Impact of the mode of data collection on the quality of survey questions in social sciences

Melanie Audrey Revilla

Research and Expertise Centre for Survey Methodology
Universitat Pompeu Fabra

Department of Political and Social Sciences

TESI DOCTORAL UPF / 2012

Directors de la tesi:

Professor Willem Saris (Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra)

Professor Peter Lynn (University of Essex)

Table of contents

Dedication / acknowledgements

January 2003, Mende (France): "I don't know what I want to do, but I know I don't want to do any Spanish anymore".

December 2007: I am seating on the benches of the Universitat Pompeu Fabra (UPF), Barcelona, **Spain,** as a student of the Master in Economics. One of the teachers asks who wants to do a PhD next year. 30 out of the 32 students put their hand up. I do not: "I don't know what I want to do, but I know I don't want to study anymore. I won't do a PhD."

September 2009: I am following my first class as a **PhD** student in the political and social sciences department of the UPF.

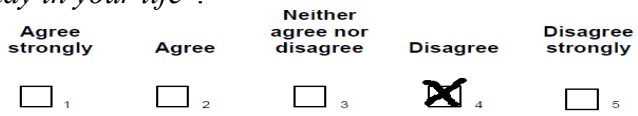
Confucius (551-479 BC) said: "*Only the wisest and stupidest of men never change*".

Barcelona's sun and reputation changed my deep negative feeling toward Spanish language. Willem Saris and Daniel Oberski changed my deep lack of motivation toward doing a PhD. Indeed, in January 2008, I started a course about multivariate statistics taught by Willem; Daniel assisted him (or should I say us?).

Confucius said: "*Everything has its beauty but not everyone sees it*".

Willem and Daniel made me discover the beauty of multivariate statistics and structural equation modelling! I started enjoying multitrait-multimethod models and dreaming about quality.

Confucius said: "*Choose a job you love, and you will never have to work a day in your life*".



So I started as a research assistant at ESADE. My first job assignment was a trip to Berlin to attend a mini-conference. It was confirmed: I started a job I love. The warmest of welcomes of the

ESADE team, which included, besides Willem and Daniel, Wiebke Weber, Lluís Coromina, Irmtraud Gallhofer, Desiree Knoppen, Joan-Manuel Batista Foguet, and soon also Paolo Moncagatta, only increased this feeling, as well as the different trips that followed in this academic year: Napols, Maastricht, Bolzano, Warsaw.

This last one gave me the opportunity to meet another important person: Peter Lynn was giving the keynote speech on the first day of the ESRA conference entitled "Mixed or muddled? Combining survey modes in the 21st century". His speech was an important source of inspiration for my future dissertation.

At the end of spring, our team moved from ESADE to the Universitat Pompeu Fabra where the Research and Expertise Centre for Survey Methodology was created. Was it because of the new offices, the new coffee machine or the view on the ostriches of the zoo? In September 2009, I finally knew what I wanted to do: "I want to continue doing research; I want to do a PhD."

Confucius said "*Life is really simple, but we insist on making it complicated*".

As the PhD started, complications did not wait long to pop up. I quickly specialised in finding problems but not solutions, and came back to my old custom of doubting of everything. Fortunately Willem, Daniel, Wiebke and Paolo were still there to support me. Other people also joined the team for shorter or longer periods over the years: André Pirralha, Maria-José Hierro, Aina Gallego, Gerardo Maldonado, Diana Zavala, Laur Lilleoja, Thomas Gruner, Raquel Andino. I also had the opportunity to visit Peter Lynn at Essex University for a few weeks and get some helpful comments. Besides, I was able to assist several ESS meetings to discuss the results of the mixed-mode experiment and the options for the future. I learned a lot from these meetings with the ESS team, even if they often ended on even more open questions than they started.

Confucius said: "*He who knows all the answers has not been asked all the questions*".

Months passed and little by little, some problems were solved, or at least I thought so. Sometimes, it later appeared that the "solutions" were only hiding new problems.

Confucius said: "*Our greatest glory is not in never falling, but in getting up every time we do*".

So I kept going, motivated by several other meetings I was nicely invited to, as the Science Po conference in Paris in 2009 or the MESS workshops every summer in the Netherlands. The different presentations and discussions were always very fruitful and gave me strength to resist the numerous distractions that offers Barcelona...

Confucius said: "*One joy dispels a hundred care*".

The first paper was published, others were rejected. Changes were made, again, and again, and again.

Confucius said: "*It does not matter how slowly you go as long as you do not stop*".

I kept going… and one day, finally, the dissertation got to an end.

Confucius said: "*By three methods we may learn wisdom: First, by reflection, which is noblest; second, by imitation, which is easiest; and third by experience, which is the bitterest*".

The elaboration and writing of the dissertation combined reflection, imitation and experience. Even if I did not learn wisdom, I learned a lot in the process, and cannot thanks enough all the people that helped me in one or another way during these years, including of course my family and friends, the other PhD students with whom I shared classes and seminars, all the researchers I met at different workshops and conferences who were kind enough to give me comments, and most of all my colleagues and thesis supervisors, with a really special thanks to Willem and Daniel without whom I would never have started this adventure!

Abstract

This dissertation studies the impact of the mode of data collection on the quality of answers to survey questions, defined as the product of reliability and validity. Using data from the Netherlands about different topics (media, social and political trust, satisfaction, political orientation, left-right self-placement, attitudes toward immigration), it shows that the quality is similar in a computed assisted face-to-face survey using show cards (the European Social Survey, ESS) and a web survey based on a probability sample (the LISS panel). This is true both at the level of single items and composite scores. It suggests that standardised relationships across variables can be compared across these two modes. On the contrary, telephone interviews lead to some differences in quality. For complex concepts, measurement equivalence also holds, meaning that means and unstandardised relationships can be compared across the face-to-face and web surveys mentioned previously.

Resumen

Esta tesis estudia el impacto que el método de recolección de datos en encuestas tiene sobre la calidad de las respuestas, definida como el producto de la fiabilidad y la validez. Utilizando datos de Holanda sobre temas diversos (utilización de los medios de comunicación, confianza social y política, satisfacción, orientaciones políticas, auto-ubicación en la escala izquierda-derecha, actitudes hacia la inmigración), se muestra que la calidad es similar en una encuesta cara-a-cara asistida con ordenador y utilizando tarjetas (la Encuesta Social Europea) y una encuesta *online* basada en una muestra probabilística (el panel LISS). Esto se cumple tanto para los indicadores simples, como para indicadores complejos. Los resultados sugieren que las relaciones estandarizadas entre variables son comparables entre los dos métodos de recolección. Al contrario, las entrevistas telefónicas producen diferencias de calidad. Para conceptos complejos, la equivalencia de las mediciones también está garantizada: las medias y las relaciones no estandarizadas son comparables en las entrevistas cara-a-cara y *online*.

General Introduction

Goal and scope of the dissertation

Conducting surveys allows collecting data about people's opinions, attitudes and behaviours. These data can be analysed and used to show people's preferences, to explain people's choices, to predict people's actions, sometimes to make decisions about future political orientations, etc. There are other ways to obtain data (e.g. by observing people), but conducting surveys is one of the most used approaches: already in 1995, 69.7% of the papers published in sociology and 41.9% of the papers published in political science used survey data (Saris and Gallhofer, 2007).

In survey research, data collection is a crucial phase as it determines to a great extent the quality of the results. But it is also a tricky step: many decisions that have to be made may impact the final findings. Therefore, it is necessary to take all these decisions into account when conclusions are drawn. One of these decisions is the mode of data collection. Note that throughout the dissertation, when we speak about the "mode of data collection", we refer to the mode used at the response stage of the survey: a "web" survey means a survey where the questions are answered on the web, just as a "face to face" survey refers to a survey in which respondents answer in face-to-face interviews, and a "telephone survey" refers to a survey in which respondents answer an interviewer on the phone. The contact with the respondent can be established in the same or in a different mode (e.g. contact letter): this does not change the way we refer to the survey.

The mode of data collection can impact the final results via its role at different levels. The total error of a survey estimate comes indeed from different sources, mainly coverage, sampling, non-response, measurement and processing errors. Different modes of data collection may lead to different levels of these several sources of errors.

For instance, if different sampling frames (e.g. postal addresses versus telephone numbers) are used depending on the mode of data collection chosen for the survey, different coverage errors may be

expected. Besides, part of the population may not have access to some of the modes, creating coverage errors.

Also, the mode used might influence the composition of the sample and so the sampling errors: for instance, one common idea is that using a Web survey will lead to a younger sample than using a telephone survey.

Even if everybody has access to the different modes, some individuals may feel uncomfortable participating in certain modes: different individuals may choose to participate or not depending on the modes of data collection proposed. If the mode has a differential impact on the response rates of different subgroups, differences in non-response errors will appear.

Moreover, even if the same individuals agree to participate in different modes, leading to identical samples, differences in measurement errors can appear just because the survey is conducted in a different mode. For instance, the presence or absence of an interviewer can result in different levels of social desirability bias (Krosnick, 1991, 1999). Also, because of memory limitations and depending on the cognitive elaboration of the question, it can be assumed that oral modes convey more recency effects whereas visual modes convey more primacy effects. This can constitute another explanation for differences across modes (Krosnick and Alwin, 1987). This dissertation focuses on this kind of errors: the measurement errors.

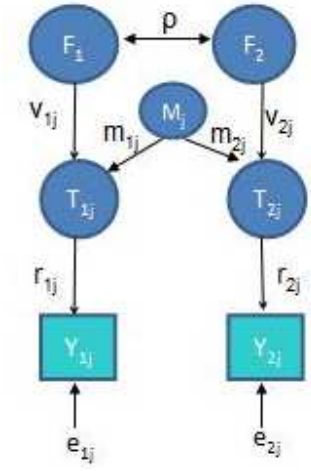
Finally, processing errors can be influenced by the mode chosen. For instance because computer assisted modes allow checks to avoid out of range numbers (error message if the number of hours a day spent in some activities is higher than 24, etc).

There is a growing body of research that focuses on the comparison of modes of data collection. A first wave of studies was linked to the growing telephone coverage: guidelines were proposed to transform questionnaires from one mode to another (Groves, 1990) or differences between telephone and mail or face-to-face were assessed (Hox, De Leeuw, 1994). These kinds of comparisons are still undertaken today (Holbrook, Green, Krosnick, 2003; Jäckle, Roberts, Lynn, 2006). A second wave was linked to the

development of computer technologies and the possibility of using computer assisted methods of interviewing (Kalfs, Saris, 1998; Lynn, 1998; Newman et al, 2002; Perlis et al, 2004). A third wave is linked to the introduction of the Internet. The same issues are addressed but adapted to this mode: many studies focus on the comparison of web surveys with surveys using more traditional modes (Forsman and Isaksson, 2003; Kaplowitz, Hadlock, Levine, 2004; Schonlau et al, 2004; Fricker et al., 2005; Lozar Manfreda et al, 2005; Faas and Schoen, 2006; Heerwegh, Loosveldt, 2008; Heerwegh, 2009).

A lot of previous research focuses on a comparison of response rates, item non-response or variables' distributions. Some also focus on satisficing and social desirability bias as an indicator of the quality (Holbrook, Green, Krosnick, 2003; De Leeuw and van der Zouwen, 2008; Kreuter, Presser, Tourangeau, 2009). However, only little research (Scherpenzeel, 1995; Scherpenzeel and Saris, 1997) has been done on comparing the quality of the measures in different modes when the quality is defined as the strength of the relationship between the latent variable of interest and the observed response. Defined that way, the quality q_{ij}^2 represents the variance of the observed variable Y_{ij} explained by the latent variable of interest F_i. It can be computed as the product of the reliability (r_{ij}^2) and the validity (v_{ij}^2) as can be seen in Figure 1: $q_{ij}^2 = r_{ij}^2 * v_{ij}^2$.

Figure 1: Illustration of the True Score model



Legend:
- F_i = the i^{th} factor or trait
- ρ = correlation between the factors
- T_{ij} = the true score for the i^{th} trait and j^{th} method
- Y_{ij} = the answer to the question for the ith trait with the j^{th} method
- M_j = the j^{th} method factor

And working with standardized variables:
- r_{ij} = reliability coefficient
- r_{ij}^2 = reliability = 1 – var(e_{ij})
- v_{ij} = validity coefficient
- v_{ij}^2 = validity
- m_{ij} = method effect coefficient
- m_{ij}^2 = method effect = 1-v_{ij}^2

Total quality of a measure: $q_{ij}^2 = (r_{ij}.v_{ij})^2$

This definition presents advantages. It allows determining which mode or combination of modes produces less errors. This information has concrete implications: it can be used in a costs-errors trade-off in order to make decisions on the procedure to follow to collect the data. Also, knowing the quality allows correcting for measurement errors, which is always an important step to get proper results, but is even more crucial in comparative research.

Why? Because if the quality varies for different groups of respondents, for instance groups answering in different modes, even if the correlation $\rho(F_1,F_2)$ between two latent variables F_1 and F_2 is the same for all groups of respondents, this does not mean that the observed correlations $r(Y_{i1};Y_{i2})$ for these groups of respondents are also the same[1]. On the contrary, if the observed correlations are similar, this does not mean that the true correlations are similar: the correlations between the latent variables can be different if the quality is different.

Therefore, if one wants to make comparisons, it is necessary to determine the quality first. If one wants to make comparisons across groups of respondents answering in different modes (e.g. surveys using different modes, waves of one survey switching modes at some point in time, groups in one single survey using a mixed-mode data collection approach), it is necessary to determine the quality in different modes. The goal of my dissertation is to compare the quality of surveys using different modes or combination of modes of data collection.

[1] The formula expressing the relationship between the latent and observed correlation is given for instance in Saris and Gallhofer, 2007, pp. 188, equation (9.1): $r(Y_{i1};Y_{i2})=r_{i1}v_{i1}\rho(F_1;F_2)v_{i2}r_{i2}+r_{i1}m_{i1}m_{i2}r_{i2}$ where $r(Y_{i1};Y_{i2})$ is the observed correlation between the variables Yi_1 and Yi_2, $\rho(F_1;F_2)$ is the true correlation between the latent factors F_1 and F_2, r_{ij} corresponds to the reliability coefficient, v_{ij} to the validity coefficient and m_{ij} to the method effect coefficient for the i^{th} method and j^{th} trait.

To study the quality defined as the product of reliability and validity, it is necessary to have data with repetitions of questions for the same respondents, i.e. data including multitrait-multimethod (MTMM) experiments. To compare the quality in different modes of data collection, it is necessary to have data with repetition of questions within each questionnaire and repetitions of the same questions in different modes. Such data are not very common. Nevertheless, we were able to get some.

Indeed, the necessity of such a research has been clearly apprehended by the coordinating team of one of the most important comparative surveys in Europe: the European Social Survey (ESS). So far the data in the ESS is collected by face-to-face interviews at respondents' house. However, because of the high costs of this data collection mode and because of the increasing difficulties for many participating countries to reach the response rates' target, the ESS decided to consider a possible switch in modes of data collection. Some countries, more advanced in their Internet coverage could switch to an online survey, whereas others could switch to telephone surveys; still others could continue with the traditional face-to-face interviews. Another idea would be to use different modes in combination within a single country. Using a mixed-mode design might reduce the costs, saving time and assuring high response rates.

This sounds very attractive, but the ESS wanted to make sure not to harm the comparability of its data (across time, across countries or, across groups of respondents within countries). Therefore, before making any decision, they first launched several experiments to study the possibility of a mixed-mode design as alternative for the current face-to-face only approach.

The ESS is a very large survey, implemented in 29 European countries since 2001 (at least one wave) and used to measure key concepts for political and social sciences, such as political efficacy, social and political trust, socio-political orientations, satisfaction with the government, the economy and the functioning of democracy, voting behaviours, attitudes towards immigration, toward the welfare state, media use, health, security, religious allegiances, social exclusion, human values, etc. Approximately 120 items are answered by around 30.000 respondents every two years.

This huge database is used by many researchers. Most of their analyses are done without the correction for measurement errors. If the mode of data collection impacts the quality of the data and this is not taken into account, switching to another mode or a mixed-mode data collection may lead to different evaluations of these central concepts, compromising the comparability across countries, across time and/or across groups. It might lead to wrong conclusions, which can impact the behaviours of individuals or institutions. Therefore, it is crucial to study possible mode effects before the ESS allows using new modes. Studying mode effects in the frame of the ESS however is not only useful for future decisions concerning the ESS data collection approach, but can also serve as a source of information for other surveys interested in switching modes or starting implementing a mixed-mode design.

Therefore, the dissertation has both a more practical goal (helping by our analyses the ESS to make its decision about the allowance in a near future of other modes of data collection than the traditional face-to-face interviews) and a broader interest (add some evidence to the literature on the effect of the mode of data collection by looking at an indicator that is missing when looking at web based surveys).

More specifically, the ESS data used in the dissertation comes from two main sources:

1. ESS round 4 (September 2008 - June 2009): this is the reference with which the alternative designs are compared.
2. ESS mixed-mode experiment (September 2008 - June 2009): made in the Netherlands in parallel to the traditional face-to-face ESS survey. It uses the same questions as round 4.

In addition, a third dataset is used in the dissertation:

3. One study answered by the Longitudinal Internet Studies for the Social Sciences (LISS) panel: this Dutch panel created in the frame of the Measurement and Experimentation in the Social Sciences (MESS) project presents each month one online study to its members. We made a proposal for one of these studies to get a repetition of the questions asked in the ESS round 4, including the repetitions with different methods of the same questions. Our

proposal was accepted and in December 2008, a questionnaire similar to the one of the ESS round 4 was answered by the LISS respondents through the web.

Using these three datasets allows studying three modes of data collection (used separately or in combination): face-to-face, telephone and web. Face-to-face is the current mode of data collection of the ESS and is still quite often used so it is our benchmark. The web is the most attractive mode in terms of costs and allows getting a huge amount of data in a very short time: introducing it would be the most interesting option in terms of costs-effectiveness. Telephone is used a lot in certain countries (e.g. Norway or Switzerland) and could save travelling expenses, so it is also attractive.

Different substantive topics are studied in the dissertation. Most chapters analyse the same topics. These are the topics for which questions were repeated using different methods, i.e. the ones for which a multitrait-multimethod experiment is available: social and political trust, political orientation, placement on a left-right scale, media use and satisfaction. In Chapters 1 and 4, some questions about attitudes toward immigration are included too. A list of the questions for each topic is available in Table 1.

Table 1: List of questions per topic studied in the dissertation

Experiment	Questions
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?
Social trust	- 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? - Would you say that most people deserve your trust or that only very few deserve your trust?
Satisfaction	- On the whole how satisfied are you with the present state of the economy in [country]? - Now thinking about the [country] government, how satisfied are you with the way it is doing its job? - How satisfied are you with the way democracy works?
Political orientation	- The government should take measures to reduce differences in income level - Gay men and lesbians should be free to live their own lives as they wish - The government should ensure that all groups in society are treated equally
Trust in institutions	How much do you personally trust each of the institutions: - [country] parliament - the legal system - the police
Left-right	In politics, people sometimes speak about "left" and "right". - Where would you place yourself on that scale? - Where would you place the party you most like? - Where would you place the party which you most dislike?
Consequences of immigration	- It is generally bad for the [country] economy that people come to live here from other countries - [Country] cultural life is generally undermined by people coming to live here from other countries - [Country] is made a worse place to live by people coming to live here from other countries
Allowance of immigrants	- [Country] should allow more people from the same race or ethnic group as most [country] people to come and live here - [Country] should allow more people of a different race or ethnic group as most [country] people to come and live here - [Country] should allow more people from the poorer countries outside Europe to come and live here

Structure of the dissertation

The different chapters of the dissertation are very much connected with each other. The datasets used, the topics studied, the method of analysis and the indicators of interest are similar in most of the chapters. However, even if the general thematic unity is clear, each chapter concentrates on a specific question, in the following way.

As pointed out by Couper and Miller (2008), two web surveys can be extremely different. The same is true for two face-to-face surveys. So a survey cannot be described only by the fact that it is a "face-to-face" or a "web" survey. Many other elements can vary: e.g. the sampling procedure, the way of approaching the respondents, the length of the survey, its main topic, etc. It is necessary to be more precise about what we are studying. Since most of the dissertation is based on the ESS round 4 data (face-to-face) and the LISS study of December 2008 (web), first, **Chapter 1** conducts a general comparison of these two surveys. It begins by a comparison of the composition of these two samples with respect to the main background variables. It shows for all variables studied that there are significant differences from the population distribution for both surveys and differences between the two surveys. However, these differences matter only if they affect the results, so then the impact on different kinds of results is investigated, as well as the effect of a correction using a weighting procedure.

Having shown that using weights or not does not change the results, we move to the central research interest, which is evaluating quality in different modes. **Chapter 2** compares the quality of answers to questions in the ESS round 4 and the LISS study of December 2008. The goal is to see if using a web instead of a face-to-face survey leads to significantly different quality coefficients. The quality is estimated using a split-ballot MTMM model in both surveys simultaneously, which allows testing for significance of differences in quality estimates across surveys. Overall, few differences are found across modes even if large differences are found depending on the methods, suggesting that face-to-face and web surveys, when done in a certain way (cf. chapter 1), may lead to similar quality of answers to questions as long as the form of the

questions is kept the same in both modes. This means that standardised relationships can be compared across these modes.

However, finding that the quality of the two samples is on average similar is not enough to conclude that the mode of data collection has no impact on the quality. It might be that the mode has a different impact on different kinds of respondents such that in average the quality of the samples is similar but at the micro level, the quality varies for respondents with certain characteristics. Chapter 2 assumes that the impact of the mode of data collection is similar for all respondents. Nevertheless, it may in fact vary depending on the respondents' characteristics. **Chapter 3** tests this implicit assumption that was made in chapter 2 by looking at the quality estimates for respondents with different background characteristics in those two samples. The quality estimates are regressed on background characteristics, the mode of data collection, the interaction between them and a series of control variables. No significant impact of the background variables are found, neither of the mode nor of the interactions terms. Therefore, it suggests that the implicit assumption of Chapter 2 holds.

Chapter 2 and 3 consider the quality at the level of single items. But many concepts in social sciences are too complex to be measured by one single indicator and are based on the combination of several items. For instance, political trust is a complex concept that is usually measured by asking a battery of questions about how much trust the respondents have in the parliament, in the legal system, in the police, etc. **Chapter 4** tests if measurement equivalence of four of these complex concepts holds across the two different modes of data collection (ESS round 4 and LISS study). Configural, scalar and metric invariance are found for all four concepts, meaning that unstandardised relationships as well as means can be compared across modes. Chapter 4 also computes the quality of the Composite Scores: quite similar quality estimates are obtained in both modes. This means that standardised relationships can also be compared across the surveys.

Because of coverage and non-response issues however, nowadays, it seems quite unlikely for many surveys to switch from traditional modes to a web-only survey. Some of the European countries participating in the ESS for instance still have an Internet coverage

lower than 40% (Hungary, Romania, Bulgaria, Greece and Turkey, cf. Eurobarometer 71.2 2009). Providing Internet access to all participants agreeing to participate will be very costly and consequently is not really an attractive option. Therefore, the idea of mixing different modes within one survey seems more viable. So **Chapter 5** focuses not on a switch from one mode to another but on a switch from a unimode design to a mixed-mode design. Is a switch to a mixed-mode design harming the comparability of the data in terms of standardised relationships? If not, is there a better way of mixing modes of data collection? In order to answer these questions, Chapter 5 compares the traditional ESS face-to-face (round 4) with the data from the ESS mixed-mode experiment (2008/2009) in the Netherlands. It shows few differences between the unimode and mixed-mode designs in terms of quality. Moreover, telephone is included in this chapter in addition to face-to-face and web. The analyses show that this mode is the most different.

Finally, a short conclusion summarised the main findings, underlines some limits and proposes some ideas for future research and future data collection.

Chapter 1

European Social Survey round 4 and LISS panel study of December 2008: A preliminary comparison

1. European Social Survey round 4 and LISS panel study of December 2008: A preliminary comparison

Before getting to the core of the dissertation and the impact on the quality of using different modes, this chapter wants to provide a few preliminary comments and analyses about the two main surveys studied in the dissertation. As mentioned in the introduction, the mode of data collection is one important element to characterise a survey data collection approach but it is far from being the only one. The web panel studied here might differ in more aspects from other online surveys (e.g. opt-in online panels where anybody can freely decide to be part of the panel) than from the face-to-face survey considered. It is important to know a bit better the two surveys in order to be able to better understand the latter results and to better evaluate to what extent the comparison between them is meaningful and to what extent results can be generalised to other surveys.

1.1. Choice of the datasets: advantages and limits

In chapters 2, 3 and 4, two surveys are compared:

- the European Social Survey (ESS), a bi-annual European survey which began in 2001 and where the data is collected by face-to-face interviews at respondents' homes using show cards
- a study completed by the Longitudinal Internet Studies for the Social Sciences (LISS) panel, a Dutch online panel created in 2007.

These two surveys have been chosen for the comparison first because they both include some multitrait-multimethod (MTMM) experiments (with the same question repeated twice to the same respondents but using a different method), and second because these experiments are similar (same questions) in the different modes of data collection. Except for the background variables, which are treated differently since the LISS is a panel, the questionnaire proposed to the LISS respondents is a simple adaptation from the face-to-face ESS version to a web version. Therefore, it offers the opportunity to compare similar questions asked at the same moment (end 2008-beginning 2009) using two modes of data collection, with repetitions of the same questions using different methods.

Since the LISS panel is a Dutch panel, even if the ESS is present in many countries, in order to avoid variations due to cultural or language differences, we focus only on the Netherlands. If the focus on the Netherlands is constrained by practical reasons, there are however other arguments in favour of working with data from this country. On the one hand, collecting the data in the traditional way is becoming more and more difficult in this country. If in the first ESS round the response rate was closed to the ESS objective (67.9% versus 70%). But in the second round it had already decreased to 64.5% and in the third it was 59.8%. In the forth, the fieldwork period had to be extended a lot (almost 10 months instead of six in the previous rounds) in order to get a level of participation of only 52.0%. Everything indicates that the face-to-face interviewing is not working well in that country, meaning that some changes would have to be implemented in a near future. On the other hand, the country beneficiates of a large coverage of the population in terms of Internet access: around 85% of the Dutch adults have access to the web, either at home or at work (most of them at home and at work, cf. ESS round 4 question). Therefore, switching to a web survey or a mixed-mode design including the Internet would make sense. The Netherlands appear as a good candidate for an impending switch of modes of data collection.

By comparing two surveys, we run the risk that two sources of errors are confounded. Indeed, such a design does not allow separating differences in sample composition and differences due to the mode per se. In that sense, a design with the same respondents answering in different modes would be more adapted. Nevertheless, the later design does not provide a good vision of what would happen in reality if different modes of data collection were used in a survey since it does not inform about the potential selection bias when different modes are used. So, besides the fact that no adequate data (including MTMM experiments) was available to study the same respondents answering in at least two modes, comparing two surveys has the advantage of providing richer information about the extent to what one or another mode is viable for future data collection.

1.2. Main characteristics of the surveys

In order to know what we are comparing when comparing two different surveys, it is useful to say a little more about them[2].

First, both surveys are using as sample frame a list of postal addresses. From this list, the sampling units are selected based on probability sampling. The sampling units therefore are households. Households agreeing to participate in the LISS panel are provided with computers and Internet access if they do not have it. In the LISS panel, all persons from the household can participate, but for the specific study of interest (of December 2008), in order to make it comparable to the ESS process, only one person per household was selected. The respondents are first contacted by a cover letter, followed by a telephone call or house visit.

It is only at the response stage (when the respondents answer the survey) that the mode of data collection differs: face-to-face for the ESS, against online completion for the LISS respondents. Consequently, an interviewer is present in the case of the ESS, but not in the case of the LISS. Also, the ESS stimulus is both oral and visual as most of the ESS questions are asked with show cards, whereas it is only visual in the LISS.

The number of observations of both surveys is high: 1,775 for the ESS and around 3,200 in the LISS. This corresponds to a response rate of 52% in the ESS. In the LISS, 65% of the panel members selected (persons that accepted to be part of the panel and where selected to be the person in the household that will get the questionnaire of interest) responded to the survey sent in December 2008 (the one used in this paper). But the panel membership rate should also be taken into account: 48% of the sampling units accepted to participate in the panel. The final response rate is therefore 65% of 48%, i.e. 31% of the initial sample, which is much lower than the ESS response rate. On the other end, the LISS panel was much quicker: one month only, against ten months for the ESS.

[2] Complete information about the surveys can be found online: For the ESS: http://www.europeansocialsurvey.org/ and for the LISS panel: http://www.centerdata.nl/en/LISSpanel
or also http://www.lissdata.nl/assets/uploaded/Sample_and_Recruitment.pdf

However, this is linked to the panel dimension and not only to the mode of data collection. For more details, we refer to Table 1.1.

Table 1.1: **Main characteristics of the ESS round 4 and LISS study**

	ESS round 4	*LISS study*
Geographic area	Around 25 European countries (focus only on the Netherlands)	The Netherlands
Contact	Letter, followed by face-to-face	letter, followed by telephone call and/or house visit
Mode	Face-to-face at respondents house (Computer assisted)	Web
Interviewer	Yes	No
Stimulus	Oral + visual (show cards)	Visual
Panel	No (but several rounds)	Yes (but panel dimension not used)
Fieldwork period	07/09/2008 to 27/06/2009 (290 days)	December 2008 (31 days)
Sample frame	Selection of addresses, list of postal delivery points	Nationwide address frame of statistics Netherlands
Selection households	Probability sample	Probability sample
Selection individuals	Only one person is selected in the household	Only one person is selected in the household /for the study of December 2008)
Number observations	1775 interviews	Complete interviews: 3194
Response rates	52.0%	Panel membership rate: 48% Response rate of our study: 65.5% * 48% = 31.44% of the initial sample
Item non response	Higher in ESS than in LISS but still usually less than 2%	Incomplete interviews: 23 = 0.5%

What this general overview reveals is that speaking about "web" or "face-to-face" is a nice shortcut for classifying a survey, but it is an extremely simplified one. Even Table 1.1 is far from being complete. It is important to keep in mind that we are always speaking about particular surveys with specific sets of characteristics.

1.3. Composition of the samples

Even if the samples are drawn randomly, the characteristics of each survey may lead to possible selection bias and so to differences in sample composition. We assume that two main elements determine

a person's decision to participate in a web relatively to a face-to-face survey: their access to Internet and their comfort with technologies. Since the LISS respondents are provided with Internet access when they do not have it, we focus on the second factor and consider some background variables that could be related to comfort with technology: gender, age, education and number of persons in the household. The question asking about the educational achievement of the respondents differs in the ESS and the LISS. In order to be able to make a comparison, we create in each survey three quite broad categories: low, middle and high level of education.

Table 1.2 summarizes the results. We are mainly interested in comparing the fourth ESS round with the LISS study, but percentages for the LISS panel and the national statistics for the Dutch population aged 16 or more (column "pop") are also presented[3].

Table 1.2: composition of the samples (in percents)

		ESS4	LISS		pop
			study	panel	
Gender	Men	46.0	44.6	49.4	49.2
	Women	54.0	55.4	50.6	50.8
Age	16-19	4.4	2.7	7.3	6.0
	20-39	28.8	27.5	32.7	32.7
	40-64	45.5	52.3	49.4	43.3
	65-79	17.0	15.5	10.0	13.4
	>80	4.3	1.9	1.0	4.6
Education	Low	37.7	35.7	33.0	33.2
	Mid	35.6	33.2	36.9	41.4
	High	26.8	31.1	30.1	25.4
Number of household members	1	27.6	25.4	23.7	35.3
	2	35.6	39.4	35.9	32.6
	3	13.2	11.3	13.5	12.5
	4	16.6	17.0	18.9	13.5
	>5	7.2	6.9	8.0	6.0

Note: "pop" stands for "Dutch population aged 16 or more"

[3] We use the national statistics reported in a paper from the CentERdata: "The representativeness of Liss, an online probability panel" (Knoef, de Vos, 2009). http://www.lissdata.nl/assets/uploaded/representativeness_LISS_panel.pdf

For gender, the biggest differences with the actual distribution of the population are found for the LISS study: men are underrepresented whereas women are overrepresented. The same trend appears in the ESS, even if the differences are smaller.

For age, people under 39 are underrepresented whereas people between 40 and 79 are overrepresented both in the face-to-face and web surveys, even if a larger difference from the distribution of the population is again found for the LISS study.

For education, we have to be cautious when comparing the ESS and the LISS since the response options were different: this may influence the position of individuals at the border between two categories. Some may have moved from low to middle or middle to high, or vice-versa, because of the different categories. However, it seems that the group with middle educational achievement is always underrepresented. In the ESS, this underrepresentation contrasts with an overrepresentation of the low educated. In the LISS on the contrary, it is contrasted with an overrepresentation of the high educated. Even using web, low educated respondents are still well (if not over) represented in the LISS. This may be related to the age distribution of the respondents, as older respondents are usually less educated. Even if the robustness of this result can be doubted, it is interesting to highlight it, because an argument is often made that web surveys discourage less educated people from participating. In this study, it is middle educated people, and not low educated ones, that are underrepresented in the web survey. In addition, they are also underrepresented in the ESS, so the mode of data collection seems not to be the main explanation.

Finally, concerning household size, both surveys show an underrepresentation of single-member households and overrepresentation of households with more members.

In addition, using a series of chi-square tests, we can conclude that all the samples' distributions are significantly different from the population distribution for all the variables. This is not surprising as the number of observations is high, but it indicates that there are differences in the composition of the samples with respect to some background variables which may be the result of a selection bias.

So far, the different samples were compared to the population. But what is more important for our purpose is to compare them between surveys. It is a very common idea that young people will be overrepresented and old people underrepresented if the Internet is used for collecting data, because young people are more used to this new technology. Nevertheless, we see in Table 1.2 that not only the 16-20 years old, but even the 20-39 years old are significantly underrepresented in the LISS study (even more than in the ESS). On the other hand, while those over 80 years old are more underrepresented in the LISS study than in the ESS, the 65-79 years old are overrepresented in the LISS study. Similarly, for gender, the common idea that web data collection elicits more men than women to participate is not found here. The link between gender, age, comfort with technology and participation in a web survey seems not to be as clear as expected.

Comparing the composition of the LISS study and the LISS panel provides some elements of understanding: the youngest people are overrepresented in the panel, which means that when they are first approached, they are more willing to accept a web survey probably because they feel more comfortable with using Internet. However, later on, they are not very involved and so the non-response for these 16-20 years old for one specific study is quite high, ultimately leading to an underrepresentation. The 60-79 years old are, in contrast, underrepresented in the panel, but their non-response rate for one study is very small: once they agree to be part of the panel, they answer the different questionnaires sent to them, such that at the end they are overrepresented in the study.

1.4. Should we correct for these differences?

Table 1.2 shows differences in sample composition with respect to four background variables. It is difficult, however, to determine from this table alone if the differences "matter": we consider that they matter if they affect the results of the analyses. The size of the differences in sample composition is one important element: if the differences are small, they will not affect the results. Also, if the variables analysed have little correlation with the background variables, even different sample compositions will not change the results. On the other hand, if different groups on one or more of the

sample compositions' variables have very different kinds of answers than the others, even a relatively small deviation in the composition of the sample from one survey to the other may have an impact on the results.

This means we need information about the relationships between our variables of interest and the background variables. Our variables of interest are 20 variables that we will analyse in the next chapters in greater detail (the 20 present in the main questionnaire of the ESS round 4, please see Appendix 1.1 for more details). They are about position toward immigration, media use, social and political trust, satisfaction, political orientation and left-right self-placement.

A first way to look at these relationships is to consider the correlations between the variables of interest and the background variables, gender (column "g" in Table 1.3), age ("a"), education ("e") and household size ("h"). Table 1.3 presents these correlations for the LISS study and the ESS round 4 as well as the absolute value of the difference in correlations between the two surveys.

Table 1.3: correlation between variables of interest and background variable

		ESS4				*LISS*				*|difference|*			
Expt	Var.	g	a	e	h	g	a	e	h	g	a	e	h
Immigration	Imsmetn	-.01	.08	-.21	-.02	.01	-.06	-.18	.05	.02	.14	.03	.07
	Imdfetn	-.02	.10	-.24	-.02	.01	-.06	-.19	.04	.03	.16	.05	.06
	Impcntr	.05	.12	-.19	-.04	.05	-.05	-.15	.04	.00	.17	.04	.08
	Imbgeco	.09	-.03	.23	.01	.06	.04	.23	-.03	.03	.07	.00	.04
	Imueclt	-.03	-.14	.26	.04	-.03	-.02	.23	-.01	.00	.12	.03	.03
	Imwbcnt	.00	-.08	.15	.04	.03	.00	.15	-.00	.03	.08	.00	.04
Media	Tvtot	-.07	.21	-.28	-.14	-.04	.13	-.23	-.10	.03	.08	.05	.04
	Rdtot	.04	.05	-.07	-.08	.07	.04	-.16	-.04	.03	.01	.09	.04
	Nwsptot	.07	.34	.06	-.10	.13	.35	-.02	-.11	.06	.01	.08	.01
Social trust	Ppltrst	.05	.01	.21	.03	-.01	.03	.17	.02	.06	.02	.04	.01
	Pplfair	-.02	.03	.16	.04	-.07	.08	.12	.01	.05	.05	.04	.03
Political trust	Trstprl	.11	-.10	.21	.06	.03	-.01	.21	.04	.08	.09	.00	.02
	Trstlgl	.11	-.09	.25	.05	.08	-.02	.26	.03	.03	-.07	.01	.02
	Trstplc	.03	-.01	.15	.02	.01	.02	.14	.03	.02	.03	.01	.01
Satisfaction	Stfeco	.12	.04	.08	.02	.06	.04	.08	.01	.06	.00	.00	.01
	Stfgov	.06	.01	.10	.04	.01	.07	.15	.01	.05	.06	.05	.03
	Stfdem	.09	-.06	.17	.06	.03	.01	.19	.04	.06	.07	.02	.02
Political orientation	Gincdif	.09	-.11	.17	.09	.06	-.11	.20	.05	.03	.00	.03	.04
	Freehms	.05	.09	-.11	.02	.07	.00	-.13	.03	.02	.09	.02	.01
Left right	lrscale	.06	.09	-.11	.00	.03	.05	-.09	.04	.03	.04	.02	.04

Note: Table 1.6 of Appendix 1.1 indicates what the variables imsmetn, etc, refer to.

The highest absolute correlations are found for education (till 0.28 in the ESS and till 0.26 in the LISS). Except for media use where the correlations with age are also relatively high (till 0.34 in the ESS and 0.35 in the LISS), the rest of the correlations are quite low. The highest differences in correlations between ESS and LISS (column "|difference|") are between age and immigration, but even these differences are quite low (between .07 and .17). This suggests there is no much interaction effect between the mode of data collection and the background variables. On the contrary, the relationships between the background variables and the variables of interest are quite similar in both modes. If the proportions of respondents in the different gender, age, education and household size groups are not too different, few differences should be found when comparing the variables of interest in the two survey samples.

Because correlations, mainly for dummy or categorical variables with few categories, such as our background variables, have many limits (e.g. are very sensitive to marginal distributions), we also compare the distributions of variables of interest for groups of different age, gender, education and household size, in order to see if groups with different background characteristics answer differently. The significance of the difference in distributions for different groups is tested by a series of Kolmogorov Smirnov tests. Table 1.4 reports for the ESS and the LISS when the difference is significant at the 5% level. For gender, we obviously compare the distributions of our variables of interest for male and female. For age, we compare the group of the youngest respondents (less than 20) with the group of the oldest respondents (80 or more), since we expect the highest differences to be found when the groups are at the two extreme points of the distribution. For the same reason, we compare the two extreme categories for education (low and high) and for household size (single person household versus more than five persons in the household).

Table 1.4: Significance of the differences in distributions

		ESS4				LISS			
Expt	Var.	g	a	e	h	g	a	e	h
Immi-gration	Imsmetn	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*
	Imdfetn	*ns*	*ns*	s	s	*ns*	*ns*	s	*ns*
	Impcntr	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*
	Imbgeco	s	*ns*	s	*ns*	s	*ns*	s	*ns*
	Imueclt	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*
	Imwbcnt	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*
Media	Tvtot	*ns*	s	s	s	*ns*	*ns*	s	s
	Rdtot	*ns*	*ns*	s	s	s	s	s	*ns*
	Nwsptot	s	s	s	*ns*	s	s	*ns*	s
Social Trust	Ppltrst	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*
	Pplfair	*ns*	*ns*	s	*ns*	s	s	s	*ns*
Political Trust	Trstprl	s	s	s	s	*ns*	*ns*	s	*ns*
	Trstlgl	s	s	s	*ns*	s	*ns*	s	*ns*
	Trstplc	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*
Satisf-action	Stfeco	s	*ns*	s	*ns*	s	*ns*	s	*ns*
	Stfgov	s	*ns*	s	*ns*	*ns*	s	s	*ns*
	Stfdem	s	s	s	*ns*	s	*ns*	s	*ns*
Political orientation	Gincdif	s	*ns*	s	*ns*	s	*ns*	s	*ns*
	Freehms	*ns*	s	s	*ns*	s	*ns*	s	*ns*
Left right	lrscale	s	*ns*	s	*ns*	s	*ns*	s	*ns*

Note: *ns* = non significant at 5%; s = significant

Table 1.4 shows that the differences are significant in 41.7% of the cases (75/160). However, the significant differences are mainly due to education: in 97.5% of the tests, low and high educated respondents are distributed differently for the variables tested. In contrast, the different age groups are significantly different only in 27.5% of the cases. Also, it seems that greater differences are found for the behavioural variables (watching television, reading newspapers) than for the attitudinal variables. This is consistent with the previous results: the highest correlations were found between education and our variables of interest and also between the media use variables and age.

Next, the correlations between variables of interest are compared: it is very important for us to check that kind of "results" because the quality analyses are based on correlations. The same gender, age,

education and household size groups are considered. However, in this case it is more difficult to evaluate if the differences matter or not: each correlation matrix contains 0.5*20*19=190 correlations. We have a correlation matrix for each group: if we focus on the two extreme groups for each variable, we have eight groups and this in each survey (ESS and LISS). So we have 190*8*2 = 3.400 numbers.

As the goal of this chapter is just to conduct a preliminary analysis to know better the datasets we are going to use later for the quality analyses, we do not want to analyse such a large amount of data in detail. Therefore, we simply report a very crude result: there are some differences between the groups with respect to the correlations in both surveys.

Because of these differences, we create weights to try to correct for the variations in sample composition. We have the national cross-table for age and gender, so we compute weights to correct for these two variables together. We also have the national figures for education and household size, and therefore we compute other weights for these variables. We compare the matrices without weight and with the different kinds of weights.

An example is presented below in the case of the LISS study for three items about political trust measured first with an 11-point scale and then with a 6-point scale. The different correlation matrices are presented on the left of Figure 1.1, and in order to see better what is going on, on the right the deviations between the unweighted matrix and the matrices using one or another kind of weights are shown.

Very few differences are found. Even for education, where the distributions for the political trust items are significantly different (see Table 1.4), weighting has almost no impact. This does not mean that different education groups have the same correlation matrix. The weights may also make no difference because the proportion of respondents in each education group in our samples is close to the population proportions (Table 1.2).

Figure 1.1: Correlation matrices and differences due to weights (LISS, trust in politics)

LISS without weights

trust in parlement 11 points	trust in legal system 11 points	trust in police 11 points	trust in parlement 6 points	trust in legal system 6 points	trust in police 6 points
1.0000					
0.7874	1.0000				
0.7119	0.7627	1.0000			
0.7464	0.5796	0.4997	1.0000		
0.5575	0.7508	0.5844	0.6266	1.0000	
0.4063	0.4712	0.7725	0.4622	0.5636	1.0000

Differences

LISS using weights gender*age

1.0000					
0.7797	1.0000				
0.7064	0.7595	1.0000			
0.7397	0.5672	0.4832	1.0000		
0.5456	0.7545	0.5736	0.6221	1.0000	
0.4008	0.4796	0.7829	0.4481	0.5599	1.0000

LISSwithout–with gender*ageweights

0					
.01	0				
.01	0	0			
.01	.01	.02	0		
.01	0	.01	0	0	
.01	-.01	-.01	.01	0	0

LISS using weights size household

1.0000					
0.7838	1.0000				
0.7107	0.7645	1.0000			
0.7563	0.5889	0.5089	1.0000		
0.5684	0.7623	0.5978	0.6405	1.0000	
0.4120	0.4866	0.7818	0.4601	0.5685	1.0000

LISSwithout–with household weights

0					
0	0				
0	0	0			
-.01	-.01	-.01	0		
-.01	-.01	-.01	-.01	0	
-.01	-.02	-.01	0	0	0

LISS using weights education

1.0000					
0.7926	1.0000				
0.7137	0.7618	1.0000			
0.7321	0.5785	0.4942	1.0000		
0.5598	0.7509	0.5866	0.6311	1.0000	
0.4039	0.4670	0.7715	0.4594	0.5621	1.0000

LISSwithout–with education weights

0					
-.01	0				
0	0	0			
.01	0	.01	0		
0	0	0	0	0	
0	0	0	0	0	0

We do the same with the data from the forth ESS round, but this time we also have post stratification weights. These post stratification weights available in the ESS are supposed to correct for gender and age. However, they are different from the weights we computed ourselves using gender and age, because they do not divide the variable age as we did. The post stratification weights are more precise, but we keep both weights since in the case of the LISS, we cannot get more precise weights for gender and age.

Figure 1.2 presents in the case of the ESS round 4 directly the differences between the unweighted correlation matrix and the weighted ones for the same six variables about trust in politics as previously.

Figure 1.2: Differences due to the weights (ESS round 4, trust in politics)

ESS4 without – with post stratification weights	ESS4 without – with gender*age weights
0 -.01 0 -.01 -.01 0 0 0 0 0 -.01 -.01 0 0 0 0 0 -.01 0 0 0	0 0 0 0 -.01 0 0 0 0 0 0 0 -.01 .01 0 0 -.01 -.01 .01 -.01 0
ESS4 without - with education weights	**ESS4 without – with size household weights**
0 0 0 0 -.01 0 0 0 0 0 .01 0 -.01 0 0 0 -.01 0 0 0 0	0 -.01 0 0 .02 0 0 -.02 -.01 0 0 0 0 -.01 0 0 .01 0 0 .01 0

The largest differences are found in the case of the household size weight. This is not surprising knowing the ESS selection procedure: since only one individual in each household can be selected, the probability of selection of one respondent is varying depending on the size of the household he/she is living in. So in the next step, we will only focus on these household size weights.

The next and final step is to look at the estimates we are really interested in at the end, i.e. the quality estimates. This is our final criteria. Table 1.5 gives the reliability and validity coefficients for the three traits (t_1, t_2 and t_3) of the political trust experiment when three different scales (methods M_1, M_2 and M_3) are used. Table 1.5 compares these estimates for the LISS study and the ESS round 4 when household size weights are or are not used to compute the ESS round 4 correlation matrices. How these estimates are obtained and all the explanations about the MTMM analyses will be described in the following chapters.

Even focusing on the weights producing the highest differences in correlation matrices, Table 1.5 shows that the differences between the reliability and validity coefficients estimated without and with weights, for the different traits and methods, are always very small (rows in italic). Only one example has been shown here, for the political trust variables. Few differences between weighted and unweighted estimates are obtained with the variables of interest related to other topics too.

Table 1.5: reliability and validity coefficients with and without household size weights in ESS round 4 (political trust experiment)

		ESS Round4						*LISS*					
Estimates		*Reliability coefficient*			*Validity coefficient*			*Reliability coefficient*			*Validity coefficient*		
	Traits	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3	t_1	t_2	t_3
M_1	without	.86	.89	.90	.92	.92	.92	.98	.97	.99	.83	.83	.85
	Wh4	.87	.89	.92	.92	.92	.93	.98	.97	.99	.84	.84	.85
	|diff|	*.01*	*.00*	*.02*	*.00*	*.00*	*.01*	*.00*	*.00*	*.00*	*.01*	*.01*	*.00*
M_2	without	.92	.94	.92	.96	.97	.96	.91	.93	.92	1	1	1
	Wh4	.91	.93	.91	.96	.96	.96	.91	.93	.91	.99	.99	.99
	|diff|	*.01*	*.01*	*.01*	*.00*	*.01*	*.00*	*.00*	*.00*	*.01*	*.01*	*.01*	*.01*
M_3	without	.93	.92	.92	.91	.92	.92	.93	.95	.94	.90	.90	.90
	Wh4	.93	.95	.94	.92	.92	.92	.93	.95	.95	.90	.91	.91
	|diff|	*.00*	*.03*	*.02*	*.01*	*.00*	*.00*	*.00*	*.00*	*.01*	*.00*	*.01*	*.01*

Note: "Wh4" means we used weights in ESS round 4 to correct for household size. "Without" means we do not use weights. |*diff*| is the absolute difference in estimates between with and without weights. M_1, M_2 and M_3 refer to the different methods; t_1, t_2 and t_3 to the different traits.

1.5. Conclusion

What this first chapter shows is that even if the data comes from two separate surveys that have a set of specific characteristics and even if the samples of the two surveys differ from the general Dutch population and between them, it seems that the differences is sample composition do not affect much the results of the analyses, or at least that using simple weights to correct from some differences in sample composition does not affect the kinds of results we are interested in.

Indeed, overall, few differences between weighted and unweighted estimates are obtained at the correlation level. This is also what is found in other studies: for instance, Lee (2006) shows that weighting procedures often affect descriptive estimates but not correlations.

Since the analyses of quality are based on covariance or correlation matrices, it is not surprising that we also found very little difference

between the reliability and validity coefficients estimated with or without using weights.

After the preliminary checks of this chapter, we decided it was reasonable to compare the ESS and LISS samples in terms of quality and go on without using weights.

Looking up into the main characteristics of the two surveys also allows identifying some key elements of these surveys that may affect the quality. In particular, the fact that the ESS is using show cards is crucial since the stimulus for the respondents is therefore not only oral but also visual, which makes it more similar to a web survey than a face-to-face survey without show cards. Also, the fact that the LISS panel is based on a probability sample and that respondents who did not had computers or Internet access were provided with it is very important to keep in mind. The composition of the samples might have been much more different in case of an opt-in panel for instance.

In contrast to the common idea about the kind of participation that a web survey may elicit, we have seen in this chapter that when based on a probability sample and because the response rates vary for different subpopulations, the participation in one specific study as the one we are studying here is such that men and young people are underrepresented whereas low educated people are overrepresented.

Appendix 1.1

Table 1.6: the 20 variables of interest

Experiment	Name	Questions
Media		On an average weekday, how much time in total:
	Tvtot	- do you spend watching television?
	Rdtot	- do you spend listening to the radio?
	Nwsptot	- do you spend reading the newspapers?
Social trust	Ppltrst	- Would you say that most people can be trusted, or that you can't be too careful in dealing with people?
	Pplfair	- 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?
Satisfaction	Stfeco	- On the whole how satisfied are you with the present state of the economy in [country]?
	Stfgov	- Thinking about the [country] government, how satisfied are you with the way it is doing its job?
	Stfdem	- How satisfied are you with the way democracy works?
Political orientation	Gincdiff	- The government should take measures to reduce differences in income level
	Freehms	- Gay men and lesbians should be free to live their own lives as they wish
Trust in institutions		How much do you personally trust each of the institutions:
	Trstprl	- [country] parliament
	Trstlgl	- the legal system
	Trstplc	- the police
Left-right		In politics, people sometimes speak about "left" and "right".
	Lrsscale	- Where would you place yourself on that scale?
Consequences of immigration	Imbgeco	- It is generally bad for the [country] economy that people come to live here from other countries
	Imueclt	- [Country] cultural life is generally undermined by people coming to live here from other countries
	Imwbcnt	- [Country] is made a worse place to live by people coming to live here from other countries
Allowance of immigrants	Imsmetn	- [Country] should allow more people from the same race or ethnic group as most [country] people to come and live here
	Imdfetn	- [Country] should allow more people of a different race or ethnic group as most [country] people to come and live here
	Impcntr	- [Country] should allow more people from the poorer countries outside Europe to come and live here

Chapter 2

A Comparison of the Quality of Questions in a Face-to-face and a Web Survey[4]

Abstract:

Web surveys are getting more and more popular these days. However, there is a risk that changing to this new mode of research could influence the comparability of the data across time, or across surveys if some surveys switch and others don't. This project compares the quality of answers to survey questions, defined as the product of reliability and validity, in one web and one face-to-face survey using comparable samples as well as identical questions and response formats. In half of the cases, no significant differences are found. In the other half, the quality is higher in the web survey, but the differences are overall quite small.

[4] Published:

Revilla, M.A. and W.E. Saris (2012) "A Comparison of the Quality of Questions in a Face-to-face and a Web Survey" *International Journal of Public Opinion Research* 2012; doi: 10.1093/ijpor/eds007
Can be retrieved from:
http://ijpor.oxfordjournals.org/cgi/content/full/eds007?ijkey=tn8QH4xUs4goiIl&keytype=ref

2. A Comparison of the quality of questions in a face-to-face and a web Survey

2.1. Introduction

Up until now, most surveys have used face-to-face interviews, postal mail, or the telephone to collect data. However, today the difficulties of carrying out surveys at reasonable costs have increased. Simultaneously, new opportunities have appeared. In particular, web surveys, which are usually cheaper, offer more flexibility, and can reach a large population in a short time, are becoming very attractive.

Nevertheless, different modes of data collection may lead to different coverage, sampling, non response, and measurement errors. We focus on the last since different modes have different properties, just because the question is asked in a different mode, a difference in responses may appear. For instance, Krosnick (1991) shows that varying levels of social desirability and satisficing biases exist depending on the mode of data collection used. This can be related to the presence of an interviewer in some modes but not in others. As a result, in order to compare data collected with different modes (across time, across countries, across groups), we first need to study the impact of modes on several parameters.

Much research already was directed to the comparison of modes (Faas & Schoen, 2006; Fricker, Galesic, Tourangeau, & Yan, 2005; Heerwegh, 2009; Kaplowitz, Hadlock, & Levine, 2004; Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008; Schonlau et al., 2004). Nevertheless, previous research mainly focused on response rates and item nonresponse (Dillman et al., 2009; Hox & de Leeuw, 1994), or on satisficing and social desirability bias (Holbrook, Green, & Krosnick, 2003; Kreuter, Presser, & Tourangeau, 2009).

These are indicators of the quality of the surveys, but there is another way of defining quality that has not been so much studied in the frame of modes comparisons and is the focus of this research. The quality can be defined as the strength of the relationship between the latent variable of interest and the observed response. It can be computed as the product of reliability and validity.

This criterion presents advantages compared for instance to one in terms of satisficing and social desirability bias: it is not only applicable to sensitive topics but also allows for the correction of measurement errors. Few studies look at the impact of computer-assisted modes of data collection on the quality of single items defined as just mentioned. Chang and Krosnick (2009) compare the reliability in telephone and Internet surveys of reports of vote choice. They find more random errors in the telephone survey. But they are not considering the quality as a whole.

On the contrary, Scherpenzeel (1995, ch. 6) uses really the same definition of quality to compare modes and concludes (Scherpenzeel, 1995, p. 110) that ‘‘the quality of the data obtained with different computer assisted data collection techniques depends on the difficulty and sensitivity to social desirability of the topic.’’ So she did not find that one mode was systematically better than the others. Two years later, in a meta-analysis, Scherpenzeel and Saris (1997, p. 360) show that the mode has a ‘‘substantial’’ effect on both validity and reliability. ‘‘In general the worst mode of data collection seems to be CATI surveys, and the best seems to be mail surveys’’ (p. 368). CASIIP (Computer Assisted Self-administered Interviewing with interviewer present) and TI (Telepanel Interview) being in the middle but without clearly one being better than the other. However, CAPI was not considered in their study and TI is only a forerunner of the web. Nowadays, people are much more used to computers and it is therefore difficult to know if the results from the TI can be extended to modern web surveys. More recently, Saris and Gallhofer (2007) pay interest to the impact of the mode on the quality in a large meta-analysis. However, they do not consider the modes in themselves but separately code their different characteristics. They found a negative impact on the quality of having an interviewer and a small positive impact of having an oral stimulus.

The goal of this article is to compare a face-to-face and a web survey in terms of quality, defined as the strength of the relationships between the latent and the observed variables. The comparison is done between the European Social Survey (ESS), where the data is collected by face-to-face interviews at respondents’ homes, and a study completed by the respondents of the Longitudinal Internet Studies for the Social Sciences (LISS)

panel. Since the LISS is an online Dutch panel, whereas the ESS is conducted in many countries, in order to avoid variations due to cultural or language differences, we focus only on the Netherlands.

A survey cannot be described only by the fact that it is a ‘‘face-to-face’’ or a ‘‘web’’ survey. As pointed out by Couper and Miller (2008), two web surveys can be extremely different. The same applies to two face-to-face surveys. Therefore, the next section gives more information about the two surveys. Then, the method is presented, followed by the topics studied and the results. Finally, the last section proposes some elements of discussion.

2.2. The surveys

The choice to compare these surveys is practical: in one of its monthly studies (December 2008), the LISS panel proposed the same questions to its respondents as posed in round 4 of the ESS. Additionally, both surveys[5] present important similarities in the way they are implemented: At the recruitment stage, the first contact is established by sending a letter, followed by a telephone call or house visit. This is possible because both sample frames are based on postal addresses. In both cases, the selection of sampling units, that is households, is based on probability sampling. Sampling units without computer or Internet access that agree to participate in the LISS panel are provided with these facilities. Even if all the members of the household can participate in the LISS panel, only one person in each household of the LISS panel has been randomly selected to complete the study of interest, in order to make it more comparable to the ESS. It is only when respondents answer questions that the mode differs: face-to-face for the ESS, against online completion for the LISS respondents. The ESS stimulus is both oral and visual since questions are asked with show cards, whereas it is only visual in the LISS. The number of observations is 1,775 in the ESS (response rate: 52%) and around 3,200 in the LISS

[5] Complete information about the surveys can be found on their Websites. For the ESS: http://www.europeansocialsurvey.org/ and for the Liss panel: http://www.centerdata.nl/en/LISSpanel or http://www.lissdata.nl/assets/uploaded/Sample_and_Recruitment.pdf

(i.e., 65% of the panel members or 65%*48% = 31% of the initial sample).

One limitation of our approach however is that by comparing different surveys, two sources of differences may confound the analysis: differences in sample composition due to selection and differences due to the mode per se. Having the same respondents answering both by face-to-face and web would allow us to distinguish what is purely the effect of the mode on the answers. It would not, however, provide any information on the potential self-selection of respondents into different modes. Therefore, besides the fact that we did not have adequate data with the same respondents answering in both modes, comparing two real surveys provides more realistic results.

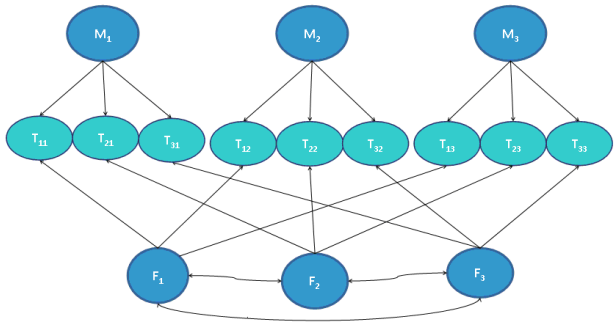
The fact that the LISS is a panel may be problematic in the sense that respondents may learn by answering questionnaires every month. Panel fatigue might appear too if respondents get bored. However, Toepoel, Das, and van Soest (2008), comparing trained and fresh respondents, find ‘‘little evidence that survey experience influences the question-answering process’’ (abstract). So all in all, the two surveys offer a good opportunity to compare identical questions asked at the same moment (the ESS round 4 was in the field from September 7, 2008 to June 27, 2009) using two modes of data collection.

2.3. A Split-Ballot Multitrait-Multimethod (SB-MTMM) approach

One common procedure used to assess the quality of measures is the multitrait-multimethod (MTMM) approach. It consists of repeating $t > 1$ questions (also called “traits”) using $m > 1$ methods: for example, the scale of the items contains five points in one method and seven in another. A $m*t$ correlation matrix among all measurements is the classic way of summarising such an MTMM dataset. Originally, Campbell and Fiske (1959) proposed to examine these matrices by comparing directly monotrait-heteromethod, heterotrait-monomethod and heterotrait-heteromethod blocs. At the beginning of the 1970s, the MTMM matrices began to be analysed using Structural Equation Models (Alwin, 1974; Jöreskog, 1970, 1971; Werts & Linn, 1970) and, in 1984 they began to be applied to

single questions by Andrews. Figure 2.1 gives an illustration of an MTMM model ($t = m = 3$).

Figure 2.1: The true score model with three traits and three methods.



Note: M_j refers to the j^{th} method factor, F_i to the i^{th} trait, T_{ij} to the true score associated with the i^{th} trait and j^{th} method.

The main limitation of this approach is that in order to get an identified model, each question needs to be repeated at least three times for the same respondent. In order to avoid memory effect, 20 minutes of similar questions needs to separate each of the repetitions (Van Meurs & Saris, 1990). If there are three methods, it means that the questionnaires have to be longer than 40 minutes. This is not always possible in practice. Besides, getting three repetitions of the same question may increase the cognitive burden of the respondents. Therefore, Saris, Satorra and Coenders (2004) propose to combine the advantages of the MTMM approach with those of the Split-Ballot (SB) approach, randomly assigning the respondents to different groups assures the comparability of the results and at the same time it reduces the number of repetitions for each respondent (only two). The model is still identified under quite general conditions. The ESS use a two-group SB design: all the respondents get method 1 (M_1) the first time. Then, one half gets method 2 (M_2) and the other half method 3 (M_3). In contrast, the LISS use a three-group design: one third of the respondents get M_1 and M_2, one third M_2 and M_3 and one third M_1 and M_3.

Using this SB-MTMM design, the reliability and validity coefficients can be obtained for each question using the true score model developed by Saris and Andrews (1991). Other models could be used (e.g. multiplicative model originally suggested by Browne, 1984) but Corten et al. (2002) showed that the additive model of

Saris and Andrews (1991) was preferable. We therefore use this model:

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad \text{for all } i,j \quad (1)$$
$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad \text{for all } i,j \quad (2)$$

Where, F_i is the i^{th} trait, M_j the variation in scores due to the j^{th} method, and for the i^{th} trait and j^{th} method, Y_{ij} is the observed variable, r_{ij} is the reliability coefficient, T_{ij} is the true score or systematic component of the response, e_{ij} is the random error associated with the measurement of Y_{ij}, v_{ij} is the validity coefficient, and m_{ij} is the method effect coefficient. We also assume that the trait factors are correlated with each other; the random errors are *not* correlated with each other, nor with the independent variables in the different equations; the method factors are *not* correlated with each other, nor with the trait factors; the method effects for a specific method M_{j*} are equal for the different traits T_{ij*} (for all i); the method effects for a specific method M_{j*} are equal across the split-ballot groups; as are the correlations between the traits, and the random errors. These assumptions are the ones we start with but if some of them do not hold, they are relaxed by estimating extra parameters.

The quality of a measure can be derived from this model as the product of the reliability (square of the reliability coefficient) and the validity (square of the validity coefficient), so: $q_{ij}^2 = r_{ij}^2.v_{ij}^2$. It corresponds to the strength of the relationship between the variable of interest F_i and the observed answer Y_{ij} expressed for the j^{th} method.

2.4. Selection of topics

An MTMM approach requires a specific dataset with repeated questions. For our purpose, different modes are also needed. Both the ESS round 4 and the LISS study included SB-MTMM experiments. We analyse all six available SB-MTMM experiments. They are about time spent on different media, satisfaction, political orientation, social and political trust and left-right orientation. Each experiment contains three items usually measured with three different methods.
Table 2.1 gives more information about the items (t_1, t_2, t_3 in "wording of the questions") and methods (M_1, M_2 and M_3). In two

cases, one of the methods was different in the two surveys: therefore the method for the ESS is mentioned in brackets. The sign Ø in the column "var" (name of the variables in the ESS dataset) means that the variable is missing in the main questionnaire of the ESS.

Table 2.1: The SB-MTMM experiments studied (different traits t_i and methods M_i)

	Var.	Wording of the questions	M_1	M_2	M_3
Media	tvtot rdtot nwsptot	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?	8 pts	Hours and min	7 pts
Satisfaction	stfeco stfgov stfdem	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?	11 pts (extr)	11 pts (very)	5 AD
Political orientation	gincdif freehms Ø	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 t_3 = The government should ensure that all groups in society are treated equally	5 AD	5 pts	5 pts *(AD in ESS)*
Social trust	ppltrst pplfair Ø	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? t_3 = Would you say that most people deserve your trust or that only very few deserve your trust?	11 pts	6 pts	2 pts
Political trust	trstprl trstlgl trstplc	How much do you personally trust each of the institutions: t_1 = Dutch parliament t_2 = The legal system t_3 = The police	11 pts batt	6 pts batt	11 pts score
Left right	lrscale Ø Ø	In politics people sometimes talk of "left" and "right". t_1 = Where would you place yourself on this scale? t_2 = Where would you place the party you most like? t_3 = Where would you place the party which you most dislike?	11 pts	11 pts (extr)	11pts extr all *(=M1 in ESS)*

Note: Pts = number of response categories; extr = labels of the end points start with "extreme(ly)", e.g. "extreme left" to "extreme right"; AD = agree-disagree scales; batt = the question is part of a battery of more questions; all = the scale is fully labelled; Ø = missing in the main ESS questionnaire. Complete formulation of the questions can be found online: http://www.europeansocialsurvey.org/

2.5. Analyses and results

For each experiment, the first step is to obtain the MTMM covariance or correlation matrices. This is done using ordinary Pearson correlations (for an explanation of this choice, see Coenders and Saris, 1995) and the pairwise deletion option in R for missing and "Don't Know" values (which are very few). The estimates are then obtained analysing these matrices with LISREL (Jöreskog and Sörbom, 1991) by Maximum Likelihood (ML) estimation for multi-group analysis. See Saris et al. (2004) for details on the estimation procedure and the distributional assumptions. In order to test if there are misspecifications, we use the JRule software (Van der Veld, Saris, & Satorra, 2009) based on the procedure developed by Saris, Satorra and Van der Veld (2009). JRule has the advantage of taking into account the power (one minus the type II errors). Also, the program tests for misspecifications at the parameter level instead of testing the model as a whole, so it is more precise. Based on the program suggestions, some corrections with respect to the general model presented earlier are introduced (extra-correlations between similar methods, unequal effects of one method on the different traits or allowing the method effects to vary across surveys)[6].

We estimate the model with five groups: three SB groups for the LISS and two SB groups for the ESS. This has two main advantages: first, it allows testing the significance of the difference between the estimates of the two surveys by adding constraints on the parameters (should be invariant[7]). Second, as some experiments are incomplete in the ESS (variables missing in the main questionnaire) but not in the LISS, it helps to identify the models and get convergence. Table 2.2 provides, for each topic, the quality measure for the three traits (t_1, t_2, t_3), as well as the mean quality among these traits, for each of the methods (usually three; two when M_3 varies), both for the ESS and the LISS[8].

[6] A list of the modifications made can be retrieved from: http://bit.ly/hbHCEG
[7] An example of LISREL input can be retrieved from: http://bit.ly/i1oSZK
[8] An overview of the reliability and validity coefficients can be retrieved from: http://bit.ly/nW3yOy

Table 2.2: Quality estimates in the ESS (2008/2009) In the Netherlands and the LISS (December 2008) for the different traits (t_i) and methods (M_i)

	Total Quality	*ESS 4*				*LISS*			
Experiments	Method	t_1	t_2	t_3	mean	t_1	t_2	t_3	Mean
Media	M_1 = 8pts	.90	.76	.90	.86	.90	.76	.90	.86
	M_2 = h/min	.30	.68	.24	.41	.30	.68	.24	.41
	M_3 = 7pts	.41	.78	.47	**.55**	.41	.80	.48	**.56**
Satisfaction	M_1 = 11 extr	.56	.73	.67	**.65**	.63	.80	.78	**.73**
	M_2 =11 very	.80	.83	.78	**.80**	.87	.89	.85	**.87**
	M_3 = 5AD	.44	.67	.57	**.56**	.48	.70	.60	**.59**
Political orientation	M_1 = 5AD	.60	.56	.60	.59	.60	.56	.60	.59
	M_2 = 5 pts	.76	.89	.66	.77	.76	.89	.66	.77
Social trust	M_1 = 11 pts	.74	.61	.81	.72	.74	.61	.81	.72
	M_2 = 6 pts	.67	.57	.68	.64	.67	.57	.68	.64
	M_3 = 2 pts	.55	.50	.57	.54	.55	.50	.57	.54
Political trust	M_1=11 batt	.63	.67	.69	**.66**	.66	.65	.71	**.67**
	M_2 = 6 batt	.78	.83	.78	**.80**	.83	.86	.85	**.85**
	M_3=11score	.72	.72	.72	.72	.70	.73	.72	.72
Left right	M_1 = 11 pts	.85	.80	.73	**.79**	.94	.88	.81	**.88**
	M_2 =11 extr	.89	.83	.85	**.85**	.94	.90	.85	**.90**

Note: Pts = number of response categories; extr = labels of the end points start with "extreme(ly)", e.g. "extreme left" to "extreme right"; "batt"=the question is part of a battery of more questions; in bold= when the average quality differs across the ESS round 4 and the LISS. Total sample size: around 1,775 in the ESS and 3,200 in the LISS.

In half of the cases, no significant differences are found. For example, for the first two methods of the media experiment, the quality is the same for the three traits. So it seems that whatever the mode, the variance in the observed variables explained by the traits of interest is identical. The same is true for all methods and traits of the political orientation and social trust experiments and for the third method of the political trust experiment. In the other cases, some differences are found between the surveys. Even when significant, however, these differences are small: for example, for M_3 of the media experiment the average quality over the traits is 0.55 in the ESS versus 0.56 in the LISS. This difference is too minimal to worry about. The experiment where the most substantial differences are found is the one about left-right positioning, with a difference of 0.09 for M_1. Even in that case, the general trend is the same: the order in quality of the methods in the experiment is the same in both surveys ($q^2_{M1} < q^2_{M2}$). It is interesting also to notice

that when differences are found, they are systematically in favor of the Internet survey: the LISS has a higher quality than the ESS.

Another important finding is that differences between methods matter much more than differences between data collection modes: in the ESS, there is a difference of 0.24 between the mean quality for satisfaction in M_2 and M_3; in the LISS this difference is even higher, 0.28. This confirms results of previous studies showing that agree-disagree scales perform poorly in terms of quality (Saris, Revilla, Krosnick, & Shaeffer, 2010). This appears to be true not only for face-to-face data collection, but also for web surveys. Similar patterns can be found in all experiments and for different scales. In some cases, there are even more extreme differences: for instance in the media experiment, the difference between M_1 and M_2 for the first trait is 0.60 both in the face-to-face and the web surveys. This is a huge difference and the quality of M_2 is very low, but this is not the result of the mode used, since a similar difference is found in both surveys. It is due to other choices made in designing the items and response scales. This means that the observed correlations between the time spent watching television and any other variable will vary significantly depending on which of these two methods is used, but will not vary because the surveys have different modes.

2.6. Discussion

In this comparison of a face-to-face and a web survey, the SB-MTMM analyses show that the quality does not vary much depending on the mode, but when it varies it is higher in the LISS. However, the quality varies much more with the method used than with the mode: for the media experiment for instance, if the time is asked in hours and minutes, the quality is only 0.41, whereas when categories are used (less than ½ hour, ½ hour to 1 hour, etc) the quality is more than double (0.86). The differences in quality between these methods are similar in the two surveys.

This is in accordance with previous research taking into account modes and characteristics of questions: for instance, Andrews (1984) finds a very small effect of the mode of data collection but important effects of the characteristics of the questions.

Scherpenzeel and Saris (1997) also find a larger effect of some question characteristics (e.g. combination of form and length of the response scale) than of the mode of data collection. However, previous research focused on different modes of data collection. Our results suggest that the same pattern applies when comparing face-to-face and web surveys, at least for the two surveys considered.

Thus, overall, we can conclude that the quality of single items seems to be quite similar in these two surveys. On the whole, the mode effect expected on the quality is not found. The results suggest that at least in the Netherlands, switching from face-to-face to web data collection could be done without threatening the comparability if one is interested in relationships. Knowing that data collection is much quicker with web and usually less expensive, a switch to web survey seems quite attractive.

This does not mean that all face-to-face and all web surveys have the same quality. We have to be careful about generalizing our conclusions because two face-to-face or two web surveys can differ greatly. The LISS survey is very different from opt-in web surveys. All the efforts made to increase the representativeness of the LISS panel (in particular, providing a computer and Internet connection if necessary to a probability sample) make it different from other web surveys. The ESS also has specific properties that make it different from other face-to-face surveys. This also does not mean that only the ESS and the LISS have the same quality. We believe that our results can be generalized to other face-to-face and web surveys, as long as they share the main properties of the ESS and the LISS (use of show cards, computer-assisted interviewing for the face-to-face survey, or probability sample for both). Furthermore, we can conclude that the mode in itself does not systematically lead to results that are not comparable and that web surveys can have a high quality of measure.

Further research is needed at different levels. In particular, we think more attention should be paid to the quality of complex and sensitive questions. As mentioned before, Sherpenzeel (1995) concluded that the quality obtained in different modes depends on the difficulty and sensitivity of the topic. We should look if this result holds when the web is compared to other modes. Except for

the topic media that is to some extent sensitive and where some methods (in hours and minutes) are a bit more complex, this note did not really consider sensitive topics, and also not complex questions. More should be done on that point.

Chapter 3

Impact of the mode of data collection on the quality of answers to survey questions depending on respondents' characteristics[9]

Abstract

The Internet is used more and more 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 respondents answer. In previous research, Revilla and Saris (2012) 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 this 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 for Web surveys whereas for others it is lower, such that on an average the quality for the complete sample is similar? Comparing the quality for different groups of respondents in a face-to-face and in a Web survey, no significant impact of the background characteristics, the mode and the interaction between them on the quality is found.

[9] Accepted (but not published yet): *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*

3. Impact of the mode of data collection on the quality of answers to survey questions depending on respondents' characteristics

3.1. Modes of data collection and quality

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

This democratisation of the survey practice has been accompanied by increasing concern about the representativeness and the quality of different surveys. If many people are able to conduct surveys, 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 that is not always the case. High quality surveys such 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. The mixed-mode approach has the advantage that the non-Internet users – that still represent a non-negligible part of the population – can participate via another mode. However, introducing the Internet may threaten the comparability of the data both across time and across groups (or 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 associated with Web surveys, an important literature began comparing Web to other modes of data collection. The comparisons focused mainly on the response rates and non-response (Fricker et al., 2005; Kaplowitzet al., 2004) and on satisficing and social desirability (Heerwegh, 2009; Kreuter et al., 2009) as indicators of quality. Satisficing and social desirability may be observed in all modes, but they are

expected to vary because of the presence of the interviewer in some modes and not in others.

Nevertheless, low response rates are only a warning of potential troubles (Couper and Miller, 2009): they do not systematically correspond to low quality. On the other hand, higher response rates imply neither higher representativeness, nor higher quality (Krosnick, 1999). The central question is whether higher response rates also mean less non-response bias (Voogt and Saris, 2005). Satisficing and social desirability are specific to certain kinds of questions and, as such, are not adapted for measuring 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 error and if necessary correct for it.

Defining quality in the same way, two papers (Revilla, 2010; Revilla and Saris, 2012) recently focused on the impact of the mode, or combination of modes, of data collection on the quality of answers to survey questions. The main result is that the quality is very similar in the face-to-face and the Web surveys compared. From that, this author concludes “that there is only a slight impact” on the quality when switching from a unimode to a mixed-mode design for the data analysed (Revilla, 2010: 163).

This conclusion may be a bit too optimistic: does the finding of a similar average quality in both modes really allow us to conclude 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 answers of respondent *i* remains the same if respondent *i* takes 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 topic of this paper. We focus on two modes: Web, because of its impressive growth during the past few 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 3.2 discusses the assumption of equal impact of the mode on all respondents. Then, section 3.3 proposes a set of hypotheses. Section 3.4 explains the model used to test these hypotheses while section 3.5 gives information on the data. Finally, section 3.6 gives the results and section 3.7 concludes.

3.2. (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 and Christian, 2005; Toepoel et al., 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[10] and find that "older respondents and those with less schooling provided the 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, we want 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.

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

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 and Loosveldt, 2009; Smythet al., 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.

3.3. 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[11] 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 et al., 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.

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

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.

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.

3.4. Method

3.4.1. 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.

Van Meurs ans Saris (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 et al. (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. 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[12] 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 et al, 2009), a software based on the procedure developed by Saris et al. (2009) that allows testing for misspecifications at the parameter level and using both type I and II errors.

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[13].

3.4.2. 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 3.1.

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.

---

[12] 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).

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

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[14]), 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".

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

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

$$q^2_{educ} = \alpha_{educ} + \beta_{0,educ}Low + \beta_{1,educ}High + \beta_{2,educ}Web + \beta_{3,educ}Low*Web + \beta_{4,educ}High*Web + \xi_{educ} \quad (5)$$

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[15] (numerical) and the kind of scales (dummy "IS" equals to one if the scale is Item Specific[16], 0 otherwise). See for example Saris and Gallhofer (2007) for more details (definitions of these terms, effects on the quality, etc).

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

[15] We call "fixed reference point" a response category that uses a label that lets no doubt about the position of the response category on the subjective scale in the mind of the respondents (for instance a label that makes it clear for all respondents that the answer category is a neutral point or is the most extreme point possible).

[16] A scale is Item Specific if the categories used to express the opinion are exactly those answers we would like to obtain for this item (by opposition for instance to Agree-Disagree scales).

3.5. Data

3.5.1. 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[17]. 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[18]. Both samples are quite similar in terms of gender, age and education distributions (see for instance Revilla and Saris, 2010).

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 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

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

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 approaches 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 3.2).

3.5.2. 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 3.1 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.

Table 3.1: traits and methods for each of the 6 MTMM experiments

Topic	*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 = 11pts battery M_2 = 6 pts battery M_3 = 11 pts 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 pts (extr) M_2 = 11 pts (very) M_3 = 5 pts AD
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 M_2 = hours-min M_3 = 7 categories
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 pts M_2 = 6 pts M_3 = 2 pts
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 AD M_2 = 5 pts
Left right	In politics people sometimes talk of "left" and "right". t_1 = Where would you place yourself on this scale?	M_1 = 11 pts M_2 = 11 pts (extr)

Note: "pts"=number of response categories; "extr"= extreme words used in the labels of the end points; "AD"= agree/disagree scale

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 3.2). 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.

3.6. Results

3.6.1. Results for gender (H2a)

Table 3.2 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).

Table 3.2: 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: * if significant at 10% level; ** if 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 3.2 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*.

3.6.2. Results for age (H1a, H2b, H3a)

Table 3.3 is similar to Table 3.2, but provides the results for age.

Table 3.3: 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 3.3 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 3.3 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.

3.6.3. Results for education (H1b, H2c, H3b)

The same information is displayed for the education analyses in Table 3.4.

Table 3.4: 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

In Table 3.4, we see no significant impact for 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.

3.6.4. Summary

In sum, the signs in the regressions ("reg 2" in Tables 3.2, 3.3 and 3.4) 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, item specific, 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.

3.7. 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 answers to 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 (2012) and Revilla (2010) was valid: at least for the different gender, age and education groups tested, 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. It is important to note nevertheless that web panels may be very different from each other: our study is based on a probability-based

web panel, and cannot be generalised to the vast majority of web surveys conducted nowadays with opt-in panels but only to other web probability-based web panels that are making a lot of efforts to get a representative sample and high quality data.

Appendix 3.1

Table 3.5: quality estimates

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

Experiment		*Social trust***			*Political orientation***		*Left Right**	
Group/Method		M_1	M_2	M_3	M_1	M_2	M_1	M_2
Liss	*men*	.65	.66	.53	.55	.83	.87	.89
	women	.65	.66	.53	**.59**	.83	*.87*	.89
Ess	*men*	.65	.66	.53	*.59*	.83	**.87**	.89
	women	.65	.66	.53	.59	.83	.80	.89
Liss	*<60*	.75	**.60**	*.53*	.57	.84	*.92*	***.92***
	≥60	*.75*	.53	*.53*	*.57*	.84	*.92*	*.90*
Ess	*<60*	**.75**	.60	.50	**.61**	.84	**.87**	**.90**
	≥60	.65	*.60*	.50	.52	.84	.83	.78
Liss	*Low*	.69	.59	.52	.57	**.85**	.92	.90
	Mid	.69	.59	.52	.57	**.85**	.92	**.92**
	high	.69	.59	.52	.57	.75	.92	**.92**
Ess	*Low*	.69	.59	.47	.57	.85	.73	.87
	Mid	.69	.59	**.52**	.57	.85	**.83**	**.90**
	high	.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.*

Appendix 3.2

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

Table 3.6: number of respondents using the Internet less than once a month and more than once a month in both surveys

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

Table 3.7: distribution of respondents' gender, age and education level in both surveys

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

Chapter 4

Measurement invariance and quality of composite scores in a face-to-face and a web survey[19]

Abstract

Measurement equivalence is a pre-requisite to be able to make comparisons across groups. In this paper we are interested in testing measurement equivalence across respondents answering surveys done using different modes of data collection. Indeed, different modes of data collection have specific characteristics that may create measurement non-equivalence across modes. If this is so, data collected in different modes cannot be compared. This would be problematic since, in order to respond to new challenges, like costs and time pressure, more and more often researchers choose to use different modes to collect their data across time, across surveys, and across countries. Studying data about trust and attitudes towards immigration, this paper shows that measurement equivalence holds across a face-to-face and a web survey done in the Netherlands (2008-2009). Moreover, the quality estimates of the Composite

[19] Under review

Scores are quite high and pretty similar in the two surveys for the four concepts considered.

4. Measurement invariance and quality of composite scores in a face-to-face and a web survey

4.1. Introduction

Measurement equivalence, if it holds, refers to the fact that two individuals with the same true opinion or attitude (or one individual at two occasions) will give the same answer when asked the same question. This may seem obvious but there are in fact a lot of reasons why measurement equivalence might not hold.

Following the terminology of Northrop (1947), a distinction can be made between concepts by postulation (CP) and concepts by intuition (CI). Concepts by postulation are complex concepts that cannot be directly measured but instead are defined by several concepts by intuition. These CPs are represented by latent variables in the models. The concepts by intuition are simple concepts that can be directly measured by items (Saris and Gallhofer, 2007). For instance, political trust is a concept by postulation, a broad concept that can be operationalized by identifying and specifying its different components. Thus, political trust can be decomposed into different CIs: trust in the parliament, trust in the legal system, trust in the police, etc. Each of these CIs can be measured by one single question.

Many concepts studied in social sciences are too complex to be measured by single items. Therefore a lot of studies are based on analyses of CPs. Measurement equivalence is usually assessed at this level. But researchers do not always work with latent variables to assess the CPs. They often combine several items (observed answers to the questions) in some kinds of average scores usually called Composite Scores (CS) or Indices (e.g. Anderson, Lepper and Ross, 1980; Peterson et al., 1982; Duckworth and Seligman, 2006; etc). Composite Scores are combinations of observed scores that are used as shortcuts to measure the CPs of interest. But these CSs are not perfect measures of the CPs. The strength of the

relationship between the CP and the CS can be computed: this corresponds to the quality of the CS (Saris and Gallhofer, 2007). It gives an indication of how well the CS measures what one really intends to measure by telling how much of the observed variance of the CS is explained by the variance of the CP.

Why should we care about measurement equivalence and quality of CS? We should care because it is a pre-requisite to be able to make comparisons between groups. Observed differences can come from true differences or from a lack of measurement equivalence. At the same time, observed similarity does not guarantee that there are no true differences: the true differences can be cancelled out by differences in the measurement leading to similar observed results. So if measurement equivalence is not assessed first, comparative research cannot be done.

Measurement equivalence is most often discussed in the frame of cross-national research (e.g. Singh, 1995; Steenkamp and Baumgartner, 1998). The idea is that countries have different cultures that make people express themselves differently. The typical cliché is that southern countries are much more willing to use extreme words and to be excessive ("fantastic", "horrible") while northern countries are famous for their understatements ("not too bad", "a bit unpleasant"). If people of different countries express themselves in different ways, then two people with the same opinion can choose different answer categories depending on which country they belong to. Besides the culture, problems in translation may also be a threat to measurement equivalence across countries or language groups (Dumka et al., 1996).

However, cross-national research is not the only context where comparisons are made. Comparisons may also be done across groups of respondents with different characteristics (Schulenberg et al., 1988; Tansy and Miller, 1997), across surveys, etc. Our interest is in comparisons across modes of data collection.

First, focusing on modes of data collection is important because different modes have different characteristics: for a more complete overview, we refer to de Leeuw (2005) or Dillman et al. (2009). Here, we only underline a few elements. One difference is that some modes are self-completed (postal mail, web) whereas in others

an interviewer is present (face-to-face, telephone). The presence of the interviewer may lead to higher social desirability bias, i.e. over-reporting of socially desirable attitudes or opinions and under-reporting of the undesirable ones. For example, Kreuter, Presser and Tourangeau (2009) find that web surveys increase the reporting of sensitive information relative to telephone surveys. Since face-to-face surveys also require the mediation of an interviewer, it can be expected that web surveys also increase the reporting of sensitive information relative to face-to-face surveys. Consequently, people with the same true score can pick different answer categories, disturbing measurement equivalence. In particular, the observed means of the variables for socially desirable (respectively undesirable) attitudes are expected to be higher (respectively lower) in presence of an interviewer than in self-completed modes.

Another difference between modes is the kind of stimuli they elicit. Some modes are associated with visual stimuli (postal mail, web) whereas others are associated with oral stimuli (face-to-face, telephone). However, a combination of both visual and oral stimuli is possible (e.g. face-to-face using show cards or web surveys with added voice). Depending on the nature of the stimuli, different ways of answering the questions can be expected. Krosnick (1991) argue that many respondents choose to satisfice, i.e. to minimize their efforts in responding to questions while providing the appearance of compliance. When the answer categories are presented visually, this may lead to primacy effects, which is a bias toward selecting earlier response options instead of considering carefully the entire set of responses. On the contrary, in oral modes, because of memory limitations, respondents are expected to choose more often the last answer categories. This is referred to as "recency effect" (Smyth et al, 1987). Again, this may threaten measurement equivalence across modes.

Secondly, studying equivalence across modes is important because, nowadays, different modes are available to conduct surveys. Each of them has some strengths and weaknesses and it is difficult to say if one is better than the others. It depends on time and costs constraints, on countries' customs for surveys, on the availability of sampling frames, on the coverage of the population for certain modes (e.g. availability of access to the Internet), and on the length of the survey, the topic, its sensitivity, etc. As a result,

several modes are regularly used nowadays. Comparing results from surveys using different modes, or results from the same survey at two different points in time after a switch of modes occurred, cannot be done without first assessing if measurement equivalence holds. Besides, some surveys try to solve the problems of low response rates by combining modes within one single survey. In this kind of mixed-mode surveys, it is again crucial to assess measurement equivalence across modes in order to be able to combine the data coming from the different modes.

Finally, there is quite some interest in comparing modes, but usually the focus is on comparing response rates (Hox, De Leeuw, 1994; Fricker, Galesic, Tourangeau, Yan, 2005) or social desirability bias (Tourangeau, Smith, 1996; Kreuter, Presser, Tourangeau, 2009). Not much is known about measurement equivalence across modes. King and Miles (1995) look at the measurement equivalence of data collected from paper-and-pencil and computerized formats. Cole, Bedeian, Field (2006), as well as De Beuckelaer and Lievens (2009), test measurement equivalence across paper-and-pencil questionnaires and web surveys. All these analyses find strong support for measurement equivalence, but they are focusing on self-completed modes with only visual stimuli. Does measurement equivalence still hold when an interviewer is present in one mode but not in the other? And when the stimuli are visual in one mode but both visual and oral in another?

The goal of this paper is to investigate whether measurement equivalence holds for different topics in two surveys, one conducted face-to-face in the respondents' house and one online. The analyses also look at the quality of different composite scores. As far as we know, research on that point is still missing from the literature, so here is a second contribution of our research to the literature. The surveys and topics are presented first, followed by some information about the method, and then the results. Finally, some general conclusions are drawn, together with limits and ideas for future research.

4.2. The surveys and topics

4.2.1. The surveys: European Social Survey (ESS) versus Longitudinal Internet Studies for the Social sciences (LISS) panel

The comparison is made between two surveys using different modes of data collection, but collecting the data in the same period (end 2008-beginning 2009) in the same country (The Netherlands[20]) and on probability-based samples drawn from a frame of postal addresses. Similar questionnaires were asked to the respondents (same wording of the questions, same scales).

The first survey is round 4 of the ESS. Many things could be said about this survey[21] but what is most relevant for our analyses is that it is a face-to-face survey conducted by an interviewer at the respondent's home and using show-cards. Slightly fewer than 1800 respondents completed the survey in The Netherlands, which corresponds to a response rate of 52%. The second survey is one completed by almost 3200 members of the LISS panel, which is a Dutch web panel[22]. This represents 65.5% of the panel members, and 31.5% of the initial sample.

The Netherlands currently have one of the highest Internet penetration rates of Europe, with 88.3% of the population having access to the Internet in 2011[23]. Compared to other countries, its population is in average more web-literate, but it is quite similar to the situation of Nordic countries (e.g. Sweden or Denmark), and within a few years we can expect more countries to present a similar profile. Therefore, it is an interesting country to investigate.

4.2.2. The topics: trust and attitude toward immigration

Four concepts related to two different topics are used for the comparison. First, the topic of trust has been chosen because many

[20] The ESS is conducted in many more countries but we focus on the Dutch data.
[21] More details can be found on the ESS website: http://www.europeansocialsurvey.org/
[22] More details can be found on the LISS website: http://www.centerdata.nl/en/MESS
[23] http://www.internetworldstats.com/stats9.htm#eu

influential scholars, from Hobbes to Weber, defend the idea that trust is essential for social, economic, and political life, at the micro and macro levels. Newton (2007, p.356) summarizes that: "trusting individuals are said to live longer, happier, and more healthy lives; high-trust societies are said to be wealthier and more democratic; trusting communities are supposed to have better schools and lower crime rates". As a consequence, trust is a central concept for political and social science research. Moreover, trust can be divided into two sub-concepts, social and political trust, because "people may trust those around them and not their political leaders" (Newton, 2007, p.344). Social trust and political trust are complex concepts. These are the first two CPs that we are going to analyse ("soctrust" and "trustin")[24].

The second topic, attitude toward immigration, gained prominence because of the growth of this phenomenon and of the problems related to it: social tensions and conflicts, racism, assimilation of new comers, etc. Most European countries (EU-15) have sizeable immigrant populations today. Consequently, the attitudes of the citizenry towards newcomers have recently been much studied (e.g. Coenders, 2001; Mayda, 2006). This topic has also been chosen because it is one of the most sensitive topics in the core questionnaire of the ESS round 4. As such, social desirability bias may be expected to be higher in a face-to-face survey than in a web survey (no interviewer). Two concepts related to attitudes toward immigration are present in the ESS and LISS data. The first measures the evaluation of the consequences of immigration: the higher the score of respondents on this variable, the more favourable are their opinions about the impact of immigration. Since the scale goes from negative to positive evaluations, we will call this variable "positive". On the contrary, the second latent variable measures the reluctance of respondents to allow more people to come to the Netherlands. The higher the score on this variable, the less willing people are to accept more immigrants. Therefore, we will call this variable "not allow". These are the third and fourth CPs that we are going to analyse.

[24] It may be argued that in fact the questions cover only a sub-concept of social trust sometimes referred to as "generalised trust" (see for instance Uslaner, 2002) and only a sub-concept of political trust that could be called "trust in institutions", but for simplification purposes, we will call them "social" and "political" trust.

Each of these four CPs has several reflective indicators. The CP of social trust has two indicators: how much the respondent thinks people can be trusted and how much he or she thinks that people try to be fair. The three other CPs have three reflective indicators. For political trust, they correspond to the trust in the parliament, in the legal system and in the police. For the evaluation of the consequences of immigration, they correspond to the opinion that immigration is good for the economy, that it enriches culture life, and that it makes the Netherlands a better place to live. Finally, for the reluctance of allowing more people to come and live in the Netherlands, each indicator asks for a different group of immigrants: people from the same race or ethnic group as most Dutch people, people from a different one and people from poorer countries outside Europe.

The names of the variables in the ESS dataset, the wording of the questions and characteristics of the scales can be found in Table 4.1.

Table 4.1: Experiments about trust and immigration

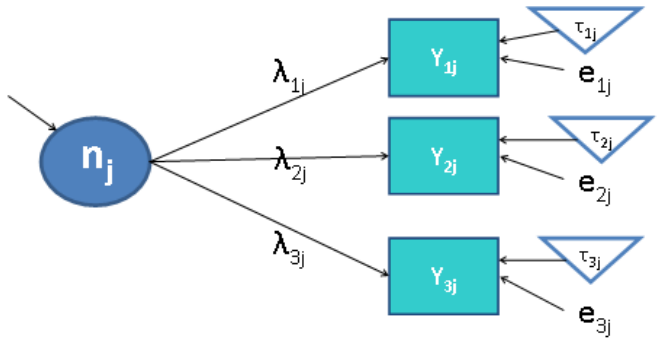
Concept	Var.	Meaning	Method
Soc trust	ppltrst	- Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?	11 points (from negative to positive)
	pplfair	- Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair?	
trust in		How much do you personally trust each of the institutions:	11 points (no trust to complete trust)
	trstprl	- Dutch parliament	
	trstlgl	- The legal system	
	trstplc	- The police	
Positive	imbgec	- It is generally bad for the Dutch economy that people come to live here from other countries	11 points (from negative to positive)
	imueclt	- Dutch cultural life is generally undermined by people coming to live here from other countries	
	imwbcn	- The Netherlands are made a worse place to live by people coming to live here from other countries	
not allow	imsmet	- The Netherlands should allow more people of the same race or ethnic group as most Dutch people to come and live here.	4 points (allow many to allow none)
	imdfctn	- The Netherlands should allow more people of a different race or ethnic group from most Dutch people to come and live here.	
	impcntr	- The Netherlands should allow more people from the poorer countries outside Europe to come and live here.	

4.3. Method

4.3.1. Testing for measurement equivalence

This section presents how to test for measurement equivalence (also called invariance) across groups for concepts with reflective indicators. The basic measurement model used is presented in Figure 4.1.

Figure 4.1: the basic measurement model



Note: η_j is the latent variable; Y_{ij} are the i observed variables for the latent trait j; λ_{ij} are the loadings; τ_{ij} are the intercepts; and e_{ij} are the random error terms.

In this model, η_j is the j^{th} latent variable of interest (the CP), the Y_{ij} are the i^{th} observed variables corresponding to the i^{th} CIs for the j^{th} latent variable of interest, the λ_{ij} are the loadings, the τ_{ij} are the intercepts and the variables e_{ij} represent the random components. It is usually recommended to have at least three indicators for each CP (e.g. see Saris and Gallhofer, 2007, Chapter 16). The model can also be expressed with a system of equations:

$$Y_{ij} = \tau_{ij} + \lambda_{ij}\,\eta_j + e_{ij} \quad \text{for all } i,j \qquad (1)$$

Each equation is similar to a regression equation, where Y_{ij} is the dependent variable that one tries to explain, η_j is the independent or explanatory variable, τ_{ij} is the intercept or value of the dependent variable when the independent variable is 0, λ_{ij} is the slope, i.e. the increase in Y_{ij} expected for each one unit increase in η_j, and e_{ij} is the error term. Basic assumptions are made: the error terms are

assumed not to be correlated with the independent variables, nor with each other. The different latent variables (η_j) are assumed to be correlated with each other.

In order to fix the scale of the latent variables, for each CP, one of the loadings, e.g. the one of the first observed variable (λ_{1j}), is fixed to 1 and one of the intercepts, e.g. τ_{1j}, is fixed to 0.

The same model is specified in the different groups that one wants to compare: in our case, the face-to-face and the web surveys. Using a multiple-group confirmatory factor analysis approach, it is possible to test for different levels of equivalence, by putting more or less constraints of equality on the parameters across groups.

We sequentially test for the three more common levels of invariance (Meredith, 1993):

- configural invariance: the same measurement model holds in all groups (i.e. in the different modes)
- metric invariance: configural invariance holds and the slopes λ_{ij} are equal in all groups
- scalar invariance: metric invariance holds and the intercepts τ_{ij} are the same in all groups

If metric invariance holds, the comparison across groups of the unstandardised relationships between variables is allowed. If scalar invariance holds, the comparison across groups of the means of the CPs is allowed. If scalar invariance holds and the means of the CPs are equal, then the means of the CS (average score based on several observed variables) are equal across groups.

The analyses are done in LISREL (Jöreskog and Sörbom, 1991) using the Maximum Likelihood estimator for multi-group analyses[25]. Pearson's correlations are used to compute the matrices to be analysed[26]. One tricky but crucial step is to assess the fit of the model. There are two main ways of looking at a model's fit.

[25] Lisrel input available online: http://bit.ly/e2wwpT

[26] Bollen and Barb (1981) show that "when as few as five categories are used to approximate the continuous variables, the correlation coefficients and their

First, one can consider the global fit of the model, using for instance the chi-square test. However, this test has important limits: it is dependent on the sample size, on the size of the parameters, and on the power, it is sensitive to deviations from normality, etc. That is why a huge range of fit indices have been developed lately (RMSEA, CFI, SRMR), but they have limits too. Saris, Satorra and Van der Veld (2009) argue that there is no proper way to test a model as a whole and that it is necessary to make the test at the parameter level.

Following them, the second option is to consider the local fit of a model. This approach is more adequate for our purpose: that is to test the equality of given parameters of the model (the loadings, the intercepts) and not only of the model as a whole. Besides, the procedure developed by Saris, Satorra and Van der Veld (2009) also takes into account the power. Therefore, by using JRule software (Van der Veld, Saris, Satorra, 2009) based on their procedure, we are able to test for specific equalities in our model and take not only type I but also type II errors into account. This software considers the modification indices, the power and when necessary the expected parameter changes in order to determine if, and where, there are misspecifications in the model. It suggests how the model can be corrected to improve its fit.

We should notice however that what is considered as a misspecification depends on what the researcher wants to detect: if he/she wants to detect a deviation of *x*, JRule tells him/her where there are deviations higher than *x*. This is what is referred to as misspecifications. We used the following values to define a misspecification: 0.10 for loadings, 0.10 for causal effects and correlations, 0.03 times the scale range for the intercepts and mean structure[27].

Measurement invariance informs us about the possibility of comparing unstandardized relationships and means across groups.

---

standard deviations for the collapsed and continuous variables are very close" (p. 232).

[27] The default values proposed by JRule are based on what is often used in practice. For instance, in the literature, it is often seen that if a loading is lower than 0.40, it is ignored by the researchers. However, we thought that the default values were too soft, so we changed them to have a stricter test.

Even if scalar invariance holds, however, the standardised estimates can be different across surveys if the variances vary. We therefore consider in the next sections the quality of the CSs, an indicator based on standardised parameters, and the correlations between the two CPs of each topic via an analysis of the external validity.

4.3.2. Computing the quality of the composite scores

Two different kinds of CS are generated using Stata version 10 (StataCorp, 2007). First, we generate what we call the "basic" CSs, which are unweighted averages of the different questions that are part of them (w_i = 1 / number of indicators). We are interested in the unweighted approach because it is the most widely used by researchers. However, more elaborated weights can be used as well. Therefore, we also generate CSs using regression weights: these weights minimize the sum of squared differences in scores between the CP and the CS (Saris and Gallhofer, 2007, p. 283). They should be computed on the pooled data (putting together the different groups), otherwise, differences can be found that come from the difference in weights.

We use LISREL to generate these CSs. For three out of the four concepts that we are studying, we estimate a simple factor model with one latent variable and three observed indicators in order to get the regression weights. For social trust, we only have two indicators. The model, therefore, is not identified. In order to get some weights, we estimate a factor model including social trust together with political trust. Since they are correlated, the model is identified and we can get regression weights. The problem is that the weights for one concept can be affected by the indicators of the second concept, because the concepts correlate, so the weights obtained may not be optimal for each concept separately. However, it may still be a better procedure than taking equal weights for the different indicators[28].

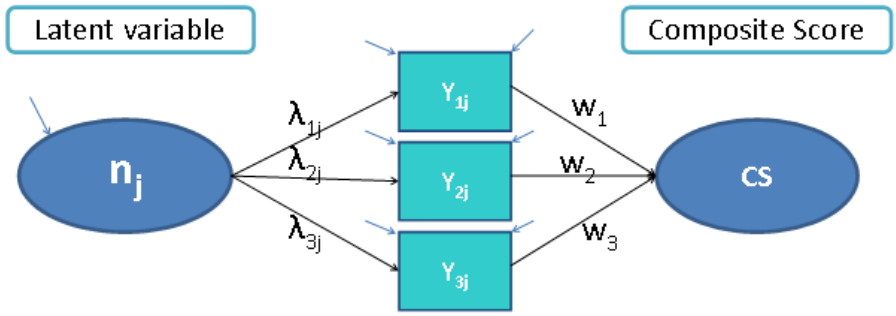
The quality of the CSs can be defined (Lawley and Maxwell, 1971; Saris and Gallhofer, 2007) in the same way as the quality of single

[28] A linear transformation is applied to all weights obtained in LISREL in order to get weights whose sum equals to one. These weights are used to compute the CSs.

items: it is the strength of the relationship between the CS and the latent variable of interest (CP).

The model in Figure 4.1 can be extended to include the CS, as shown in Figure 4.2. The intercepts and error terms have not been explicitly specified, but the small arrows represent them.

Figure 4.2: extension of the model to the CS



Note: η_j is the latent variable; Y_{ij} are the observed variables; λ_{ij} the loadings; w_i is the i^{th} weight; the arrows represent the intercepts and error terms.

The quality, or strength of the relationship between the latent variable (η_j) and the CS, can be computed as the correlation squared between the latent variable of interest and the CS. For the exact formula and details on the procedure, we refer to Saris and Gallhofer (2007, p.284).

Discussing the significance of the differences in quality of the CSs across the two surveys is a bit tricky since a formal test would require computing the standard errors of the quality estimate, which is quite complex. Instead, we focus on the relevance of the difference. We consider that a difference in quality of the CSs across surveys is relevant if it changes significantly (0.10 or more, criterion used in JRule) the observed correlations we get when the true correlation is the same.

Both the unweighted CSs and the CSs based on regression weights are considered. For three out of four concepts, the values used for the loadings (λ_{ij}) are the ones obtained in LISREL by running a simple factor model for one concept with three indicators separately for each survey. For social trust, as the regression weights are taken from the combined analysis of this concept together with political trust, the loadings are also taken from such a combined analysis, but estimated separately for the ESS and the LISS.

4.3.3. External validity

Different types of validity can be distinguished. We focus here on what is called "criterion-related validity" or "external validity". In Alwin's (2007, p.23) terms: "Criterion-related validity refers to the predictive utility of a measure or set of measures - do they predict or correlate with other theoretically relevant factors or criteria? For example, the criterion-related validity of SAT scores are typically assessed in terms of their ability to predict college grades (Crouse and Trusheim, 1988)". Or a few pages later: "Criterion validity is simply defined as the correlation of the measure Y with some other variable, presumably a criterion linked to the purpose of measurement" (Alwin, 2007, p.47).

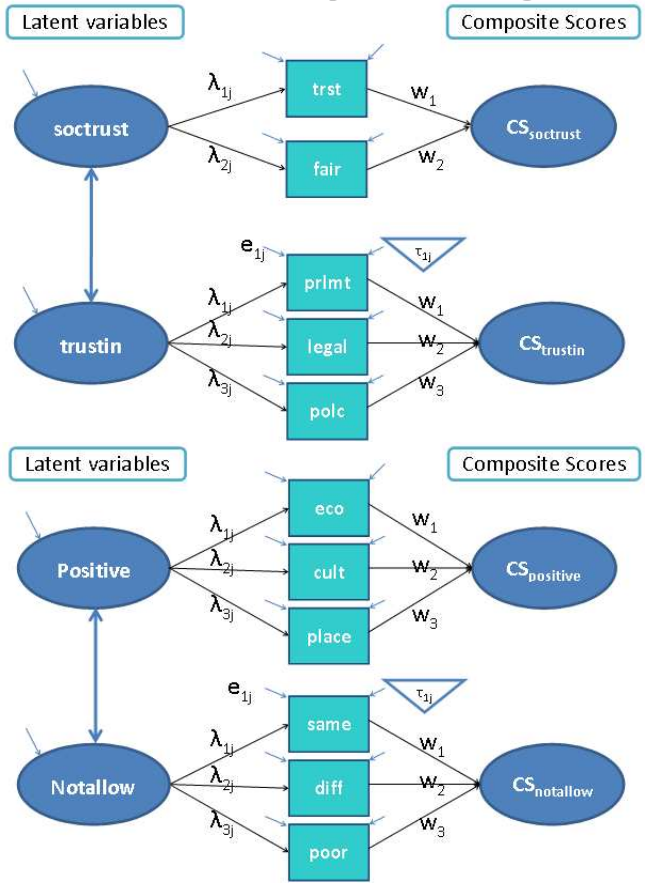
In our case, the criterion validity is quantified by looking at the correlation between the two concepts of one of the topics. The more similar this correlation is to the expected value, the better the external validity.

In order to test if these correlations are significantly different in one mode than in the other, we add in LISREL a constraint of equality on these specific parameters. Then we look in JRule whether the program indicates a misspecification for these parameters when they are constrained to be the same. If not, we conclude that the parameters are not significantly different across modes, and therefore that the external validity is similar in the face-to-face and web surveys. On the contrary, if JRule indicates a misspecification, we conclude that the parameters cannot be constrained to be the same without damaging the fit of the model and therefore, that external validity varies significantly in function of the mode used.

4.3.4. Application

The model presented in Figure 4.2 is applied to the two topics of interest, trust and attitude towards immigration, in the way described in Figure 4.3.

Figure 4.3: model for the trust example and the immigration example



In one given survey (ESS or LISS) and for one given topic (trust or immigration), the model is composed of two latent variables that each has several reflexive indicators. These same items are also used to create the CSs (both for the unweighted model and using regression weights). The CSs are called: "$CS_{soctrust}$", "$CS_{trustin}$", "$CS_{positive}$" and "$CS_{notallow}$" since they respectively intend to measure the CPs of social trust, trust in politics, evaluation of the

consequences of immigration and reluctance towards allowing more immigrants. The external validity is tested by looking at the correlation between the two latent variables for one given topic.

For trust, although empirical research does not always find a correlation significantly different from zero between these two CPs (Newton, 2007), and although it is necessary to make a distinction between them, theoretically it makes sense to argue that people that tend to trust other individuals also tend to have higher trust in politics such that some positive correlation should be found between them.

For immigration, it is expected that respondents thinking that immigration has negative consequences for the country will also be reluctant to allow more immigrants to come and live in the Netherlands, whereas respondents thinking immigration has positive consequences will be favourable towards allowing more immigrants. The two concepts should therefore be negatively correlated.

4.4. Results

4.4.1. Measurement equivalence

First, testing for configural invariance, no misspecifications were detected by JRule, so for both topics, the same model holds in the face-to-face and the web surveys. We can notice that although several or all items are measured with the same method, the testing of the model does not suggest that we need to introduce a method factor, since no misspecifications are detected, in particular, no correlation between the error terms are suggested.

Since configural invariance holds, we went on with testing the second level of invariance, which is metric invariance (equal slopes). For trust as well as for attitudes towards immigration, JRule does not indicate any misspecification for the parameters constrained to be equal across surveys, i.e. the loadings. In addition, the power is 0.99 in most cases, which means that there is a 99% chance that the test will detect the misspecification if the true difference for one parameter is bigger than the minimal difference we want to detect. Therefore, metric equivalence cannot be rejected,

and unstandardized relationships between variables can be compared across the ESS and the LISS surveys for the topics of trust and attitudes toward immigrants.

Finally, scalar invariance is tested by adding equality constraints on the intercepts. Again, JRule does not indicate any misspecification for the parameters of interest (loadings and intercepts now) although there is high power (0.99).

Since scalar invariance holds, it is possible to compare the means of the CPs. So far, we allowed them to be free in each survey. This led to very similar but not equal means of the latent variables in both surveys. In order to see if the differences are statistically significant, we add the constraint in LISREL that they should be invariant across surveys. Using JRule again, even if the power is very high, we cannot reject this hypothesis, meaning that the means of the CPs studied are equal in the two surveys. Their values are given in Table 4.2 (model for scalar invariance with additional constraint on the means of the CPs).

Table 4.2: Means of the CPs in both surveys

	Immigration				Trust			
	Not allow		Positive		Soctrust		Trustin	
	ESS4	LISS	ESS4	LISS	ESS4	LISS	ESS4	LISS
Mean LV	2.31		5.31		5.87		5.90	

Note: LV stands for "Latent Variable"

We start with attitudes toward immigration. Table 4.2 shows that whatever the mode of data collection used, the CP's means is 2.31 for "not allow" (measured on a 4-point scale) and 5.31 for "positive" (11 point scale): on average the Dutch population thinks that some or a few more people should be allowed to come and live in the country and that immigration is positive for the country across different domains. However, on average the Dutch population is almost neutral (close to the middle of the scale).

For trust, the means of the latent variables are 5.87 for "soctrust" and 5.90 for "trustin" (both measured on an 11-point scale): so the means social and political trust in the Netherlands are in the positive half of the scale, but again close to the midpoint.

The means of the CSs can differ from the means of the CPs, but since scalar invariance holds and the means of the CPs are equal in the LISS and the ESS, the means of the CSs should be similar across the two surveys. So even if the surveys use different modes of data collection, the means of both the CPs and the CSs can be compared across the ESS and LISS[29].

4.4.2. Quality of CS

Table 4.3 gives the quality of both the unweighted and regression-weights based CSs. The regression weights obtained for the different CSs are the following:

- for "not allow": 0.21, 0.70 and 0.09 respectively for traits 1, 2, 3
- for "positive": 0.31, 0.40 and 0.29 respectively for traits 1, 2 and 3
- for "soctrust": 0.54 and 0.46 respectively for traits 1 and 2
- for "trustin": 0.21, 0.61 and 0.18 respectively for traits 1, 2 and 3

Table 4.3: Quality of the Composite Scores

	Immigration				Trust			
	Not allow		Positive		Soctrust		Trustin	
	ESS4	LISS	ESS4	LISS	ESS4	LISS	ESS4	LISS
$q^2_{basic\ CS}$	.90	.90	.77	.82	.75	.78	.79	.88
$q^2_{regweights\ CS}$	.94	.95	.77	.83	.75	.78	.83	.91
$\lvert q^2_{basic} - q^2_{regweights} \rvert$	.04	.05	.00	.01	.00	.00	.04	.03
$\lvert q^2_{basicESS} - q^2_{basicLISS} \rvert$	.00		.05		.03		.09	
$\lvert q^2_{regweightsESS} - q^2_{regweightsLISS} \rvert$	.01		.06		.03		.08	

[29] If each observed variable Y_{ij} is a linear function of the CP, the mean of the unweighted CS is: $E(CS)=1/3\ E(Y_{11}+Y_{21}+Y_{31})= 1/3\ (\lambda_{11} + \lambda_{21} + \lambda_{31})E(\eta_1)+ E(\tau_{11} + \tau_{21} + \tau_{31})+E(e_{11} + e_{21} + e_{31})$. If we assume that the mean of the error terms is 0 and if scalar invariance holds, then a difference in the means of the CS can only comes from a difference in the means of the CPs. If the means of the CPs are equal across groups, then the CSs should also be equal across groups. Still the means of the CSs may vary from the means of the CPs, for instance if the sum of the loadings is different from 3.

First, if we compare the quality estimates for the basic and the more elaborated CSs, it seems not to matter much. There is only a difference for "not allow", but it is quite small (0.04) and it is the same in both surveys.

For "not allow" the quality is quite high and very similar in both surveys (around 0.90 for the basic CSs and 0.95 for the ones based on regression weights). For the other concepts, the quality is not so high but it is still higher than 0.75. Besides, the differences are larger but they stay small. In order to determine the relevance of these differences, we use the estimates of quality found in the different surveys to examine what differences in observed correlations appear, given a true correlation, due to a variation in quality across surveys. We focus on the topic of trust and on the basic CS since it is there that the greatest differences between the face-to-face and the web surveys are found (0.09 for "trustin" and 0.03 for "soctrust"). The observed correlation between the CS for social and political trust can be expressed as the product of the true latent correlation times the quality coefficient for the CS of social trust times the quality coefficient for the CS of political trust, that is:

$$r(CS_{soctrust}, CS_{trustin}) = \rho \times q_{CS_{soctrust}} \times q_{CS_{trustin}} \qquad (2)$$

where ρ is the true correlation between the CPs of social trust and political trust. So the difference between observed correlations across surveys is:

$$r(CS_{soctrust}, CS_{trustin})^{LISS} - r(CS_{soctrust}, CS_{trustin})^{ESS}$$
$$= \rho(q^{LISS}_{CS_{soctrust}} \; q^{LISS}_{CS_{trustin}} - q^{ESS}_{CS_{soctrust}} \; q^{ESS}_{CS_{trustin}}) \qquad (3)$$

The difference is a linear function of the true correlation: the higher the true correlation, the higher the difference in observed correlations. So in order to see the maximum difference we take a correlation of one. Then, the highest difference for trust is still lower than the value we set as criterion for misspecification. Indeed, we have:

$$r(CS_{soctrust}, CS_{trustin})^{LISS} - r(CS_{soctrust}, CS_{trustin})^{ESS} = 1 \times (.88 \times .95 - .86 \times .89)$$
$$= .0706 < .10$$

In the next section, we will see that the true correlation between the concepts by postulation of social and political trust is around 0.50 (cf. Table 4.4). If this is so, it means, knowing the quality of the two CSs, that we expect an artificial difference of 0.0706/2 = 0.0353 in the observed correlations of the face-to-face and the web surveys. This is small enough to not worry about. For the other cases (other topic and/or CSs based on regression weights), the differences are even lower.

All in all, the similarities dominate: the quality of the different CSs is close enough in the face-to-face and web surveys for the different concepts to not disturb the cross-survey analyses of standardised relationships. It seems also that basic CSs can perform almost as well as more elaborated CSs.

4.4.3. External validity

The last result we want to stress concerns the external validity of the four concepts analysed. As argued previously, for the two concepts about immigration, "positive" and "not allow", we assume a negative and quite strong correlation. On the contrary, for the topic of trust, a positive correlation is expected between social and political trust. This correlation should not be too close to 1 (otherwise it would mean that social and political trust are the same concept, which is not what the literature shows). It may even be relatively low, but still it should be significant and positive.

To test for external validity, we run the models again but constraining the parameter of the covariance between the two CPs to be the same in the face-to-face and the web survey. This does not lead to misspecification according to JRule, which was expected since we found metric invariance for both CPs. However, our interest here is to consider the standardised relationships and not the unstandardised ones. Therefore, Table 4.4 presents the correlations between CPs for the two topics.

Table 4.4: testing for external validity

	Immigration		Trust	
	ESS4	LISS	ESS4	LISS
Corr(CP_1, CP_2)	-.64	-.64	.47	.52
External validity	ok	ok	ok	ok

One can notice that differences in the correlations are found: for the topic of trust, although the unstandardised parameters are equal, the standardised ones vary. Nevertheless, the difference found across surveys is lower than the criterion used to specify a misspecification ($0.05<0.10$), so we conclude that even if one survey is using face-to-face interviews whereas the other use the web, this does not impact significantly the correlation between the two CPs considered in each of the two topics.

Moreover, the correlations are in line with what we expected for both topics. Indeed, the correlation between "positive" and "not allow" is -0.64 in the ESS and in the LISS: so it is a quite strong negative correlation. On the contrary, the correlation between social and political trust is positive: 0.47 in the ESS and 0.52 in the LISS. This is quite high compared to some past results (Newton, 2007) but in line with others (Saris and Gallhofer, 2007), and this is still much lower than 1 and seems quite probable. Overall, the analyses suggest that the external validity is similar in the ESS and in the LISS surveys, when face-to-face is used and when web is used.

4.5. Conclusion

Measurement equivalence needs to be assessed in order to be able to make meaningful comparisons across groups. In this paper we were interested in groups of respondents that are completing surveys using different modes of data collection. Modes have a set of properties (interviewer or self-completion, visual or oral stimuli) that may influence the way people express themselves when answering a survey. Thus, there is a risk that the mode of data collection could threaten the measurement equivalence of the questions.

Comparing a face-to-face survey (ESS) and a web survey (LISS) for four different concepts related to the two topics of trust and attitude

towards immigration, we found that configural, metric and scalar invariances all hold across the two surveys and for all four concepts. Since metric invariance holds, one can compare the unstandardized relationships of the concepts with each other across modes. That scalar invariance holds too suggests that one can also compare the means of these four concepts across different modes of data collection.

But scalar invariance only tells us about unstandardised relationships. Standardised relationships may still vary. The quality estimates of the CSs, since they are computed using standardised estimates, may therefore vary across surveys even if scalar invariance has been assessed. However, our results show that the quality estimates are comparable across surveys. Therefore, we can compare standardized measures across the ESS and the LISS for the concepts tested. We also find that using a basic CS or one based on regression weights does not really make a difference.

We looked at one particular standardised measure to illustrate this concept, the correlation between the two CPs of interest. This allows us to check for external validity simultaneously. The correlations are equivalent across surveys and the external validity seems to hold too since the correlations found between the two CPs within each topic go in the expected direction and are relatively large.

The analysis however focuses on only four concepts about two topics, considers only two modes, and is based on data from only one country, the Netherlands. Therefore, much more evidence would be needed before being able to generalise our conclusions. Still, overall, the results are quite encouraging, since they show that even using different modes of data collection, as long as the exact same wording and scales are used and the samples are drawn randomly from the population, equivalent measurements can be obtained and CSs of similar and quite high quality can be constructed using the data. The use of show cards in the face-to-face survey is probably also an important element explaining the similarity across the two surveys studied. Further, our results are in line with previous research about measurement equivalence across different modes (King and Miles, 1995; Cole, Bedeian, Field, 2006; De Beuckelaer and Lievens, 2009).

Chapter 5

Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach [30]

Abstract

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.

[30] Published:

Revilla, M. (2010). "Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach". *Survey Research Methods*, 4(3):151-164. ISSN 1864-3361
Can be retrieved from: http://www.surveymethods.org

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

5.1. 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, don't 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.

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, p. 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, p. 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, p. 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, p. 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.

5.2. The European Social Survey (ESS)

5.2.1. ESS round 4

The ESS is a biannual cross-national project designed to measure changing social attitudes and values in Europe[31]. 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

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

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 before 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.

5.2.2. 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-mode, 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 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)[32] 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 ½ hour).

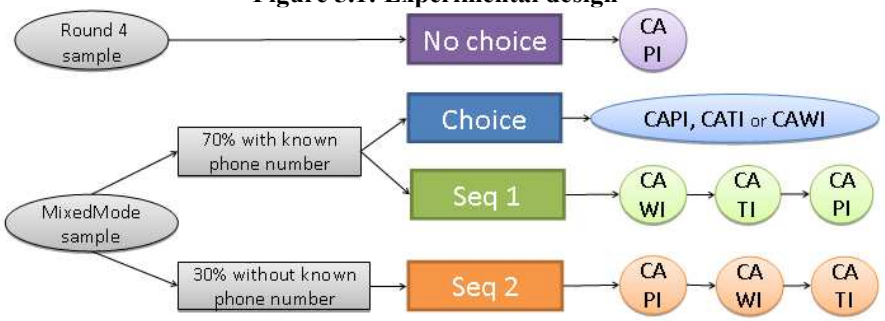
As Figure 5.1 shows, the general design of the experiment is however more complex, since a separation is done between people with and without known phone number. This is because of the nature of the sampling frame and mode of contact used. The sampling frame consisted of postal addresses, but the contact 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

[32] See for instance Eurobarometer 71.2 2009

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.

Figure 5.1: Experimental design



5.2.3. 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 5.1 summarizes the experiments analysed.

Table 5.1: Questions and methods analyzed

Expt	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)	11 points (very)
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

***Note**: "extreme"= extreme used in the labels of the end points, "very" = very used in the labels of the end points*

5.3. 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).

5.3.1. 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 5.2 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.

Table 5.2: 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%)

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 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 5.3 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).

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

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 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.

5.3.2. 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 modes 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[33], 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 5.4 gives this information both when dividing respondents by modes and by designs.

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

Table 5.4: 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 tel	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

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, but some people may have used a mobile phone to answer the interview. Besides, the ESS question used to obtain the numbers in Table 5.4 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 5.4 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 5.5: 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 5.5 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 5.3 are probably influenced by the variations in telephone and Internet coverage of the different age groups.

5.3.3. 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 5.6 gives the coefficients of this regression: basic coefficients and coefficients expressed on the odd ratios scale.

Table 5.6: 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 *	

Table 5.6 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, different 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.

5.4. Estimation of the quality

5.4.1. 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 ½ 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[34]. 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

[34] 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.

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).

5.4.2. *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.

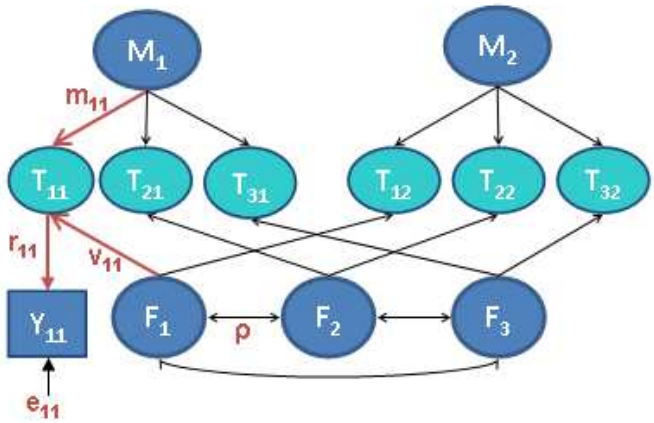
Each experiment is studied separately. Figure 5.2 shows the model used for six variables. It contains three correlated traits (F_1, F_2 and F_3), each measured with two methods (M_1 and M_2). It is assumed that the methods are not correlated with each other, nor with the traits, and that the effects 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 5.2 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):

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \qquad \text{for all } i,j \qquad (1)$$
$$T_{ij} = v_{ij} F_i + m_{ij} M_j \qquad \text{for all } i,j \qquad (2)$$

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.

Figure 5.2: The MTMM model for 3 traits and their repetitions



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[35]. 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

[35] An example of Lisrel input is available online: http://docs.google.com/Doc?docid=0AbQWMcvxTKZGQ3Mm10MzRfMTY1ZGN0YjZtY3Q&hl=en

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[36].

5.5. Main findings

5.5.1. Comparison of the quality estimates by designs

Table 5.7 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.

Table 5.7: Estimates in the different designs

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 pts	1	.79	.91	.98	.97	.98	.96	.59	.80	**.78**
		h-min	.68	1	.62	.96	.98	.95	.43	.96	.35	**.58**
	concurrent +sequential	8 pts	1	.82	1	.96	.94	.96	.92	.59	.92	**.81**
		h-min	.73	1	.71	.91	.96	.90	.44	.92	.41	**.59**
Satisfaction	ESS 4	11 extr	.83	.96	.91	.93	.95	1	.60	.83	.83	**.75**
		11 very	.92	.94	.91	.90	.90	.87	.69	.72	.63	**.68**
	concurrent	11 extr	.83	.96	.91	.94	.96	1	.61	.85	.83	**.76**
		11 very	.93	.94	.91	.86	.86	.83	.64	.65	.57	**.62**
	sequential	11 extr	.82	.95	.91	.97	.98	1	.63	.87	.83	**.78**
		11 very	.92	.94	.91	.89	.89	.86	.67	.70	.61	**.66**
Social trust	ESS 4 + concurrent + sequential	11 pts	.86	.83	*na*	1	1	*na*	.74	.69	*na*	**.71**
		6 pts	.91	.84	*na*	.89	.87	*na*	.66	.53	*na*	**.60**
Political trust	ESS 4	11 pts	.84	.91	1	.98	.98	.99	.68	.80	.98	**.82**
		6 pts	.90	.92	.84	.88	.88	.85	.63	.66	.51	**.60**
	concurrent + sequential	11 pts	.87	.96	1	.92	.89	.94	.64	.73	.88	**.75**
		6 pts	.88	.92	.86	.83	.84	.80	.53	.60	.47	**.53**

***Note**: "h-min"=time asked in hours and minutes, "extr"= 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), "pts"=number of points*

[36] 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

Table 5.8: 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

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 5.8 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 separates methods one and two in the media as well as in the political trust experiment. This is much higher than the differences across 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.

5.5.2. 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 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 5.9 provides the same estimates as Table 5.7 but focusing on the mixed-mode data and differentiating the groups of people answering by CAPI, CATI and CAWI. Table 5.10 gives the differences in mean quality between modes.

Table 5.9: Estimates in the different modes

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 pts	1	.79	1	.96	.94	.96	.92	.55	.92	**.80**
		h-min	.72	1	.77	.89	.94	.90	.41	.88	.48	**.59**
	CATI	8 pts	1	.80	1	.96	.94	.96	.92	.57	.92	**.80**
		h-min	.75	1	.70	.89	.94	.90	.45	.88	.40	**.58**
	CAWI	8 pts	1	.88	1	.93	.90	.93	.86	.63	.86	**.79**
		h-min	.74	1	.73	.89	.94	.90	.43	.88	.43	**.58**
Satisfaction	CAPI	11 extr	.85	.96	.93	.96	.97	.97	.67	.87	.81	**.78**
		11 very	.92	.94	.89	.85	.85	.83	.61	.64	.55	**.60**
	CATI	11 extr	.83	.95	.84	.96	.97	.97	.63	.85	.66	**.72**
		11 very	.80	.90	.82	.92	.92	.91	.54	.69	.56	**.59**
	CAWI	11 extr	.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 pts	.89	.86	*na*	1	1	*na*	.79	.74	*na*	**.77**
		6 pts	.91	.82	*na*	.91	.89	*na*	.69	.53	*na*	**.61**
	CATI	11 pts	.81	.87	*na*	.96	.96	*na*	.60	.70	*na*	**.65**
		6 pts	.92	.84	*na*	.85	.82	*na*	.61	.47	*na*	**.54**
Political trust	CAPI	11 pts	.84	.94	.96	.93	.86	.94	.61	.65	.81	**.69**
		6 pts	.90	.90	.83	.92	.93	.91	.69	.70	.57	**.65**
	CATI	11 pts	.95	.91	.95	.77	.83	.94	.54	.57	.80	**.63**
		6 pts	.81	.87	.87	.89	.89	.87	.52	.60	.57	**.56**
	CAWI	11 pts	.87	.98	1	.93	.86	.94	.65	.71	.88	**.75**
		6 pts	.88	.92	.82	.97	.97	.96	.73	.80	.62	**.71**

***Note**: "h-min"=time asked in hours and minutes, "extr"= 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), "pts"=number of points*

Table 5.10: 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

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 separate 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.

5.6. 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 interviewer. 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 5.2, 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 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 5.3 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, 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.

General Conclusion

Every researcher conducting a survey has to deal, in between many other choices and decisions, with the issue of choosing a mode of data collection. In the past, the choice was limited to relatively few available modes (mainly face-to-face, mail, telephone). Nowadays, with the development of new modes of data collection (e.g. computer-assisted data collection modes, web), the decision becomes more complex.

If a "mode effect" exists, i.e. if the answers of the respondents vary depending on the mode in which they are answering, then it is crucial to know it and take it into account; otherwise wrong conclusions may be drawn when comparing data coming from different modes. As Merton already stated in 1959 (p. *xiii*): "Before social facts can be 'explained', it is advisable to ensure that they actually are facts [...] explanations are sometimes provided for things that never were". To ensure that facts exist, it is necessary to check that observed differences (respectively similarities) are not resulting from differences in errors, linked for instance to the use of several modes, but are "true" differences (respectively similarities).

Therefore this dissertation dealt with the impact of the mode of data collection on the quality of answers to survey questions. The focus on the quality is motivated by the gap in previous research on this particular aspect and the importance of filling it in.

Overall, our results shed light on a point that was almost completely missing from the literature but is really important, by giving a first idea of what can be expected when different modes are used with respect to the quality of the answers to survey question defined as the product of reliability and validity. By contrast to the common ideas about the low quality of web surveys, it shows that the use of a web survey instead of a face-to-face one does not systematically impact the quality: if the web survey is done in a similar way as the LISS study, then the quality even tends to be a bit better than the one of a face-to-face, even if the differences are usually not significant.

In brief, the main results of the dissertation suggest that means, unstandardised (chapter 4) and standardised relationships (chapters 2 and 5) can be compared across the face-to-face and web surveys studied, but also across groups of respondents with specific profiles (chapter 3). However, the telephone yields more differences so the introduction of this mode may be more problematic (chapter 5). In general, the method used (number of points of the scale, labels, middle point present, etc) seems to impact much more the quality than the mode of data collection.

The dissertation also has practical implications for the European Social Survey. The size and scope of this survey makes it one of the most important in the world, and as such, it deserves a special attention. As mentioned in the introduction, the ESS team is interested in switching from the traditional face-to-face only data collection to a new data collection approach allowing the introduction of other modes. In that context, the results of the dissertation have been discussed with the ESS team and have been incorporated in the ESS report (Eva et al., 2010) on which the decisions for future allowance of new modes will be based, together of course with other kinds of analyses about sample composition, variables distributions, cost per question, response rates (see for instance Lynn, Revilla, Vannieuwenhuyze, forthcoming), etc.

Our overall impression at the end of this dissertation is that for a high quality survey that aims to draw conclusions for the general population of a country, an alternative unimode design than the face-to-face is not viable or cannot nowadays lead to the same level of quality in any country, except if it takes the form of a web panel similar to the LISS. For a few years now a trend is appearing in Europe toward the creation of web panels. In the light of the success of the LISS, a few countries started thinking about or even launching web panels in a very similar way. Germany (Gathmann and Blom, 2011) and France (Lesnard, 2011) are in the process of creating such panels. They could be the future for collecting survey data.

Another option could be found in a mixed-mode approach. In that case, we would recommend not to allow the use of telephone, because of the differences in quality we found, but we tend to think

that a combination of face-to-face and web could help getting more or as representative samples at a reduced cost and without damaging the quality and comparability of the results as long as the form of the questions is kept as similar as possible in the face-to-face and web questionnaires.

However, the analyses are based only on a few datasets that were available in one given country. So we should be careful about drawing general conclusions from our results alone. Nevertheless, we believe that the results are not only specific to this particular case but hold for surveys with similar characteristics to the ones studied and in countries relatively similar to the one studied: in particular, the face-to-face survey should use show cards and both surveys should be based on a probabilistic sample. For surveys that do not share these characteristics, more research is needed. Concerning the country, the Internet coverage should be quite high and the general attitude of the population towards surveys should present similarities. For very different countries (e.g. Asiatic or African countries), we believe more research is needed too.

Also, selection and measurement effects could not be separated in our analyses because of the design of the surveys (we did not have the same respondents answering in different modes). But a recent contribution by Vannieuwenhuyze, Loosveldt and Molenberghs (2010) worked out a way of disentangling selection and measurement effects in a design similar to the one we have using an instrumental variable approach. Combining this approach with the MTMM analyses will allow getting more precise results than the ones of the thesis and is a project for future research (Vannieuwenhuyze and Revilla, forthcoming).

Finally, more research would be needed for complex or sensitive questions since the dissertation only focused on "normal" questions. Nevertheless, it could be argued that for complex questions, the mode of data collection plays a more important role and that it is especially for these kinds of questions that differences across modes can be expected: the more complex the question, the more difficult it can be to answer it in a oral mode, creating more random errors in these modes; the more complex the question, the more difficult it can be to answer it without the help of an interviewer, creating more random errors in self-completed modes. For sensitive questions too,

biggest differences across modes could be expected because of social desirability: respondents may feel more uncomfortable to admit behaviours that are condemned by law or by society in presence of an interviewer. It would therefore be interesting to test these hypotheses.

Bibliography

Alwin, D.F. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H.L. Costner (ed.), *Sociological Methodology,* 79-105. San Francisco: Jossey-Bass.

Alwin, D.F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken NJ: Wiley.

Alwin D.F., and J.A. Krosnick (1991). The Reliability of Survey Attitude Measurement. *Sociological Methods and Research* 20(1):139-181. doi: 10.1177/0049124191020001005

Anderson, C.A., Lepper, M.R. and L. Ross (1980). Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information. *Journal of Personality and Social Psychology*, 39(1-6): 1037- 49.

Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 46:409-42. doi: 10.1086/268840 Reprinted in W.E. Saris and A. van Meurs (1990). *Evaluation of measurement instruments by metaanalysis of multitrait multimethod studies*. Amsterdam: North-Holland

Bollen, K.A., and K.H. Barb (1981). Pearson's r and Coarsely Categorized Measures. *American Sociological Review* 46(2):232-239. http://www.jstor.org/stable/2094981

Brambilla, D.J., and S.M. McKinlay (1987). A Comparison of Responses to Mailed Questionnaires and Telephone Interviews in a Mixed-mode Health Survey. *American Journal of Epidemiology*, 126:962-971. http://aje.oxfordjournals.org/content/126/5/962.short

Browne, M.W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and*

Statistical Psychology, 37:1-21. doi: 10.1111/j.2044-8317.1984.tb00785.x

Campbell, D.T. and D.W. Fiske (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 6:81-105. doi: 10.1037/h0046016

Coenders, M. (2001). *Nationalistic Attitudes and Ethnic Exclusionism in a Comparative Perspective: An Empirical Study of Attitudes Toward the Country and Ethnic Immigrants in 22 countries*. Publisher: Radboud University Nijmegen

Coenders, G., and W.E. Saris (1995). Categorization and measurement quality. The choice between Pearson and Polychoric correlations. In W.E. Saris, *The MTMM approach to evaluate measurement instruments* (pp. 125-144).

Cole, M.S., Bedeian, A.G., and H.S. Field (2006). The Measurement Equivalence of web-Based and Paper-and-Pencil Measures of Transformational Leadership: A Multinational Test. *Organizational Research Methods* 9:339-368. doi: 10.1177/1094428106287434

Corten, I.W., Saris, W.E., Coenders, G., van der Veld, W.M, Aalberts, C.E., and C. Kornelis (2002). Fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, 9(2):213-232. doi:10.1207/S15328007SEM0902_4

Couper, M.P, and P.V. Miller (2008). Introduction to the special issue. *Public Opinion Quarterly*, 72(5): 831-835 doi: 10.1093/poq/nfn066

Crouse, J. and D. Trusheim (1988). *The case against the SAT*. Chicago, IL. University of Chicago Press.

De Beuckelaer, A., and F. Lievens (2009). Measurement Equivalence of Paper-and-Pencil and Internet Organisational Surveys: A Large Scale Examination in 16 Countries. *Applied Psychology: an international review* 58(2):336-361. doi: 10.1111/j.1464-0597.2008.00350.x

De Heer, W., de Leeuw, E.D, and J. van der Zouwen (1999). Methodological Issues in Survey Research: a Historical Review. *BMS: Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 64:25-48. doi: 10.1177/075910639906400104

De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*. 21(2):233-255. http://igitur-archive.library.uu.nl/fss/2011-0314-200305/EdL-to%20mix%202005.pdf

De Leeuw ED, and J. van der Zouwen (1988). Data quality in telephone and face-to-face surveys: a comparative meta-analysis. In: Groves RM, Biemer PP, Lyberg LE et al. eds. *Telephone survey methodology*. New York: John Wiley and Sons, 1988.

Dillman, D.A., and L.M. Christian (2005). Survey Mode as a Source of Instability in Responses Across Surveys. *Field Methods* 17(1):30-52. doi: 10.1177/1525822X04269550

Dillman, D.A., Clark, J.R., and K.K. West (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., and B.L. Messer (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. doi: 10.1016/j.ssresearch.2008.03.007

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

Duckworth A.L, and M.E.P Seligman (2006). Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. *Journal Educ. Psychol*. 98(1):198–208

Dumka, L. E., Stoerzinger, H. D., Jackson, K. M., and M.W. Roosa (1996). Examination of the cross-cultural and cross-language

equivalence of the parenting self-agency measure. *Family Relations*, 45: 216-222

Eva, G., Loosveldt, G., Lynn, P., Martin, P., Revilla, M., Saris, W., and J. Vannieuwenhuyze (2010). *ESSPrep WP6 – Mixed Mode Experiment. Deliverable 21: Final Mode Report*. London: European Social Survey.

Faas, T., and H. Schoen (2006). Putting a questionnaire on the Web is not enough- A comparison of Online and Offline survey conducted in the context of the German Federal Elections 2002. *Journal of Official Statistics* 22:177-191. http://www.business.aau.dk/~csp/Gang_08/Putting_a_quest_on_web.pdf

Fricker, S., Galesic, M., Tourangeau, R., and T. Yan (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly* 69:370-92. doi: 10.1093/poq/nfi027

Gathmann, C., and A. Blom (2011). Concept and Structure of the German Internet Panel. Presented at the MESS workshop, 12-08-2011, at Oisterwijk (The Netherlands)

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. doi: 10.1093/ijpor/edn054

Heerwegh, D., and G. Loosveldt (2009). Face-to-Face Versus Web Surveying in a High-Internet-Coverage Population. Differences in Response Quality. *Public Opinion Quarterly* 72.5: 836-846 doi:10.1093/poq/nfn045

Holbrook, A.L., Green, M.C., and J.A. Krosnick (2003). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires. Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly* 67:79-125. doi: 10.1086/346010

Hox, J.J., and E.D. De Leeuw (1994). A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys:

Applying Multilevel Models to Meta-analysis. *Quality and Quantity* 28:329-44. doi: 10.1007/BF01097014

Jöreskog, K.G. (1970). A general method for the analysis of covariance structures. *Biometrika*, 57:239-51. doi: 10.1093/biomet/57.2.239

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36:109-133. doi: 10.1007/BF02291393

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

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

Kaplowitz, M.D., Hadlock, T.D., and R. Levine (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly* 68(1):94-101. doi: 10.1093/poq/nfh006

King, W.C, and E.W. Miles (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology* 80(6):643-651. doi: 10.1037/0021-9010.80.6.643

Knoef, M., and K. de Vos (2009). The representativeness of LISS, an online probability panel. CentERdata. http://www.centerdata.nl/export/sites/CentERdata/en/TopMenu/Projecten/MESS/Knoef_DeVos-2008-MESS_workshop-paper.pdf

Kreuter, F., Presser, S., and R. Tourangeau (2009). Social Desirability Bias in CATI, IVR and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72(5): 847-865. doi:10.1093/poq/nfn063

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology* 5:213-36. doi: 10.1002/acp.2350050305

Krosnick, J.A. (1999). Survey Research. *Annual Review of Psychology* 50:537-567. http://www.annualreviews.org/doi/pdf/10.1146/annurev.psych.50.1.537

Krosnick, J.A. and D.F. Alwin (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly* 51:201-219. doi: 10.1086/269029

Lawley D.N., and A.E. Maxwell (1971). *Factor Analysis as a Statistical Method*. London: Butterworth. http://www.jstor.org/stable/2986915

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer panel Web Surveys. *Journal of Official Statistics* 22(2): 329-349. http://www.jos.nu/Articles/article.asp

Lesnard, L. (2011). ELIPSS: A New Mobile Web Panel for Social Scientists. Presented at the MESS workshop, 12-08-2011, at Oisterwijk (The Netherlands)

Lozar Manfreda, K., Bosnjak, M., Haas, I., and V. Vehovar (2005). A meta-analysis of response rates in Web surveys compared to other survey modes. In ESF workshop on internet survey methodology. Dubrovnik, Croatia.

Lynn, P., Häder, S., Gabler, S., and S. Laaksonen (2007). Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience. *Journal of Official Statistics* 23(1):107-40. http://konference.fdvinfo.net/rc33/2004/Data/PDF/stream_08-19.pdf

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

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

Mayda, A.M. (2006). Who Is Against Immigration? A Cross-Country Investigation of Individual Attitudes toward Immigrants. *The review of Economics and Statistics* 88(3): 510-530. doi:10.1162/rest.88.3.510

Merton, R. K., L. Broom, and L. S. Cottrell. (1959). *Sociology Today: Problems and Prospects*. Basic Books.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58(4):525-543. doi: 10.1007/BF02294825

Newton, K. (2007). Social and Political Trust. In *The Oxford Handbook of Political Behavior* edited by Dalton, Russell J. and Klingemann, Hans-Dieter, Chapter 18: 342-359.

Northrop F.S.C. (1947). *The Logic of the Sciences and the Humanities*. New York: World Publishing Company.

Peterson, C., Semmel, A., von Baeyer, C., Abramson, L.Y., Metalsky, G.I, and M.E.P. Seligman (1982). The attributional style questionnaire. *Cognitive Therapy and Research*, 6: 287-300

Revilla, M. (2010). Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach. *Survey Research Methods* 4(3):151-164. http://w4.ub.uni-konstanz.de/srm/article/view/4278

Revilla, M., and W.E. Saris (2010). A comparison of surveys using different modes of data collection: ESS versus LISS panel. RECSM working paper 13. http://www.upf.edu/survey/_pdf/RECSM_wp013.pdf

Revilla, M., and W.E. Saris (2012). A Comparison of the Quality of Questions in a Face-to-face and a Web Survey. *International Journal of Public Opinion Research*. Advance Access published April 17, 2012. doi:10.1093/ijpor/eds007

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):68-73. doi:10.1136/jech.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. http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf

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. (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, July 10-12, 2008.

Saris, W.E. and F.M. Andrews (1991). Evaluation of measurement instruments using a structural modeling approach. In P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-597). New York: Wiley. doi: 10.1002/9781118150382.ch28

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

Saris, W.E., Revilla, M., Krosnick, J.A., and E.M. Shaeffer (2010). Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options. *Survey Research Methods,* 4(1):61-79. http://w4.ub.uni-konstanz.de/srm/article/view/2682

Saris, W.E., Satorra, A. and G. Coenders (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology* 34:311-347. doi: 10.1111/j.0081-1750.2004.00155.x

Saris, W.E, Satorra, A., and W.M. Van der Veld (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal,* 16(4):561-582. doi: 10.1080/10705510903203433

Scherpenzeel, A. (1995). Meta Analysis of a European Comparative Study. In Saris, W.E., Münnich, A. (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments* (pp. 225-242). Eötvös University Press, Budapest

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. Zeist. Available on the Web: http://www.centerdata.nl/export/sites/CentERdata/nl/TopMenu/Projecten/MESS/Scherpenzeel-2008-MESS_workshop-paper.pdf

Scherpenzeel, A., and W.E. Saris (1997). The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. *Sociological Methods and Research* 25(3):341-83. doi: 10.1177/0049124197025003004

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H.J., Turner, R., and S. Berry (2004). A comparison between a propensity weighted Web survey and an identical RDD Survey. *Social Science Computer Review* 21:1-11. doi: 10.1177/0894439303256551

Schulenberg, J. E., Shimizu, K., Vondracek, F.W., and M. Hostetler (1988). "Factorial invariance of career indecision dimensions across junior high and high school males and females". *Journal of Vocational Behavior*, 33: 63-81.

Shettle, C., and G. Mooney (1999). Monetary Incentives in US Government Surveys. *Journal of Official Statistics* 15(2): 231-250. http://www.jos.nu/Articles/abstract.asp?article=152231

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, 26:597-619.

Smyth, J.D., Christian, L.H. and D.A. Dillman (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. doi: 10.1093/poq/nfn005

Smyth, M.M., Morris, P.E., Levy, P., and A.W. Ellis (1987). *Cognition in Action*. London: Erlbaum.

StataCorp. (2007). *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.

Steenkamp, J.E.M., and H. Baumgartner (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research*, 25:78-90.

Tansy, M., and J.A. Miller (1997). The invariance of the self-concept construct across White and Hispanic student populations. *Journal of Psychoeducational Assessment*, *15:* 4-14.

Toepoel, V., Das, M., and A. Van Soest (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

Toepoel, V., Das, M., and A. van Soest (2008). Effects of Design in Web Surveys. Comparing Trained and Fresh respondents. *Public Opinion Quarterly* 72(5):985-1007. doi:10.1093/poq/nfn060

Tourangeau, R., and T.W. Smith (1996). Asking Sensitive Questions: the Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly* 60(2):275-304. doi: 10.1086/297751

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

Uslaner, E.M. (2002). The moral foundations of trust. NewYork: Cambridge University Press.

Van der Veld, W.M., Saris, W.E., and A. Satorra (2009). Jrule 2.0: User manual (Unpublished Manuscript, Internal Report). Radboud University Nijmegen, the Netherlands.

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

Vannieuwenhuyze J., Loosveldt G., and G. Molenberghs (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly* 74(5), 1027-1045

Voogt, R.J.J., and W.E. Saris (2005). Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects. *Journal of Official Statistics* 21:367-87. http://www.jos.nu/Articles/abstract.asp?article=213367

Werts, C.E., and R.L. Linn (1970). Path analysis: Psychological examples. *Psychological Bulletin* 74:194-212. doi: 10.1037/h0029778