucial because there are always errors in the measurement and if this is not taken into account, the conclusions drawn may be wrong. The presence of random errors can attenuate the observed correlations between variables. The presence of systematic errors can lead to overestimated observed correlations. Different groups can have different levels of both random and systematic errors, forbidding any direct comparison across groups. It is therefore useful to look at the quality, defined as the strength of the relationship between the latent variable of interest and the observed answer, to get an idea of the potential measurement errors and if necessary correct for them.

Defining quality in the same way, two papers (Revilla and Saris, 2010; Revilla, 2010) recently focused on the impact of the mode or combination of modes of data collection on the quality of survey questions. Their main result is that the quality is very similar in the face-to-face and the Web surveys compared. From that the author concludes "that there is only a slightly impact" on the quality of switching from a unimode to a mixed-mode design for the data analysed (Revilla, 2010, p.163).

This conclusion may be a bit too quick: does the finding of a similar average quality in both modes really allow concluding that the mode of data collection does not have an impact on the quality?

What is true at the aggregate level is not necessarily true at the micro-level. If the average quality of a sample of face-to-face respondents equals the average quality of a sample of Web respondents, does that mean that the quality of respondent *i* remains the same if respondent *i* gets a face-to-face or a Web interview? An implicit assumption made by the authors is that the impact of the mode of data collection is the same for all respondents. But what if for some respondents the quality is higher in Web than in face-to-face interviews, whereas for others it is the contrary?

The goal of this paper is to test if the assumption of equal impact of the mode of data collection on all respondents does or does not hold. Investigating in each mode if

differences are found between different kinds of respondents is a second issue of the paper. We focus on two modes: Web, because of its impressive growth in the past decades and the huge possibilities it offers; and face-to-face, because it is still nowadays seen as the gold standard for survey research.

Section 2 discusses the assumption of equal impact of the mode on all respondents. Then, section 3 proposes a set of hypotheses. Section 4 explains the model used to test these hypotheses while section 5 gives information on the data. Finally, section 6 gives the results and section 7 concludes.

2) (In)equal impact of the mode of data collection on the quality depending on the respondents' characteristics?

The assumption of equal impact of the mode on all the respondents is in line with a view of quality used for instance by Saris and Gallhofer (2007). In this view, the quality is considered to be a property of the questions per se. Therefore, the quality may be influenced by elements such as: use of battery or separate questions, number of response categories, use of labels, etc. The topic and the visual presentation of the question (horizontal versus vertical scale, use of images) are also considered as potentially influencing its quality (Dillman, Christian, 2005; Toepoel, Das, van Soest, 2005).

Nevertheless, one could argue that the quality depends not only on the question's properties but also on how these properties are perceived by the respondents. The quality may therefore be seen as the result of an interaction between a question's properties and the characteristics of the respondent. If an interviewer is present, a third side may even be considered.

Some research has already been done on the impact of respondents characteristics on the quality. For instance, Alwin and Krosnick (1991) use a simplex model to look at the impact of schooling and age on the psychometric concept of reliability[1] and find that "older respondents and those with less schooling provided the

[1] They define it as the "correlational consistency of response, independent of true individual change". It is therefore "limited to random errors" (Alwin, Krosnick, 1991, p.142)

least reliable attitude reports" (abstract). Their results suggest that characteristics of the respondents may be an element to consider when studying quality. However, they only consider reliability and not the total quality (q^2), i.e. the product of reliability (r^2) and validity (v^2). Besides, they do not take the mode of data collection into account.

A study by Andrews (1984) does consider the mode of data collection and separate validity from method effects and residual errors. Andrews concludes that "respondent characteristics were not a major predictor of variation in the quality of measurement in these data" (p. 433). Nevertheless, some effects of age and education are found. Also, Andrews reports a very small effect of the mode of data collection. But the comparison was between group-administered questionnaires, telephone and face-to-face interviews.

Following this idea, our goal is to investigate if the mode of data collection interacts with some characteristics of the respondents to determine the quality, such that for respondents with some characteristics, switching from a face-to-face to a Web survey would increase the quality of their answers whereas for respondents with other characteristics, it would decrease. If this is the case, a similar quality in average across samples interviewed with different modes does not imply that the mode has no impact on the quality. It may have a different impact on different groups.

Why is it important to know if this is happening? It is important because the correlations and the analyses based on correlations may be biased if differences in quality exist across respondents or for the same respondent across time. Different situations may be though of where problems could appear due to that variation of quality. A few examples are presented below.

First, imagine that one wants to study time series using respondents that at time *t-1* answered by face-to-face and at time *t* answer online and that depending on their level of schooling the quality for some respondents increases (high educated) when switching to the Internet whereas for others (low educated) it decreases. Then, when comparing the answers of one respondent at times *t-1* and *t*, one would get confounding effects of variations in modes and true variations in opinion of the respondent.

Second, one can think about what could happen if one does a survey of a specific population: for example, it is quite usual, for practical reasons, to conduct surveys on a population of students only (e.g. Heerwegh, Loosveldt, 2009; Smyth, Christian, Dillman, 2008). Then, even if the quality in different modes is similar for samples representative of the whole population, if different subpopulations have different qualities when answering in different modes, studies focusing on these subpopulations may suffer from a switch in modes. It may be so that using a face-to-face interview or a Web interview will not lead to the same quality for a student-based survey if students (because of their age or level of education) react differently to the different modes.

Finally, even using a population-based sample, if different modes are used for different respondents of the sample (mixed-mode survey) and if respondents with different backgrounds have the tendency to choose different modes, then it may be problematic to study relationships conditional on that background variables. For instance if one wants to study in a mixed-modes survey relationships conditional on age and the quality varies in different modes for different age groups and these different age groups choose mainly different modes (e.g. younger people choose the Web and older people face-to-face), the conclusions may be incorrect if no correction for variation in modes is done.

3) Hypotheses

First, we should mention that we focus on what we call “normal questions”, meaning questions that are not very complex neither very sensitive. These questions may have different characteristics that impact the quality. But for complex and sensitive questions, more differences in quality can be expected across modes.

In face-to-face interviews, the skills that the respondents need to answer normal questions are quite limited. They have to understand the question and give a response. But the respondents should only say their answer, they do not have to do any manipulation (e.g. check a box): the interviewer is doing this for them. Therefore the second part of the task, providing a response, is simplified.

The first part of the task, understanding the question, is also simplified in face-to-face: indeed, if respondents have problems understanding one question, the interviewers can help them, explaining unknown terms or giving examples to illustrate and clarify the meaning of the question. Therefore, we do not expect large differences between different groups of respondents.

Nevertheless, the analyses of Alwin and Krosnick (1991) and Andrews (1984) suggest that age and education have some impact on the quality. Even for normal questions, the cognitive abilities of the respondents might affect the quality. Also, other factors, as the capacity of concentration, the mental distraction or the motivation of the respondents, may lead to differences in quality: even if all respondents are ideally able to answer with a similar quality, in practice, some may not be motivated enough to provide the effort maximum. Some may be inattentive or may satisfice (Krosnick, 1999). Therefore, even if all respondents have the cognitive ability to reach the same level of quality, it may happen that some groups (e.g. low educated people) are more willing to satisfice than others (e.g. high educated people), which would lead to different qualities of the same question for different groups of respondents. So following previous results, we assume that:

H1a: Eldest respondents have a lower quality in face-to-face surveys than younger respondents.

H1b: Less educated respondents have a lower quality in face-to-face surveys than more educated respondents.

In Web surveys, there are two main aspects that differ and may play a role in determining the quality.

First, Web surveys are self-completed, so the respondents have to do the entire task by themselves. They need to be able to read and understand what the questions mean. They need to understand how to give an answer and how to go to the next question. They need to keep themselves motivated to continue the questionnaire and not skip items. Such surveys are therefore much more demanding.

Second, compared to other self-completed modes, Web surveys require the use of a computer[2] and the Internet. This has both advantages and disadvantages. On the one hand, the branching for example, that may be quite burdensome for the respondents in paper-and-pencil surveys, can be done automatically in Web surveys. Automatic checks can also be made in Web surveys to substitute some of the checks an interviewer could make. Some extra help may also be added more easily to Web surveys than to paper questionnaires (e.g. adding links opening windows with extra definitions). All these possibilities make the Web closer to a face-to-face interview than a paper questionnaire. On the other hand, Web surveys require more skills than paper-and-pencil questionnaires since the respondents have to be able to use a computer and the Internet.

How can these aspects of the Web surveys interact with respondents characteristics? Some authors defending the idea that a "digital divide" exists (e.g. Rhodes, Bowie, Hergenrather, 2003) argue that Web surveys incite more men and young people to participate, and on the contrary discourage women and older people. Besides this potential difference in participation, we want to see in this paper if once they have agreed to participate, we get differences in the quality of the answers of such subpopulations.

In Europe, we believe that nowadays women and men are on average able to understand normal questions without the help of an interviewer and have all in average reached the minimum degree of computer and Internet familiarity required to answer a Web survey.

However, we assume that the eldest respondents are not in general familiar enough with the Internet, such that for them completing Web surveys creates an additional burden and leads to more measurement errors. So we expect the differences in quality between eldest and younger respondents to be higher in Web surveys than in face-to-face ones.

[2] Web surveys can also be completed via a Smartphone or a tablet, but to keep it simple we only speak about "computer".

Another variable of interest is the respondents' education. Because of the self-completed aspect of the Web, we assume that the quality will be lower in a Web than in a face-to-face interview for respondents with a lower level of education, since the absence of interviewer makes their task more difficult. At the same time, because people can choose the moment of the interview and can complete it at their own space, we assume that the quality will be higher in a Web survey for people with high level of education. Concerning the use of the computer and Internet, it can be seen as an extra burden for the respondents with low level of education. On the contrary, since it allows extra checks or to use more friendly designs, it can improve the quality for high educated respondents, by lowering the random errors or increasing their motivation.

So to summarize, we propose the following hypotheses:

H2a: Women and men have similar levels of quality in Web surveys (and a fortiori in face-to-face surveys).

H2b: the difference in quality between eldest and younger respondents (with lower quality for the eldest) is higher in Web than in face-to-face surveys.

H2c: the difference in quality between less and more educated respondents (with lower quality for low educated) is higher in Web than in face-to-face surveys.

Putting together these hypotheses and the fact that previous research does not find relevant differences in the average quality of a face-to-face and a Web survey, it appears that an increase in one group should be compensated by a decrease in another, so we formulate one final set of hypotheses:

H3a: *when switching from face-to-face to Web, the quality increases for the younger respondents and decreases for the eldest.*

H3b: *when switching from face-to-face to Web, the quality increases for the high educated respondents and decreases for the low educated.*

The hypotheses could be specified more precisely: for instance, topics of more interest to the respondents may lead to higher quality. The complexity of the question may also have an impact: for very basic questions, there is little reason to think that the quality depends on respondents' characteristics. Nevertheless, it seems reasonable that

mainly in self-completed modes, when the questions get more complicated, differences appear. The degree of social desirability could play a role too: if different education groups for instance grant different levels of sensitivity to the same questions, then the level of social desirable answers may vary across groups, leading to more variations on the quality estimates for questions seen as differently sensitive for the different groups. But as mentioned earlier, the paper focuses on not very complex and sensitive questions.

4) Method

4.1 Getting the quality estimates

The multitrait-multimethod (MTMM) approach consists in repeating several questions (called "traits") with several "methods" (Campell and Fiske, 1959). To avoid random variations due to the sample, the repetitions should be asked to the same respondents. To avoid possible changes in true opinions or attitudes, they should be asked in a short period of time, preferably in the same questionnaire to guarantee there is no possible communication of the respondents with other persons that could make them change their mind. However, if people are asked several times the same question in a very short period of time, this may lead to memory effect: respondents are not processing the question the second time but instead they are remembering what they answered and saying it again, adapting the answer to the scale if necessary.

Saris and van Meurs (1990) show that after 20 minutes of similar questions respondents usually do not remember their answer anymore. Therefore the different methods should be proposed to the respondents with at least a 20 minutes interval to avoid memory effects. Since at least three methods are necessary for identifying the model, long questionnaires are required. This can increase the cognitive burden of the respondents and may also not always be possible in practice because of costs or time's constraints.

That is why Saris, Satorra and Coenders (2004) propose the split-ballot multitrait-multimethod (SB-MTMM) approach, which combines the MTMM with a

split-ballot (SB) approach, meaning that respondents are randomly assigned to different groups, each group getting a different combination of only two methods.

The true score model proposed by Saris and Andrews (1991) is used. In this model, it is assumed that there is a "true score" T_{ij}, which is a function of the i^{th} trait F_i (with a coefficient equals to the validity coefficient v_{ij}) and of the j^{th} method M_j (with a coefficient equals to the method effect m_{ij}). Then, the observed variable corresponding to the i^{th} trait and the j^{th} method (Y_{ij}) is expressed as a linear function of the true score T_{ij}. The slope corresponds to the reliability coefficient r_{ij}, and the intercept to the random error component e_{ij} associated with the measurement of $Y_{ij.}$ As a starting point, we assume that the traits are correlated with each other but the methods are not correlated with each other neither with the traits and the error terms are not correlated with each other neither with any of the independent variables.

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \qquad \text{for all } i,j \qquad (1)$$

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \qquad \text{for all } i,j \qquad (2)$$

This model allows to separate systematic errors (due to method effects) from random errors and to estimate reliability and validity coefficients. The product squared of these coefficients is the total quality of the question. This total quality for the i^{th} trait and the j^{th} method is denoted $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$.

The maximum likelihood estimation for multiple group[3] analyses of LISREL (Jöreskog and Sörbom, 1991) is used to estimate the model. The model is estimated separately for different gender groups, age groups and level of education groups. The basic model constrains the parameters to be invariant across all groups. The model is tested each time using JRule (Van der Veld, Saris, Satorra, 2009), a software based on the procedure developed by Saris, Satorra and Van der Veld (2009) that allows testing for misspecifications at the parameter level and using both type I and II errors.

[3] Different variables are used to split up the respondents accordingly to our hypotheses: gender, age and education. For instance for gender, the analyses contain 10 groups: men in the Web survey (three split-ballot groups), men in the face-to-face survey (two SB groups), women in the Web survey (three SB groups) and women in the face-to-face survey (two SB groups).

The model is corrected (mainly releasing constraints of invariance across groups or adding extra correlation between two similar methods) till we get an acceptable model according to the JRule test for misspecifications. A list of the modifications made to the initial model is available online[4].

4.2 Using the estimates to test our hypotheses

Since we consider different experiments, with each time several traits and methods, in two surveys and for different background groups, quite a lot of quality estimates are obtained. A table presenting the average quality for the different traits for each method and group can be found in Appendix 2.

Since it is difficult to make conclusions directly from these estimates, in order to test our hypotheses and look at the impact of several potential causes on the quality, we run regressions with the quality estimates as dependent variable.

We cannot run a unique regression with everything because it is the same data that is analyzed when cutting the sample in gender, age and education groups (dependence of the estimates), so we run one regression for each cutting variable.

As independent variables, we first include only the cutting variable (one dummy for men in the first one, one dummy for the eldest respondents in the second one, two dummies, one for low and one for high level of education in the third one[5]), the mode of data collection (dummy for Web), and the interaction between the cutting variable and the mode. So we have the three equations below. From now on, we refer to this first set of equations as "Reg1".

$$q^2_{gender} = \alpha_{gender} + \beta_{1,gender}\,\text{Men} + \beta_{2,gender}\,\text{Web} + \beta_{3,gender}\,\text{Men*Web} + \xi_{gender} \quad (3)$$

$$q^2_{age} = \alpha_{age} + \beta_{1,age}\,\text{More60} + \beta_{2,age}\,\text{Web} + \beta_{3,age}\,\text{Men*Web} + \xi_{age} \quad (4)$$

$$q^2_{educ} = \alpha_{educ} + \beta_{0,educ}\,\text{Low} + \beta_{1,educ}\,\text{High} + \beta_{2,educ}\,\text{Web} + \beta_{3,educ}\,\text{Men*Web} + \xi_{educ} \quad (5)$$

[4] Please see http://bit.ly/rySyUU

[5] The analyses were repeated using "low" education as the reference category instead of "middle" but this does not change the results.

In the second set of regressions ("Reg2" from now on), we add to equations 3 to 5 some independent variables that have been shown to have an impact on quality. It includes the topic of the questions (dummy for each experiment), and three variables about the characteristics of the methods: the number of response categories (numerical), the number of fixed reference points (numerical) and the kind of scales (dummy "IS" equals to one if the scale is Item Specific, 0 otherwise). See for example Saris and Gallhofer (2007) for more details (definitions of these terms, effects on the quality, etc).

5) Data

5.1 European Social Survey (ESS) and Longitudinal Internet Studies for the Social sciences (LISS) panel

The data needed for our analyses has to have several characteristics: first, it is necessary to have repetitions of several questions in one survey for the same respondents in order to use the true score model. Then, all the characteristics of the question varying from one mode to the other can cause differences in the quality that could be confounded with mode effects. To avoid this potential source of difference, we should have the exact same wording of the questions and answer categories in the different modes.

Such datasets are not so common but the ESS round 4 (2008/2009) and one questionnaire completed in December 2008 by the LISS panel's respondents can be used since in both datasets similar SB-MTMM experiments are included. The ESS is done in 25 to 30 European countries every two years since 2002. The interview is conducted at the respondents' home[6]. The LISS panel is a Dutch online panel based on probability sample. Respondents that agree to participate are provided with a computer and Internet access if they do not already have it[7].

These datasets present some limits: first, the LISS panel is a Dutch panel only, so for the comparison we cannot use all the ESS data but we focus only on the Netherlands. Second, since the LISS respondents are members of a Web panel, they all

[6] For more details, please see http://www.europeansocialsurvey.org/
[7] Please see http://www.centerdata.nl/en/MESS

have at least some minimal level of computer skills. It would be preferable to have respondents that are never using the Internet answering to the Web survey since it is for such respondents that we expect the highest differences in quality.

However, these limits are not as problematic as they may look. First, the Netherlands has a high Internet coverage and at the same time experiences a large decrease in its face-to-face response rates, so it would be a good candidate for a switch of data collection approach in a near future. Even if not representative of all European countries, it presents many common characteristics with the Nordic countries in its Internet coverage and response rates.

Second, the method of recruitment of the LISS panel members is such that even people without previous computer and Internet access are integrated in the panel. Since they are proposed questionnaires every month, even if they had no experience at the beginning they are each time getting a bit more trained. But looking at the question about the frequency of use of the Internet we see that still 7.37% of the LISS respondents are using the Internet only once a month or less. So there is still a non negligible part of the LISS respondents that may have a very limited level of computer skills. However, because of the split-ballot design of the LISS survey, for a given SB group in a given experiment, there are too few respondents using the Internet once a month or less to directly test the impact of using frequently Internet on the quality (Appendix 1).

5.2 Choice of the variables

A first set of variables are the ones for which we are going to compute the quality. Once the dataset is decided, we do not have much freedom. Indeed, the surveys only count six MTMM experiments. Table 1 gives, for each one, details about the traits (t_i) and methods (M_i) for which the comparison between the LISS and the ESS could be made.

Table1

Ideally each experiment would count three traits and each of the traits would be repeated using three methods. This is the case for the experiments about media, satisfaction and political trust. However, in the experiments about political orientation, social trust and left-right positioning, one or two of the traits are only measured with M_2 and M_3 (but not with M_1): these traits are used for the estimation but are not considered when looking at the results. Besides, for political orientation and left-right positioning, the third method varies between the LISS and the ESS: in these experiments, the questions asked using M_3 are therefore not considered in the results' section.

The second set of variables consists in the variables used to make the splits. According to our hypotheses, we need variables to measure gender, age and education. Since these variables are used to split the samples in different groups for which the quality is computed, the variables cannot be continuous or even have a large number of categories. Because we think that the difference for age stays between really the eldest respondents and the others, we cut the sample into two subgroups. To get a sufficient number of observations in each group however, we fixed the cutting age at 60 even if it may have been better to cut at a more advanced age (Appendix 1). Concerning education, we separated "low" (lower secondary or less), "middle" (upper secondary and post secondary non tertiary) and "high" (first and second stages of tertiary) levels of education. We made three categories to see the effects both of a low and a high education and see if the effect is progressive or if the opposition is between low on the one hand and middle and high on the other hand (what we expect), or between low and middle on the one hand and high on the other hand.

6) Results

6.1 Results for gender (H2a)

Table 2 gives the results of the regressions with the quality for the different gender groups as dependent variable. The table also gives the regression coefficients when disaggregating the quality into reliability and validity coefficients but only for the regressions with all the explanatory variables. The traits are treated separately for all these analyses. This allows having more observations: 156 for the regressions of gender

and age, and 234 for the regression for education (because we split the data into more groups for education).

Table2

Table 2 indicates that there is no significant impact of gender, neither of the interaction between gender and mode, when considering the quality, or when considering the reliability and validity coefficients separately. We can notice that in "reg1", where only the variables of main interest for us are included, no significant effects are found at all, and the R^2 is almost null. However, by including the topic and some questions' characteristics as independent variables, the R^2 is going up quite a lot. The same is true for the regressions on the validity and reliability separately. We have to be careful about the meaning of the R^2 and the tests of significance because it is linked to the number of observations which is quite low in our analyses. So we should look at the size of the estimates too: for gender and for the mode, they are all really small. So overall, the results seem to support *H2a*.

6.2 Results for age (H1a, H2b, H3a)

Table 3 is similar to Table 2, but provides the results for age.

Table3

Table 3 shows that in the regressions of the quality, but also the ones of the reliability and validity, the coefficient for age is not significant, neither is the one of the interaction between age and mode. This is true both when including only a few independent variables (reg1) and when controlling for the topic and some questions' characteristics (reg2). All the estimates for the variables of interest are almost zero. Only the topic and questions' characteristics have significant effects. Therefore, we cannot accept *H1a*, neither *H2b*.

Besides, Table 3 shows that the mode does not have a significant impact on the quality, reliability or validity coefficients, and we already said that the interaction between age and mode is not significant, so *H3a* is also not supported.

6.3 Results for education (H1b, H2c, H3b)

The same information is displayed for the education analyses in Table 4.

Table4

In Table 4, we see no significant impact of education, and neither of the interaction between education and mode. This is true when using the quality as dependent variable and when using the reliability and the validity coefficients. So *H1b* and *H2c* are rejected. Also, as for *H3a*, the results suggest that *H3b* does not hold.

6.4 Summary

In sum, the signs in the regressions ("reg 2" in Tables 2, 3 and 4) of the coefficients for more than 60 years old (negative), low educated (negative) and Web (positive) seem to support some of our hypotheses. But in fact, all these estimates are really small and none of the variables we are interested in (i.e. gender, age, education, mode of data collection and the interaction between the first three and the mode) has a significant effect on the quality. Therefore, we can conclude that in the data analysed there is no effect on the quality of having a Web instead of a face-to-face interview, that there is no effect of being a man instead of a woman, no effect of being above 60 instead of under 60, no effect of having a low or a high education instead of a middle one. The picture is similar when considering reliability and validity coefficients separately.

On the contrary, almost all the other explanatory variables (topics, IS, number of answer categories and number of fixed reference points) have significant effects. Besides, the size of the effects is sometimes quite large: e.g. for left-right, it is around .20 in the three regressions. So it seems that the most determining for the quality are the properties of the questions.

7) Conclusion

Building on previous results comparing the quality in different modes of data collection, this paper wanted to go one step further, challenging the implicit assumptions made that the impact of the mode is similar for all the respondents, independently of their own characteristics. The fact that the average quality is similar in face-to-face and Web surveys is not sufficient to conclude that the mode has no impact on the quality of survey questions. One of the reasons is that it is possible that the quality is higher in Web surveys for some groups of respondents whereas it is lower for others, leading to the same average. From this main idea different hypotheses were proposed and tested.

The analyses show that when comparing one face-to-face survey, the ESS round 4, with its specificities (use of show-cards is an important one), to one Web survey completed by the LISS respondents, also with its specificities (e.g. probability-based panel), no significant impact of the mode of data collection on the quality is found, but also no impact of gender, age or education, and no impact of the interaction between the mode and these background variables. Therefore, it seems that the hypothesis *H2a* (no differences between men and women in both modes) is supported by our results, whereas hypotheses *H1a* (lower quality for eldest respondents in face-to-face), *H1b* (lower quality for low educated in face-to-face), *H2b* (highest difference in quality between age groups in Web), *H2c* (highest difference in quality between education groups in Web) and *H3a* and *b* (quality increases when switching from face-to-face to Web for younger and higher educated respondents; decreases for eldest and low educated) are not.

This suggests that the implicit assumption made in Revilla and Saris (2010) and Revilla (2010) was valid: at least for different gender, age and education groups, the analyses do not show significant differences in quality for the two modes. This is an attractive finding: it means that switching from one mode to the other can be done (if done "properly") without disturbing the comparison of correlations between observed variables for these different groups. It also means that it is not necessary if we are interested in the quality and in standardised relationships to correct for differences in background between samples since this has no effect.

However, it could be argued that the nature of the data used for the Web survey is problematic. Because the LISS respondents are members of a panel, the part of the population that really has the lowest computer skills is missing from our data. This is one limit to the study. But the rarity of datasets with repetitions of different traits with different methods into the same survey allowing estimating the quality in the way we defined it does not let much freedom. Besides, it seems that there is a trend in different European countries towards the creation of Web panels and we think that if Web surveys are going to be used in the future for high quality surveys, it will probably be via Web panels. Our results in that sense are closer to what might be the future situation.

References

Alwin, D.F., and J.A. Krosnick (1991). "The Reliability of Survey Attitude Measurement". *Sociological Methods and Research* 20 (1):139-181.

Andrews, F. (1984). "Construct validity and error components of survey measures: A structural modeling approach." *Public Opinion Quarterly*, 46:409-42. Reprinted in W.E. Saris & A. van Meurs. (1990). *Evaluation of measurement instruments by metaanalysis of multitrait multimethod studies*. Amsterdam: North-Holland

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

Couper, M.P, and Miller, P.V (2009). "Introduction to the special issue". *Public Opinion Quarterly*, 72(5): 831–835

Dillman, D. A., and L. M. Christian. (2005). "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17(1):30.

Fricker, S., Galesic, M., Tourangeau, R., Yan, T. (2005). "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly* 69:370–392.

Heerwegh, D. (2009). "Mode differences between face to face and Web surveys: an experimental investigation of data quality and social desirability effects". *International Journal of Public Opinion Research*, 21(1):111-119.

Heerwegh, D., and 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. and Sörbom, D. (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.

Kaplowitz, M.D., Hadlock, TD, Levine, R. (2004). "A Comparison of Web and Mail Survey Response Rates." *Public Opinion Quarterly* 68:94–101

Kreuter, F., Presser, S., and 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. (1999). "Survey Research". *Annual Review of Psychology*, 50: 537-567

Revilla, M. (2010) "Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach" *Survey Research Methods* 4(3):151-164

Revilla, M., and 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., and K. C. Hergenrather. (2003). "Collecting Behavioural Data using the World Wide Web: Considerations for Researchers." *Journal of Epidemiology and Community Health* 57(1):68.

Saris, W.E. (2008) "Something has to be done to protect the public against bad web surveys". WAPOR conference Cadenabbia VII: On Misapprehended Quality Criteria, Online Polls and Horoscopes, Lake Como, Italy.

Saris, W.E. and F.M. Andrews (1991). "Evaluation of measurement instruments using a structural modeling approach". *Measurement Errors in Surveys*. Ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. New York, NY: John Wiley & Sons, Inc. 1991. 575-597.

Saris, W.E. and I. Gallhofer (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York, NY: John Wiley & Sons, Inc. 2007.

Saris, W.E., Satorra, A. and Coenders, G. (2004). "A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design". *Sociological Methodology* 34:311-347

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

Scherpenzeel, A. (1995). "Meta Analysis of a European Comparative Study." *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments* 225-242.

Scherpenzeel, A. (2008). "Online interviews and data quality: A multitrait-multimethod study". Draft paper to be presented at the MESS Workshop, 22-23 August 2008.

Smyth, J.D, Christian, L.H, Dillman, D.A. (2008). "Does 'yes or no' on the telephone mean the same as 'check-all-that-apply' on the Web?" *Public Opinion Quarterly*, 72(1):103–113

Toepoel, V., Das, M. and Van Soest, A. (2005). "Design of Web Questionnaires: A Test for Number of Items per Screen" CentER Discussion Paper No. 2005-114. Available at SSRN: http://ssrn.com/abstract=852704

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

Van Meurs, A. and Saris, W.E. (1990). "Memory effects in MTMM studies". In W.E. Saris and A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. 134-146. Amsterdam: North Holland.

Voogt, R.J.J, and Saris, W.E (2005). “Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects.” *Journal of Official Statistics* 21:367–87.

Tables

Table 1: traits and methods for each of the 6 MTMM experiments

Experiments	*Traits*	*Methods*
Political trust	How much do you personally trust each of the institutions: t_1 = Dutch parliament t_2 = The legal system t_3 = The police	M_1 = 11points battery M_2 = 6 points battery M_3 = 11 points score
Satisfaction	How satisfied are you with: t_1 = the present state of the economy in NL? t_2 = the way the government is doing its job? t_3 = the way democracy works?	M_1 = 11 points (extreme) M_2 = 11 points (very) M_3 = 5 points Agree/Disagree
Media	On an average weekday, how much time, in total: t_1 = do you spend watching television? t_2 = do you spend listening to the radio? t_3 = do you spend reading the newspapers?	M_1 = 8 categories hours M_2 = in hours and minutes M_3 = 7 categories general
Social trust	t_1 = Would you say that most people can be trusted, or that you can't be too careful in dealing with people? t_2 = 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?	M_1 = 11 points M_2 = 6 points M_3 = 2 points
Political orientation	t_1 = The government should take measures to reduce differences in income level t_2 = Gay men and lesbians should be free to live their own life as they wish	M_1 = 5 Agree/Disagree M_2 = 5 points
Left right	In politics people sometimes talk of "left" and "right". t_1 = Where would you place yourself on this scale?	M_1 = 11 points M_2 = 11 points (extreme)

Table 2: estimates from different regressions' models for gender

		Reg1 qual	Reg2 qual	Rel coeff	Val coeff
Background	Men	.0185	.0185	-.0028	.0146
Mode	Web	.0303	.0303	.0087	.0118
Interactions	Men*web	-.0018	-.0018	.0082	-.0095
Topic	Pol. trust		.0689**	.0501**	-.0187**
	Satisfaction		.1015**	.0371*	.0081
	Media		-.0816**	-.0555**	-.0262**
	Pol.Orientation		.1912**	.0643**	.0588**
	Left-right		.1798**	.0539*	.0351**
Questions properties	IS		.1711**	.0691**	.0304**
	No. points		.0097**	.0078**	.0016
	Fixedref		.0345**	.0209**	.0043**
	Constant	.6595	.3552**	.7164**	.8800**
	No. observations	156	156	156	156
	R^2	.0111	.5619	.5766	.3842
	Adjusted R^2	-.0084	.5284	.5442	.3372

Note: *significant at 10% level; **significant at 5% level
IS = item specific; Fixedref = number of fixed reference points;
qual= quality, rel=reliability; val=validity
Social trust is used as reference category (experiment with the smallest differences)

Table 3: estimates from different regressions models for age

		Reg1 qual	Reg2 qual	Rel coeff	Val coeff
Background	More60	-.0233	-.0233	-.0118	-.0036
Mode	Web	.0023	.0023	.0056	-.0041
Interaction	More60*web	.0120	.0120	.0082	.0002
Topic	Pol. trust		.0695*	.0096	.0243*
	Satisfaction		.0587	-.0039	.0217
	Media		-.0917**	-.0939**	.0044
	Pol.Orientation		.1483**	.0254	.0699**
	Left-right		.1960**	.0259	.0766**
Questions properties	IS		.0852*	.0588**	-.0278*
	No. points		.0088*	.0101**	-.0003
	Fixedref		.0351**	.0205**	.0052*
	Constant	.6820	.4746**	.7505**	.9243**
	No. observations	156	156	156	156
	R^2	.0034	.4115	.5418	.2386
	Adjusted R^2	-.0163	.3666	.5068	.1804

Table 4: estimates from different regressions models for education

		Reg1 qual	Reg2 qual	Rel coeff	Val coeff
Background	Low	-.0236	-.0236	-.0085	-.0064
	High	.0102	.0102	.0020	.0069
Mode	Web	.0185	.0185	.0069	.0049
Interactions	Low*web	.0215	.0215	.0082	.0049
	High*web	-.0118	-.0118	-.0026	-.0069
Topic	Pol. trust		.0776**	.0414**	.0072
	Satisfaction		.1213**	.0325**	.0379**
	Media		.0183	-.0494**	.0298**
	Pol.Orientation		.2133**	.0563**	.0872**
	Left-right		.2210**	.0726**	.0582**
Questions properties	IS		.1894**	.0751**	.0428**
	No. points		.0099**	.0072**	.0012
	Fixedref		.0341**	.0190**	.0035**
	Constant	.6833	.3374**	.7268**	.8528**
	No. observations	234	234	234	234
	R^2	.0084	.5540	.5678	.3101
	Adjusted R^2	-.0133	.5276	.5422	.2693

Appendix 1

Number of observations in one slip-ballot group (group 1) for the ESS and the LISS for different cuts of the data

	Frequency of use of Internet	
	Once a month or less	Several times a month or more
ESS	140	434
LISS	24	319

	Gender		Age				Education		
	men	women	<60	>=60	<65	>=65	low	middle	high
ESS	260	315	403	172	448	125	217	208	153
LISS	143	200	249	94	282	61	126	105	112

Appendix 2

Table: quality estimates

Experiment		*Political trust*			*Satisfaction*			*Media*			*Social trust***			*Political orientation***		*Left Right**	
Group/Method		M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_1	M_2
Liss	*men*	**.77**	**.82**	**.77**	**.77**	.88	***.63***	*.84*	.40	*.59*	.65	.66	.53	.55	.83	.87	.89
	women	*.69*	.79	*.74*	*.75*	*.88*	*.55*	*.84*	.40	*.59*	.65	.66	.53	**.59**	.83	*.87*	.89
Ess	*men*	.64	**.82**	.72	.67	**.88**	**.55**	.81	.40	.56	.65	.66	.53	*.59*	.83	**.87**	.89
	women	.64	.79	.72	.67	.75	.50	.81	.40	.56	.65	.66	.53	.59	.83	.80	.89
Liss	*<60*	*.66*	.84	**.75**	.61	.83	.63	**.84**	.38	.55	.75	**.60**	*.53*	.57	.84	*.92*	***.92***
	≥60	*.66*	.84	.69	*.61*	.83	.63	.81	.38	.55	*.75*	.53	*.53*	*.57*	.84	*.92*	*.90*
Ess	*<60*	**.65**	.84	.75	**.61**	.83	.63	.84	.38	.55	**.75**	.60	.50	**.61**	.84	**.87**	**.90**
	≥60	.59	.84	*.75*	.55	.83	.63	*.84*	.38	.55	.65	*.60*	.50	.52	.84	.83	.78
Liss	*Low*	***.75***	.81	.73	*.70*	*.89*	*.55*	*.88*	.44	.60	.69	.59	.52	.57	**.85**	.92	.90
	Mid	*.67*	.81	.73	***.78***	*.89*	*.55*	*.88*	.44	**.62**	.69	.59	.52	.57	**.85**	.92	**.92**
	high	*.70*	.81	.73	***.78***	*.89*	.55	*.88*	.44	**.62**	.69	.59	.52	.57	.75	.92	**.92**
Ess	*Low*	.60	.81	.73	.66	.74	.51	.83	*.46*	.60	.69	.59	.47	.57	.85	.73	.87
	Mid	**.65**	.81	.73	.69	**.85**	.51	.86	*.46*	**.62**	.69	.59	**.52**	.57	.85	**.83**	**.90**
	high	**.65**	.81	.73	.69	**.85**	***.58***	**.87**	***.51***	**.62**	.69	.59	**.52**	.57	*.85*	**.83**	**.90**

*Note: ** based on 2 traits; * based on one trait; in **bold** if the quality for a given method in a given experiment and a given mode is strictly higher in the corresponding group; in italic if for a given method and experiment and group (gender or age or education group) the quality is higher in the corresponding mode. LISS is the Web survey, ESS the face-to-face one.*