UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 13

2010

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

A COMPARISON OF SURVEYS USING DIFFERENT MODES OF DATA COLLECTION:

EUROPEAN SOCIAL SURVEY VERSUS LISS PANEL

Melanie Revilla, Willem E. Saris [1],

[1] Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra, Edifici França, Passeig de Circumval·lació, 8, 08003 Barcelona. Spain. Corresponding author: send email to melanie.revilla@upf.edu

Abstract

Web surveys are becoming more and more popular in survey research, mainly because of their lower costs. With the increase of the Internet coverage in most European countries, the response rates are becoming high enough to collect huge amount of data in a short period of time. However, there is a risk that changing to this new mode would lead to data incomparable with data collected in the past. Therefore it is necessary to check if data collected using Web and data collected with more traditional modes (mainly mail, telephone, face-to-face) produce similar results. This paper compares one survey completed by the Longitudinal Internet Studies for the Social sciences (LISS) panel (Web panel based on probability sample) in December 2008 with the same questions asked in the frame of the European Social Survey (face-to-face) in the Netherlands. Focusing on the quality of single items and composite scores, we find few differences between these two surveys.

Keywords

Web and face-to-face surveys, quality of questions, MultiTrait-MultiMethod approach, measurement invariance

Acknowledgments

We are very grateful to the Measurement and Experimentation in the Social Sciences (MESS) project that accepted our questionnaire proposal and provided us the data needed.

Introduction

The data collection is a crucial phase in survey research since it determines for a large part the quality of the results. But it is also a tricky step, since many decisions have to be taken which may impact the final findings so that it is necessary to take them into account. The mode of data collection is one of these decisions that have to be considered.

So far, most surveys used face-to-face or telephone in order to collect data. But today, with the increasing difficulties of achieving surveys at reasonable costs and in a short period of time, more and more people argue that using different modes of data collection could offer an interesting alternative. With the development of new technologies, new opportunities appear. In particular, the idea to switch to Web surveys, which are usually cheaper, offer more flexibility and can reach a large population in a short time, is becoming very attractive.

However, different modes of data collection may lead to different coverage, sampling, nonresponse and measurement errors. Part of the population may not have access to some of the modes (mainly telephone or Web). Even if all have access to the different modes, some persons may feel uncomfortable to participate in certain modes but not in others: different people can therefore select themselves into participating depending on the modes of data collection proposed. Finally, even if the same persons agree to participate leading to perfectly identical samples in different modes, a mode effect, i.e. a difference in responses resulting from the fact that the question is asked in a different survey mode, can still appear. For instance, Krosnick (1991, 1999) shows that varying levels of social desirability and satisficing biases exist depending on the mode of data collection used. So switching from one mode to another cannot be done without studying first the impact of different data collection strategies on several parameters. In order to compare data collected with different modes (across time, across countries, across groups), many elements, including the quality of the data collected and the equivalence of measures across modes have to be assessed. Therefore it is important to determine what are the exact mode effects, and if they vary, to find a way to correct for this difference.

Quite some research has yet been done on the comparison of two or more modes of data collection, looking at variables as diverse as response rates, item non response, costs, presence of satisficing behaviours and social desirability bias, or socio-demographic characteristics of the respondents. The first wave of studies on this topic is linked to the expansion of the telephone coverage: researchers developed guidelines to transform questionnaires from one mode to another (Groves, 1990) or tried to assess the difference between telephone and mail or face-to-face data collection (Hox, De Leeuw, 1994). These kinds of comparisons are still present today in the literature (Holbrook, Green, Krosnick, 2003; Jäckle, Roberts, Lynn, 2006). The second wave is linked to the development of computer technologies and the possibility to use computer assisted methods of interview (Kalfs, Saris, 1998; Lynn, 1998; Newman et al, 2002; Perlis et al, 2004). The third wave is linked to the introduction of the Internet. The same issues are addressed but adapted to this new mode: a lot of studies focus on the comparison of Web surveys with surveys using more traditional modes of data collection (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; Kreuter, Presser, Tourangeau, 2009). Some research has also been done more particularly on mixing modes of data collection (Schonlau, Asch, Du, 2003; De Leeuw, 2005; Dillman et al., 2009; Jäckle, Roberts, Lynn, 2008).

Nevertheless, most of the previous research focuses on a comparison of costs, responses rates and eventually variables distributions. But few (Scherpenzeel, 1995; Scherpenzeel and Saris, 1997) have been done on comparing the quality of the measures in different modes. Therefore the goal of this paper is to study the impact of using different modes of data collection on the quality of survey questions, by comparing a face-to-face and a Web survey with respect to the quality of their measures, both at the single item level and at the composite score level.

It is important to notice that we are interested in comparing surveys where different modes are used *at the same stage* in order to accomplish the *same task*. More precisely, we are interested in the case where different modes are used at the response stage. So when we refer to "Web" survey (respectively "face-to-face" survey), we always mean "a survey where the questions are answered on the Web" (respectively in a face-to-face interview). The contact with the respondent can be established in the same mode or in a different mode (e.g. by a contact letter), this does not change the way we refer to the survey.

The comparison will be based on the analysis of the European Social Survey (ESS) which is administered by face-to-face and a study completed on the Web by the respondents of the LISS panel. Because a survey cannot be only described by the fact that it is a "face-to-face" or a "Web" survey, since a lot of other elements can vary, we will first conduct a general comparison of the surveys (ESS and LISS study). As pointed out by Couper and Miller (2008), two Web surveys can be extremely different, so it is necessary to be more precise about what we are studying. This will be the first section of this paper. In a second section, we will focus on the quality of the measures in the two surveys both at the single item and composite score (CS) level. By composite score, we mean an average score constructed by combining several items (questions).

1. General Comparison

1.1. The surveys

Two surveys are compared in this paper: the ESS, a bi-annual European survey which began in 2001 and where the data is collected by face-to-face interviews at the respondents' home, and a study completed by the LISS panel, an online panel of 5.000 Dutch households created in 2007. The choice of comparing these two surveys is very practical: in one of its monthly studies, the LISS panel presented to its respondents the same questionnaire as the one of the round 4 of the European Social Survey. Except for the background questions, which are treated differently since the LISS is a panel, the rest of the questionnaire was adapted from the face-to-face version to a Web version, keeping unchanged as many elements as possible. It offers therefore the opportunity to compare similar questions asked at the same moment (end 2008-beginning 2009) using two different modes of data collection, with also repetitions of the same questions using different methods. Since the LISS panel is a Dutch panel, even if the ESS is present in

more than 25 countries, in order to avoid variation due to cultural or language differences, we focus only on The Netherlands.

One limit of our approach is that by comparing different surveys, two sources of differences may be confounded: 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 in that sense be preferable to distinguish what is purely the effect of the mode on the answers. However, this same point that constitutes an advantage by allowing to isolate pure mode effect presents also a negative side: since such a design would not give any information on the potential selection when proposing different modes of data collection to sampling units, it would not give a good idea of what would happen in practice if a switch from a face-to-face to a Web survey were implemented. Besides the fact that we did not had adequate data with the same respondents answering in both modes, comparing two real surveys provides more realistic results. It is closer to what could effectively happen in case of a switch of modes. This is an important advantage of this study.

In order to know what we are comparing however, it is important to say a little more about these two surveys. Several characteristics of the surveys can be identified and mentioned[2]. First, in both cases the 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, i.e. households, is based on probability sampling. Even if all the persons in an household can participate to the LISS panel, only one person in each participating 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 selection procedure of individuals, which only allows one person per household (the one whose the birthday is next) to respond the survey. It is only when respondents answer the questions that the mode differs: face-to-face for the ESS, against online completion for the LISS respondents. This has several consequences, in particular that 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.

Even using only one country, the number of observations is high: between 1.770 and 2.370 for each ESS round, and around 3.200 in the LISS. This corresponds to response rates between 52% and 68% in the ESS rounds. In the LISS, the response rate of the study that we are studying is 65%: it means that 65% of the panel members (persons that accepted to be part of the Web panel) responded to the survey sent in December 2008 to them (which is the one used in this paper). But the panel membership rate should also be taken into account: only 48% of the sampling units accepted to participate to 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 rates. Even considering that the item non response is a little higher in the ESS (in the LISS, it is quasi inexistent: 23 interviews out of the 3200), it cannot compensate the lower general response rate. On the other end, the LISS panel was much quicker: one month only, against six months for

[2] 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 also http://www.lissdata.nl/assets/uploaded/Sample_and_Recruitment.pdf

each of the three first ESS rounds and ten months for the fourth ESS round. So the LISS panel seems much more efficient at this level, which is clearly linked to the panel dimension and not only to the mode of data collection. We refer to Table 1 for more details.

Table 1: **Some elements of comparison ESS-LISS**

	ESS	*LISS study*
Geographic area	Around 25 European countries, but 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** (respondents house)	**Web**
Interviewer	Yes	No
Stimulus	Oral + visual (show cards)	Visual
Panel	No (but several rounds)	Yes (but panel dimension not used)
Fieldwork period	R1: 01/09/2002 to 24/02/2003 (**176** days) R2: 11/09/2004 to 19/02/2005 (**162** days) R3: 19/09/2006 to 15/03/2007 (**177** days) R4: 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 of households	Probability sample	Probability sample
Selection of individuals	Only one person is selected in the household	Only one person is selected in the household
Number observations	R1: **2 364** interviews R2: **1881** R3: **1889** R4: **1775**	Complete interviews = **3194**
Response rates	R1: **67.9**% R2: **64.5%** R3: **59.8%** R4: **52.0%**	Panel membership rate = 48% total sample 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 appears from this general overview is that, as we briefly mentioned in the introduction, speaking about "Web" or "face-to-face" survey is a nice shortcut for classifying a survey, but it is an extremely simplified one. There are many different characteristics for one survey which can influence the results: even Table 1 is far from being complete and much more aspects could be compared. Therefore it is important to keep in mind that generalizing from the results of the two specific surveys which we are using in these analyses to "Web" or "face-to-face" surveys in general is not necessarily possible. In this study, we are always speaking about particular surveys which have sets of specific characteristics.

1.2. Composition of the samples

Even if the samples are drawn randomly, the characteristics of each survey may lead to possible selection bias and so differences in sample composition. We assume that usually two main elements determine people's decision to participate in a Web survey 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 focused on the second factor and considered some background variables that we thought could be related to the comfort with technology: gender, age, education and the number of persons in the household. Unfortunately, the question asking for the educational achievement of the respondents differs in the ESS and the LISS. Even if there are two measures of education in the ESS, none corresponds to the formulation of the LISS. In order to still be able to make a comparison, we create in each survey three quite large categories: low, middle and high level of education. We compare the different rounds of the ESS (1, 2, 3 and 4), the LISS study, the LISS panel and the national statistics for the Dutch population aged 16 or more[3].

We are mainly interested in comparing the fourth ESS round with the LISS study, since both have been collected at the same period. Therefore, we do not really need the first three rounds of the ESS for our purpose. Nevertheless, we include them when we can (when we have appropriate data) in order to get an idea of the variations that can appear from one wave to the next. Even when the procedure is similar, differences may appear in the composition of the samples, as can be clearly seen for instance for the gender composition when comparing the second ESS round (less men) with the three others (and this cannot be due to changes over time: the population composition in terms of gender does not vary so quickly). Regarding that, the difference in men's proportion between the ESS round 4 and the LISS study appears to be small.

Table 2: composition of the samples (in percents)

	ESS				*LISS*		*Pop*
	Ess1	Ess2	Ess3	Ess4	Study	*Panel*	NL
Gender							
men	44.1	**41.6**	45.9	46.0	44.6	*49.4*	49.2
women	55.9	**58.4**	54.1	54.0	55.4	*50.6*	50.8
Age							
16-19	4.5	3.8	3.4	4.4	**2.7**	*7.3*	6
20-39	31.1	28.9	31.7	28.8	**27.5**	*32.2*	32.7
40-64	45.6	46.0	44.2	45.5	**52.3**	*49.4*	43.3
65-79	14.5	**16.6**	16.0	17.0	15.5	*10.0*	13.4
>80	4.3	4.6	4.7	4.3	**1.9**	*1.0*	4.6
Education							
low	42.8	43.7	38.8	37.7	**35.7**	*33.0*	33.2
middle	33.9	31.6	36.5	35.6	**33.2**	*36.9*	41.4
high	23.3	24.7	24.7	26.8	**31.1**	*30.1*	25.4
Number of members in the household							
1	**22.9**	27.3	30.1	27.6	25.4	*23.7*	35.3
2	35.7	36.4	36.2	35.4	**39.4**	*35.9*	32.6
3	**14.8**	13.7	12.1	13.2	11.3	*13.5*	12.5
4	**17.8**	14.9	14.1	16.6	17.0	*18.9*	13.5
>5	**8.9**	7.7	7.5	7.2	6.9	*8*	6

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

Table 2 summarizes the results in percentages. For gender, as just mentioned, the biggest differences with the population distribution are found for the second round of the ESS: men are underrepresented whereas women are overrepresented. The same tendency appears in ESS rounds 1, 3, 4 and also in the LISS study, even if the differences are smaller.

For age, people till 39 years old are underrepresented in these five samples, whereas people between 40 and 79 are overrepresented. This trend is shared by the face-to-face and Web surveys, even if a larger difference from the population distribution is found for the LISS study.

For education, we have to be more cautious about the comparison since the response options are different in the ESS and in the LISS: this may have influenced the position of people 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 in all surveys, the group with middle educational achievement is underrepresented. In the ESS, mainly in the first rounds, this underrepresentation is opposed to an overrepresentation of the low educated. In the LISS on the contrary, it is opposed to an overrepresentation of the high educated. So using Web as a mode of data collection may tend to favour the participation of higher educated people, but 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 lower educated. Even if the robustness of this result can obviously be doubted, it is interesting to underline it, because it is often argue against Web surveys that they will discourage low educated people to participate. In our study, it is middle educated people, and not low educated ones, that in fact are underrepresented in the Web survey. Besides, they are also underrepresented in the face-to-face survey, so the mode of data collection is probably not the main explanation.

Finally, concerning the household size, all surveys show an underrepresentation of single households and overrepresentation of the households with more members.

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

So far the different samples were compared to the population. But the samples can also be compared between surveys. Some elements of Table 2 give information in that direction. It is a quite 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 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 side, indeed the more than 80 years old are more underrepresented in the LISS study than in the ESS, but the 65-79 years old are overrepresented in the LISS study (15.5 versus 13.4 in the Dutch population). The same kind of comments could be made 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 gives some elements of understanding: the youngest people are overrepresented in the panel, which probably means that when they are first approached, they are more willing to accept a Web survey because they feel more comfortable with using Internet. But then, they are not very involved, very "serious", and so the nonresponse for these 16-20 years old for one specific study (as the one we are interested in) is quite high, leading finally to an underrepresentation. The 60-79 years old on the contrary are underrepresented in the panel, but their nonresponse rate to a particular study is very small (once they agree to be part of the panel, they answer the different surveys sent to them), such that at the end they are overrepresented in the study.

1.3. Should we correct from these differences?

Table 2 shows differences in sample composition with respect to four background variables. It is difficult however to determine only from this table if the differences matter or not: we consider that the differences "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, then, they will not affect the results. But if the variables analysed correlate very little 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 other groups, even a relatively small deviation in the composition of the sample from one mode to the other may have an impact on the results.

From Table 2, we already know the sample composition. What we miss is information about the relationships between our variables of interest and the background variables. Our variables of interest are 20 variables that we will analyse more in details latter on. These are variables about position toward immigration (six questions), media use (three questions), social (two questions) and political trust (three questions), satisfaction (three questions), political orientation (two questions) and left-right self-placement (one question).

A first way to look at these relationships is to consider the correlations between the 20 variables of interest and the background variables gender (column "g" in Table 3), age ("a"), education ("e") and household size ("h"). Table 3 presents these correlations for the LISS study and the fourth round of the ESS (which is the most comparable to the LISS since both were conducted during the same period), as well as the absolute value of the difference in correlations between the two surveys.

Table 3: correlation between variables of interest and background variables

		ESS4				*LISS*				*|difference|*			
Expt	Var.	g	a	e	h	g	a	e	h	g	a	e	h
Immi-gration	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
Satisf-action	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

From the first two sets of columns, it is clear that the highest correlations are found for education. Except for media use where the correlations with age are also relatively high, the rest of the correlations are quite low. The highest differences in correlations between the face-to-face and the Web surveys (third column) are between age and immigration but even these differences are quite low. This suggests there is no 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 in both surveys, few differences should therefore be found when comparing the variables of interest in the two survey samples.

Because correlations, mainly for dummy variables or categorical variables with few categories as most of our background variables here, have many limits (e.g. very sensitive to marginal distributions), we also look at the relationships in another more precise way: considering three kinds of "results" and trying to see if they are influenced by the differences in sample composition.

First, we compare the *distributions of variables of interest* for different groups of 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 4 reports for the four ESS rounds and for the LISS when the difference is significant at the 5% level ("s") or not ("ns"). For gender, we obviously compare the distributions of the 20 variables for male and female. For age, we compare the group of the youngest respondents (less than 20) with the one of the oldest respondents (80 or more), since we expect the highest differences to be found when the groups at the two extreme points of the distribution are compared. 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). The last row and column indicates the number of differences that turn out to be significant in the corresponding row or column.

Table 4: Significance of the differences in distributions

		ESS1				*ESS2*				*ESS3*				*ESS4*				*LISS*				
Expt	Var.	g	a	e	h	g	a	e	h	g	a	e	h	g	a	e	h	g	a	e	h	# s
Immi-gration	Imsmetn	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	6
	Imdfetn	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	s	*ns*	*ns*	s	*ns*	7
	Impcntr	*ns*	s	s	*ns*	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	7
	Imbgeco	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	10
	Imueclt	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	7
	Imwbcnt	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	5
Media	Tvtot	s	s	s	s	s	s	s	s	s	s	s	s	*ns*	s	s	s	*ns*	*ns*	s	s	17
	Rdtot	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	s	s	s	s	*ns*	8
	Nwsptot	s	s	s	s	*ns*	s	s	s	s	s	s	s	s	s	s	*ns*	s	s	*ns*	s	17
Social Trust	Ppltrst	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	s	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	6
	Pplfair	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	s	s	s	*ns*	7
Political Trust	Trstprl	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	s	s	s	*ns*	*ns*	s	*ns*	11
	Trstlgl	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	s	s	*ns*	s	s	s	*ns*	s	*ns*	s	*ns*	12
	Trstplc	*ns*	*ns*	s	*ns*	*ns*	s	s	*ns*	*ns*	*ns*	s	*ns*	ns	*ns*	s	*ns*	*ns*	*ns*	s	*ns*	6
Satisf-action	Stfeco	s	s	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	11
	Stfgov	*ns*	s	*ns*	*ns*	*ns*	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	*ns*	s	s	*ns*	8
	Stfdem	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	s	s	*ns*	s	*ns*	s	*ns*	11
Political orientation	Gincdif	s	*ns*	s	*ns*	s	*ns*	s	s	s	*ns*	s	s	s	*ns*	s	*ns*	s	*ns*	s	*ns*	12
	Freehms	*ns*	s	s	*ns*	s	s	s	*ns*	s	*ns*	s	s	*ns*	s	s	*ns*	s	*ns*	s	*ns*	12
Left right	lrscale	s	*ns*	s	*ns*	*ns*	s	s	s	*ns*	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	s	*ns*	10
	# s	9	9	19	2	8	6	20	5	10	3	20	4	9	7	20	4	10	4	19	2	190

ns= non significant at 5%; s= significant; #s= number of significant differences in row / column

Table 4 shows that the differences are significant in 190 out of 400 cases which correspond to 47.5%. However, the significant differences are mainly due to education: in 98% of the tests, low and high educated respondents are distributed differently for the variables tested. On the contrary, the different age groups for instance are significantly different only in 29% of the cases. Also, it seems that more differences are found for the behavioural variables (watching television, reading newspapers) than for the opinions variables. This is consistent with the previous results: highest correlations were found between education and our variables of interest and also between the media use variables and age.

Secondly, we compare the *correlations between variables of interest*: it is very important for us to check that kind of "results" because the quality analyses are based on correlations. As for the distributions, we look at different gender, age, education and household size groups. But in this case, more data has to be considered and 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 two groups for gender, two for age, two for education, two for household size, and this in each survey (four ESS rounds and the LISS study, i.e. five surveys). So we have 190*8*5 = 7 600 numbers. As the goal of this paper is to compare quality of questions asked in different modes, we are not going to analyse such a huge data into details. We simply want to report a very crude result: there are often some differences between the groups with respect to the correlations in each of the five surveys. Because of these differences, we create weights in order to try to correct for the variations in sample composition. We have the national cross-table for age and gender, so we can easily 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, 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.

Figure 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

LISS without – with gender*age weights

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

LISS without – 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

LISS without – with education weights

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

Very few differences are found. Even for education, where the distributions for the three political trust items are significantly different (see Table 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 proportions of respondents in each education groups in our samples are quite close to the population proportions (see Table 2).

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 round 4 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 2 presents directly the differences between the unweighted correlation matrix and the weighted ones for the same six variables about trust in politics in the case of the ESS round 4.

Figure 2: Differences due to the weights (ESS round 4, trust in politics)

ESS4 without – with post stratification weights

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

ESS4 without – with gender*age weights

```
   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

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

ESS4 without – with size household weights

```
   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 design weights.

The final step is to look at the *estimates of interest* in at the end. This is our final criterium. Table 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 (M_1, M_2 and M_3) are used. Table 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 analyses will be described in the following section.

Table 5: reliability and validity coefficients with and without household size weights (political trust)

		ESS Round4						*LISS*					
Estimates		*Reliability coeff.*			*Validity coeff.*			*Reliability coeff.*			*Validity coeff.*		
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
	Weighth4	.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
	Weighth4	.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
	Weighth4	.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*

Even taking the weights producing the highest differences in correlation matrices, Table 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 other variables of interest. Besides, Table 5 focuses on estimates at the single item level, but we also did analyses with and without weights at the Composite Score level, and again, few differences were found. Therefore, we decided for the rest of the paper not to use weights.

2. Quality of the measures

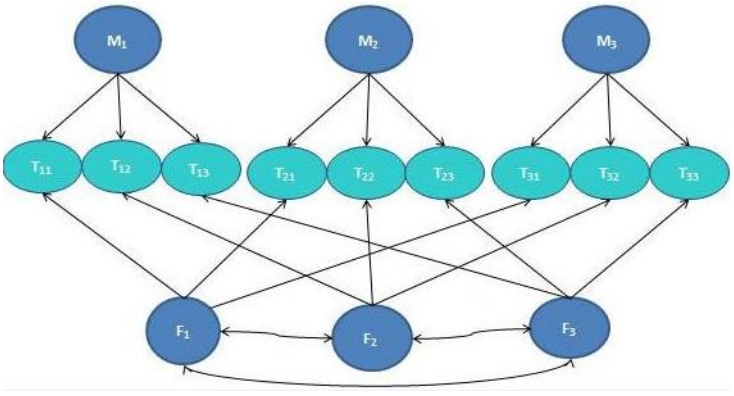
2.1. Single item level

A split-ballot multitrait-multimethod (SB-MTMM) approach

One of the most common procedures in order to assess the quality of measures is the multitrait-multimethod (MTMM) approach. As the name indicates, the MTMM designs consist in repeating *t (t*>1) questions (also called "traits") using *m (m*>1) methods: e.g. the scale of the items can contain five points in one method and seven points in another method. 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 this kind of matrices by comparing directly monotrait-heteromethod, heterotrait-monomethod and heterotrait-heteromethod blocs. It is only at the beginning of the 1970's that the MTMM matrices began to be analysed in a more elaborated way using Structural Equation Models (Werts and Linn, 1970; Jöreskog, 1970; Althauser, Herberlein and Scott, 1971; Alwin, 1974) and in 1984 that they began to be applied to single question by Andrews.

Figure 3 gives an example of an MTMM model for three traits and three methods.

Figure 3: MTMM model with 3 traits and 3 methods



The main limit of this approach is that in order to get an identified model, each question needs to be repeated at least three times. Because this can lead to memory effects and increase the cognitive burden of the respondent (Van Meurs and Saris, 1990), Saris, Satorra and Coenders (2004) propose to combine the advantages of the MTMM approach with the ones of the Split-Ballot (SB) approach: assigning *randomly* the respondents to different groups assures the comparability of the results and at the same time it allows limiting the number of repetitions for each respondent (two methods only). The model is still identified in that case under quite general conditions.

Using this design and structural equation modelling techniques, the reliability, validity and quality coefficients can be obtained for each question, estimating for instance the true score model developed by Saris and Andrews (1991) [4]:

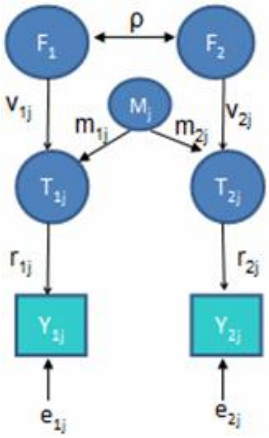
$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad \text{for all } i,j \quad (1)$$
$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad \text{for all } i,j \quad (2)$$

Where:

- Y_{ij} is the observed variable for the i^{th} trait and j^{th} method
- r_{ij} is the reliability coefficient for the i^{th} trait and j^{th} method
- T_{ij} is the true score or systematic component of the response Y_{ij}
- e_{ij} is the random error component associated with the measurement of Y_{ij} for the i^{th} trait and j^{th} method
- v_{ij} is the validity coefficient for the i^{th} trait and j^{th} method
- F_i is the i^{th} trait
- M_j represents the variation in scores due to the j^{th} method
- m_{ij} is the method effect for the i^{th} trait and the j^{th} method

Figure 4 gives a visual representation of the relations described in the equations *(1)* and *(2)* for a simplified model of two traits measured with a single method.

Figure 4: the true score model for 2 traits and 1 method



The model needs to be completed by some assumptions:

- 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

[4] Other models could be used (e.g. multiplicative model originally suggested by Browne, 1984) but Corten et al. (2002) showed, analyzing many data sets, that the additive model of Saris and Andrews (1991) should be preferred. We therefore use this model.

From this model the quality of a measure can be derived, as the product of the reliability (which is the square of the reliability coefficient) and the validity (which is the 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 corresponding observed answer Y_{ij} expressed for the j^{th} method.

Selection of topics

An MTMM approach requires a specific dataset with repeated questions. Both the ESS rounds and the LISS study included SB-MTMM experiments. However, the number of items in these experiments is quite limited in each survey (maximum six experiments with three traits and three methods, i.e. 54 items), so the possibilities of analyses are limited too. Moreover, the experiments are done on different topics in the different rounds. However, the round 4 of the ESS and the LISS study contain similar MTMM experiments. Therefore we focus on their comparison. This has also the advantage of avoiding issues linked to a potential time effect since the fieldwork of these two surveys was done in the same period (2008/2009). The six experiments analyzed are about:

- time spent on different media on an average weekday ("media")
- satisfaction ("satisf")
- political orientation ("polor")
- social trust ("soctrust")
- political trust ("trustin")
- left-right orientation ("leftright")

Each experiment contains three items usually measured with three methods. Table 6 gives more information about the different items (t_1, t_2, t_3 in "wording of the questions") and methods (*M1*, *M2* and *M3*). In one case, one of the methods was different in the two surveys: therefore the method for the ESS is mentioned into brackets. Because of the Split-Ballot design, each respondent gets only two out of the three methods (combined in different ways: *M1+M2*, *M1+M3*, *M2+M3*). The question mark 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 6: The six SB-MTMM experiments

Table 6	Var.	Wording of the questions	*M1*	*M2*	*M3*
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 (hour)	Hours and min	7 pts
satisf	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
polor	gincdif freehms ?	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 ESS)*
	ppltrst	t_1 = Would you say that most people can be trusted, or that			

soctrust	pplfair ?	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
trustin	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
leftright	lrscale ? ?	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)*

Pts = points, number of response categories; extr = labels of the end points start with "extremely"; AD = agree-disagree scales; batt = questions asked in battery; all = fully labelled; ? = variable missing in the main questionnaire of the ESS

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[5] and the pairwise deletion option of R for missing and "Don't Know" values (which are very few). Due to the Split-Ballot design, these matrices are incomplete: only the blocs (i.e. correlations or covariances) for the specific methods that each group receives are non-zero. The estimates are then obtained analysing these matrices with Lisrel by Maximum Likelihood estimation for multi-group analysis. 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 both type I (reject the null hypothesis when it was true) and type II errors (accept the null hypothesis when it is false), since it considers the power (reject the null hypothesis when it is false), which is basically one minus the type II errors. The program also tests for misspecifications at the parameter level (i.e. it tests if each specific parameter is misspecified and it does not test the model as a whole). Based on the program suggestions, in some cases corrections are introduced with respect to the general model presented earlier. Principally, the changes consist in adding a correlation between two methods when they are really similar or allowing 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 round 4. 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, some experiments being incomplete in the ESS round 4 (variables missing in the main questionnaire) but not in the LISS, it helps identifying the models and getting convergence. Table 7 gives for each topic the quality for the three traits (t_1, t_2, t_3), as well as the mean quality over these

[5] For the reasons of this choice, see Coenders and Saris (1995): "If the researcher is interested in measurement-quality altogether (…),the Pearson correlations should be used" (pp: 141)

[6] A list of all the modifications made can be found online: http://docs.google.com/Doc?docid=0AbQWMcvxT-2KZGQ3Mm10MzRfMTcwY25nZjMzczg&hl=en

[7] An example of Lisrel input is available online: http://docs.google.com/Doc?docid=0AbQWMcvxT-2KZGQ3Mm10MzRfMTY4YzVqOHRmYzk&hl=en

three traits, in each of the methods (usually three: M_1, M_2, M_3, but sometimes, only two when the third method varies), both in the ESS round 4 and in the LISS.

Table 7: Quality estimates in the ESS round 4 and the LISS study

	Quality	*ESS 4*				*LISS*			
Expt	Method	t_1	t_2	t_3	mean	t_1	t_2	t_3	Mean
media	M1 = 8pts	.90	.76	.90	.86	.90	.76	.90	.86
	M2 = h/min	.30	.68	.24	.41	.30	.68	.24	.41
	M3 = 7pts	.41	.78	.47	**.55**	.41	.80	.48	**.56**
satisf	M1 = 11 extr	.56	.73	.67	**.65**	.63	.80	.78	**.73**
	M2 = 11 very	.80	.83	.78	**.80**	.87	.89	.85	**.87**
	M3 = 5AD	.44	.67	.57	**.56**	.48	.70	.60	**.59**
polor	M1 = 5AD	.60	.56	.60	.59	.60	.56	.60	.59
	M2 = 5 pts	.76	.89	.66	.77	.76	.89	.66	.77
soctrust	M1 = 11 pts	.74	.61	.81	.72	.74	.61	.81	.72
	M2 = 6 pts	.67	.57	.68	.64	.67	.57	.68	.64
	M3 = 2 pts	.55	.50	.57	.54	.55	.50	.57	.54
trustin	M1 = 11 batt	.63	.67	.69	**.66**	.66	.65	.71	**.67**
	M2 = 6 batt	.78	.83	.78	**.80**	.83	.86	.85	**.85**
	M3 = 11 score	.72	.72	.72	.72	.70	.73	.72	.72
leftright	M1 = 11 pts	.85	.80	.73	**.79**	.94	.88	.81	**.88**
	M2 = 11 extr	.89	.83	.85	**.85**	.94	.90	.85	**.90**

Pts = points, number of response categories; extr = labels of the end points start with "extremely"; AD = agree-disagree scales; batt = questions asked in battery

In half of the cases, no significant differences were found. The other cases where there are significant differences are indicated in bold in the table (cf. columns with the means). But we can see that in general even when significant the differences are quite small (e.g. 0.55 versus 0.56 for the third method of the media experiment, or 0.66 versus 0.67 for the first method of the political trust experiment) and in favour of the LISS. The experiment where the highest differences are found is the one about satisfaction, with a difference of 0.12 for method 1.

But even in that case, differences between methods matter much more than differences between surveys: indeed, in the ESS, there is a difference of 0.14 between the mean quality in method 2 and method 3; in the LISS study this difference is even 0.28. This confirms results found in previous studies showing that agree-diagree scales are performing quite 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.

At the same time, this shows that even if some of the methods (e.g. third method of the satisfaction experiment or second method of the media experiment) lead to a relatively low quality, this is not a result of the mode of data collection used. It is due to other choices made in designing the items and response scales. Overall, we can conclude that the quality of single items seems to be quite similar in these two surveys

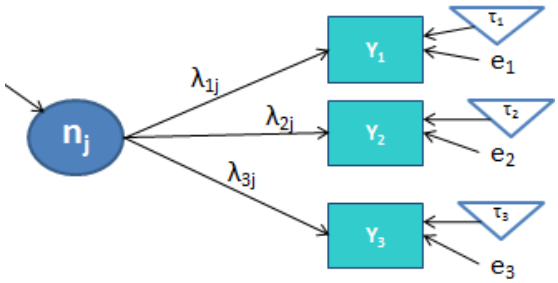
2.2. Composite score level

Most of the major concepts studied in social sciences are too complex to be measured by single items. Therefore a lot of studies are based on analyses of Composite Scores (CS). Each CS represents a concept by postulation one wants to study. This concept by postulation is defined by several concepts by intuition which are measured by items (Saris and Gallhofer, 2007). The analyses of the quality of single items should therefore be completed by an analysis at the CS level.

Test of invariance

In order to compare these CS (across time or groups), their invariance and their quality is studied. This section presents a test of invariance across groups, using the three common criteria of measurement invariance: configural, metric and scalar invariance (Meredith, 1993). Our interest is in comparing surveys collecting data in different modes. Consequently, the groups are not, as often, different countries but different surveys. Contrary to the MTMM approach, which required a very specific dataset, the CS analyses can be done on most datasets. The same questions have to be asked in the different surveys, but they do not need to be repeated within a survey. But for identification, it is recommended to have at least three indicators for each concept by postulation. Moreover, the model is different for reflexive and formative indicators. We analyse in this paper concepts with reflective indicators. Focusing on the main questionnaire of the ESS, it is possible to compare all five surveys: first, second, third and forth rounds of the ESS and LISS study. However, only a limited number of concepts are compared because few concepts by postulation have in fact three reflective indicators.

Figure 5: the basic measurement model



The basic measurement model used is presented in Figure 5. In this model, η_j is the latent variable of interest, the Y_i are the observed variables, the parameters λ_{ij} are the loadings, the parameters τ_i the intercepts and the variables e_i represent the random components in the relationships. The model can also be expressed by a system of equations:

$$Y_i = \tau_i + \lambda_{ij}\eta_j + e_i \quad \text{for all } i,j$$

In order to fix the scale of the latent variable, one of the loadings, usually λ_{1j}, is fixed to 1 and one of the intercepts, usually τ_1, is fixed to 0.

Equivalence of measures is usually decomposed in three requirements: configural (same measurement models), metric (same loadings λ_{ij}) and scalar (same loadings λ_{ij} and same intercepts τ_i) invariance. If metric invariance holds, the

comparison of the relationships between variables is allowed. If scalar invariance holds, the comparison of the means of the concepts by postulation is allowed. If scalar invariance does not hold, still partial metric or scalar invariance may hold true and allow some comparisons. For example, Byrne, Shavelson and Muthén (1989) state that consistent estimates of the means of the latent variables are obtained if at least two indicators are scalar invariant. But Saris and Gallhofer (2007) show that all indicators on which the CS are based have to be scalar invariant if one wants to compare the means of these CS.

Application: trust and attitude toward immigration

This procedure is applied to two topics: trust and attitude towards immigration. The topic of trust has been chosen because many influencing scholars, from Hobbes to Weber, passing by Smith or Durkheim, defend the idea that trust is essential for social, economic, and political life, at the micro and macro levels. Newton (2007, p.356) states 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 sciences research. Trust is also a complex concept, which can be divided in two main sub-concepts: social and political trust, quite complex themselves. Social and political trust can be seen as two correlated concepts by postulation, even if empirical research does not always find any correlation (Newton, 2007). Each of these concepts by postulation is measured by two or three items.

The second topic, attitude toward immigration, gained recently high interest in political and social sciences because of the growth of this phenomenon. Most of the countries in Europe (EU-15) have today sizeable immigrant population. Consequently, attitudes of the citizens towards with new comers have recently been studied a lot (e.g. see Coenders, 2001, or Mayda, 2006). This topic has been chosen so because of its interest but also because it is one of the most sensitive topics available in the ESS: social desirability bias may be expected to be higher in a face-to-face interview than in a Web questionnaire (self-completed, no interviewer). Finally, both topics have been chosen for practical reasons: the exact same questions (same wording, same scale) are present in the four ESS main questionnaires as well as in the LISS study. Figure 6 and **Figure 7** represent the model analysed in the case of the trust and the immigration examples. For clarity reasons, the intercepts and error terms have not all been explicitly specified, but the small arrows are here to represent them.

Figure 6: model for the trust example

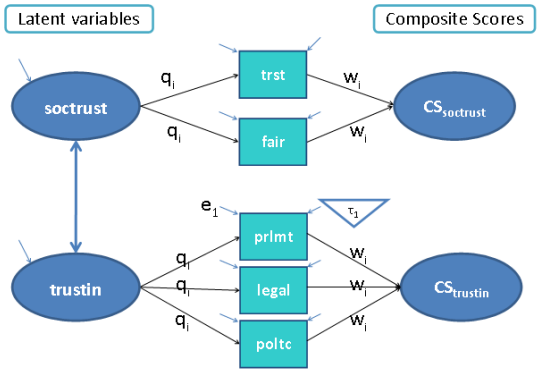
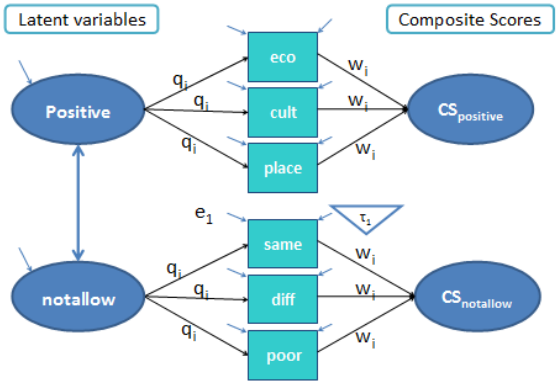


Figure 7: model for the immigration example



This model is composed of two latent variables. In the trust example, these latent variables correspond to social trust ("soctrust") and political trust ("trustin"). A correlation between them is specified, even if we expect it to be low (Newton, 2007). In the immigration example, the first latent variable ("positive") measures the positivity of the attitude towards immigration: the higher the score of respondents on this variable, the more favourable are their opinions toward immigration. On the contrary, the second latent variable ("notallow") measures the reluctance of respondents to allow more people to come and live in the Netherlands. The higher the score on this second latent variable, the less willing people are to accept more immigrants. Therefore a negative correlation is expected between these two latent variables. Each latent variable has two or three reflexive indicators. The right part of Figure 6 and Figure 7 shows that in each case these items can be used in order to create two CS. We use an unweighted model, so w_i = ½ or ⅓ depending on the number of indicators. We could use different weights (more elaborated) but we are interested in the unweighted model because it is largely used by researchers. By doing so, we want to detect whether researchers can use this kind of simple CS. The exact wording and scales of the items can be found in Table 8.

Table 8: Experiments about trust and immigration

	Var.	Meaning	Method
soctrust	ppltrst pplfair	- 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 (from negative to positive)
trustin	trstprl trstlgl trstplc	How much do you personally trust each of the institutions: - Dutch parliament - The legal system - The police	11 points (no trust to complete trust)
positive	imbgec imueclt imwbcn	- It is generally bad for the Dutch economy that people come to live here from other countries - Dutch cultural life is generally undermined by people coming to live here from other countries - The Netherlands are made a worse place to live by people coming to live here from other countries	11 points (from negative to positive)
not allow	imsmet imdfctn impcntr	- The Netherlands should allow more people of the same race or ethnic group as most Dutch people to come and live here. - The Netherlands should allow more people of a different race or ethnic group from most Dutch people to come and live here. - The Netherlands should allow more people from the poorer countries outside Europe to come and live here.	4 points (allow more to allow none)

Analyses and results

The three-step analysis (configural, metric, scalar invariance) is done using the multi-group estimation in Lisrel (Maximum Likelihood estimator) where some parameters (loadings, loadings/intercepts) are specified to be invariant across groups, i.e. in the different surveys[8]. As for the single item analyses, the testing is done using JRule. The software does not indicate any misspecification for the parameters of interest, except for one: the intercept associated to the indicator of trust in the Dutch parliament, which is misspecified in ESS round 2. In all other cases, we cannot reject

[8] The Lisrel input provided online gives more details: http://docs.google.com/Doc?docid=0AbQWMcvxT-2KZGQ3Mm10MzRfMTY3eGRrd214aG0&hl=en

the invariance. The very high power (0.99 in most cases) guarantees that the absence of misspecifications does not come from the incapacity of drawing conclusions: with such a power, even small misspecifications can be detected. But we do not find misspecifications. Consequently, this tests shows that a comparison of the means (of the latent variables as well as of the CS) is possible except for political trust in ESS 2.

According to Saris and Gallhofer (2007), for political trust in ESS 2, the CS means cannot be compared if the three items are used to compute this CS. However, the difference between this item with deviating intercept and the other items intercept value is only 0.65, while in the calculation of the CS a weight of ⅓ is used. That means that the bias in the mean would only be ⅓ of 0.65, which is around 0.22. This is such a small difference that even this item is used in the calculation of the means of the CS. For the latent means this lack of scalar invariance is not a problem. As long as there are at least two scalar invariant items the estimate of the mean of the latent variable will be consistent (Byrne et al, 1989).

The means obtained for the latent variables are very similar. In order to see if the observed differences are statistically significant, we add in the Lisrel input the constraint that they should be invariant across groups. Using JRule again to test this hypothesis, we cannot reject it while the power of the test is again very high. So the means of the latent variables seem to be equal across surveys for our four concepts.

Table 9: Comparison of the means

	Ess1	Ess2	Ess3	Ess4	LISS	Ess1	Ess2	Ess3	Ess4	LISS
Immigration	*Not Allow*					*Positive*				
mean CS	2.42	2.46	2.50	2.32	2.44	5.16	5.07	5.48	5.55	5.36
(Rank)	(4)	(2)	(1)	(5)	(3)	(4)	(5)	(2)	(1)	(3)
mean LV	2.33	2.33	2.33	2.33	2.33	5.10	5.10	5.10	5.10	5.10
Trust	*Soctrust*					*Trustin*				
mean CS	5.94	6.02	5.99	6.11	5.98	5.46	5.35	5.72	5.92	5.86
(Rank)	(5)	(2)	(3)	(1)	(4)	(4)	(5)	(3)	(1)	(2)
mean LV	5.81	5.81	5.81	5.81	5.81	5.68	5.68	5.68	5.68	5.68

Table 9 shows that these means are 2.33 for “notallow” (measured on a 4 point scale), 5.10 for “positive”, 5.81 for “soctrust” and 5.68 for “trustin” (all three measured on an 11 point scale). These means are somehow different from the ones of the CS, but the differences are small. For instance, just using the simple mean CS (unweighted) of “not allow” with the LISS data, one would get 2.44. The real mean one is interested in is in fact the mean of the latent variable, i.e. 2.33. So the size of the error is 0.11 (a 2.75% of the total scale), which can be considered most of the time acceptable. Using the data of the ESS4 for this same topic, the mean of the CS is 2.32, so the error is only 0.01. So at the end, using a simple CS based on the face-to-face or the Web data lead in both cases to a very acceptable proxy of the mean of the latent variable of interest. For the other concepts, the differences are a bit larger, but the scales are also longer, so in percentages of the scales, the error is in fact quite small. The higher deviation is for “positive” when using the ESS round 4 (0.45, i.e. 4.09% of the scale). So it seems that researchers can use simple CS and compare them across the ESS and the LISS.

We can finally notice that the social desirability bias expected is not found. Respondents do not show a more negative attitude toward immigration in the absence of

an interviewer. Also, the LISS ranks from 2 to 4. This means that the ESS values are in some rounds higher than the LISS value and in other rounds lower. So there is no clear tendency opposing the LISS and the ESS rounds.

Quality of the CS

Saris and Gallhofer (2007) define the quality of CS in the same way as the quality of single items, i.e. the quality of a CS is the strength of the relationship between the CS constructed using the observed variables and the latent variable of interest. It can be computed as the correlation squared between the latent variable of interest and the CS, using the following formula:

$$q^2_{CS} = \rho^2(LV, CS) = \left(\frac{\sum_{i=1}^{n} q_i w_i}{\sqrt{var(CS)}} \right)^2 \quad (3)$$

Where: $$\sqrt{var(CS)} = \sum_i w_i^2 var(item_i) + 2\sum_{i,j} w_i w_j cov(item_i, item_j) \quad (4)$$

Table 10 gives the results obtained by doing so. For "not allow" the quality is quite high and very similar in all surveys (around 0.90). For the three others, the quality is not so high but still higher than .70 and the differences are larger (maximum 0.12 in "trustin"). The main differences are found between the LISS and the first ESS round, which may be due to the combination of a time (round 1 done six years before the LISS study) and mode effects. In the three cases where differences are found, the LISS survey is the one which performs the best: it is quite encouraging for the future of Web surveys. Moreover, if we limit the comparison to the round 4 and the LISS (only ones where we can really assume that they should be equal if there is no mode effect), then the differences in quality between these two surveys are really small (0.01, 0.04, 0.03 and 0.06).

Table 10: Quality Composite Scores

	Ess1	Ess2	Ess3	Ess4	LISS	Ess1	Ess2	Ess3	Ess4	LISS
Immigration	*Not Allow*					*Positive*				
q^2_{CS}	.90	.87	**.91**	.89	.90	.71	.77	.77	.78	**.82**
Trust	*Soctrust*					*Trustin*				
q^2_{CS}	.72	.74	.71	.75	**.78**	.79	.79	.79	.85	**.91**

A last point about the analyses of these CS can be mentioned: describing the model earlier, we made the assumptions of a negative correlation between "positive" and "not allow". Not surprisingly, this is confirmed in all studies. Constraining the estimates to be the same in the different surveys does not lead to misspecification. The assumption cannot be rejected, so we keep it and find finally a standardized estimate around -0.6. For the trust experiment, according to the literature, a small correlation is expected between social and political trust. The standardized estimates of this correlation are around 0.5 in all surveys, which is quite high regarding past results (Newton, 2007).

Discussion / conclusion

In this comparison of a face-to-face and a Web survey, it appears that in both surveys the sample composition varies from the population distribution with respect to the main background variables, suggesting a potential selection bias. However, these differences matter only if they change the results. Looking at different results, we found that indeed differences exist in the distributions of our variables of interest and in correlation matrices when comparing different gender, age, education or household size groups. However, correcting by weighting (post-stratification for gender and age, education, and household size) does not change the correlations, neither the final estimates. Therefore, we continued the analyses without weighting. We could also have tried more complex weights combining more variables. But the differences in sample compositions seem to be relatively low, and the correlations between our variables of interest and the background variables too, so we did not expect a large effect of weighting. More complex weights probably would as well lead to very similar results, but more research on this point could nevertheless be interesting to confirm this hypothesis.

Even without corrections, the CS analyses show that the measurement instruments for the four complex concepts considered (about trust and immigration) are scalar invariant across the different ESS rounds (face-to-face) and the LISS study (Web). Therefore, one can compare means across modes and surveys, since scalar invariance holds. Because scalar invariance holds, one can also compare unstandardized relationships of the concepts with each other across modes. Besides, the results with respect to the quality of the CS show that these quality estimates are in general comparable across surveys. That means that one can also compare correlations and other standardized measures across modes.

In the case of single items, the quality is in general lower and varies a bit more depending on the modes. But 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 twice higher (0.86). The differences in quality between these methods are similar in the two surveys.

So, on the whole, the mode effect expected is not really 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 means and relationships. Knowing that the data collection is much quicker with Web and usually less expensive, a switch to Web survey seems quite attractive.

However, this study has also important limitations with respect to concepts and countries of implementation. So further research with other concepts (in particular, when complex questions or sensitive topics are used), in other countries (with different Internet coverage and different patterns of population's comfort with technologies) and comparing more modes (e.g. telephone) is necessary. Besides, more research is needed in order to confirm our conclusions because, as mentioned previously, Web surveys can be extremely different. The LISS survey is probably much more different from those opt-in Web surveys. All the efforts made to increase the representativeness of the LISS panel (in particular, provide a computer and Internet connection if necessary) are very

specific to this Web panel and do not allow generalization to other Web surveys. However, from our findings, we can conclude that the mode in itself does not systematically lead to incomparable results and that Web surveys can have a quite high quality of measure. The way of implementing the survey (whatever the mode) might make more differences.

References

- Althauser, R.P., Heberlein, T.A., Scott, R.A. (1971). "A causal assessment of validity: the augmented multitrait-multimethod matrix". In *Causal Models in the Social Sciences,* ed. H. M. Blalock Jr., pp. 374-99. Chicago: Aldine

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

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

- Browne, M. W. (1984). The decomposition of multitraitmultimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37, 1-21.

- Byrne, B.M., Shavelson, R.J. and Muthén, B. (1989). "Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance". *Psychological Bulletin* 105(3): 456-466.

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

- 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 Saris, W.E. (1995). "Categorization and measurement quality. The choice between Pearson and Polychoric correlations". In W.E. Saris, *The MTMM approach to evaluate measurement instruments* (1995), Chapter 7, 125-144.

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

- Couper, M.P, Miller, P.V (2008). "Introduction to the special issue". *Public Opinion Quarterly*, Vol. 72, No. 5, pp. 831–835

- De Leeuw, E.D (2005). "To Mix or Not to Mix Data Collection Modes in Surveys". *Journal of Official Statistics*. Vol. 21, No. 2, 2005, pp. 233-255

- Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., Messer, B.L. (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.

- Faas T. and H. Schoen (2006). "Putting a questionnaire on the Web is not enough- A comparison of Online and Ofline survey conducted in the context of the German Federal Elections 2002". *Journal of Official Statististics*, 22 ,177 - 191

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

- Forsman, G. and A. Isaksson (2003). "A comparison between using the Web and using the telephone to survey political opinions", Report nr. 20 from the project Modern statistical survey methods.

- Groves, R.M (1990). "Theories and Methods for Telephone Surveys". *Annual review of Sociology*, 1990. 16:221-40

- Heerwegh, D., Loosveldt, G. (2008). "Face-to-Face Versus Web Surveying in a High-Internet-Coverage Population. Differences in Response Quality". *Public Opinion Quarterly* 2008, 72: 836 - 846

- 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*, Vol. 21, No. 1, pp. 111-119. Oxford University Press.

- Holbrook, A.L., Green, M.C, Krosnick, J.A. (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.

- Hox, J.J., De Leeuw., ED (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

- Jäckle, A., Roberts, C., and Lynn, P. (2006). "Telephone versus Face-to-Face Interviewing: Mode Effects on Data Quality and Likely Causes. Report on Phase II of the ESS-Gallup Mixed Mode Methodology Project", ISER Working Paper 2006-41. Colchester: University of Essex

- Jäckle, A., Roberts, C., Lynn, P. (2008). "Assessing the Effect of Data Collection Mode on Measurement". ISER Working paper No 2008-08, February 2008

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

- Kalfs, N., Saris, W. (1998). "Large Differences in Time Use for Three Data Collection Systems". Social Indicators Research. Volume 44, Number 3, July 1998, pp: 267-290. Springer Netherlands.

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

- Knoef, M., de Vos, K. (2009). "The representativeness of LISS, an online probability panel". CentERdata

- Kreuter, F., Presser, S., Tourangeau, R. (2009). "Social Desirability Bias in CATI, IVR and Web Surveys: The Effects of Mode and Question Sensitivity". *Public Opinion Quarterly*, Vol. 72, No. 5 2008, pp. 847–865. Publisher by Oxford University Press.

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

- Krosnick, J.A. (1999). "Survey Research". *Annual Review of Psychology*, 1999, 50:537-67

- Lozar Manfreda, K., Bosnjak, M., Haas, I., and Vehovar, V. (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. (1998). "Data Collection Mode Effects on Responses to Attitudinal Questions". *Journal of official statistics*, Vol. 14, No. 1, pp. 1-14.

- Lynn, P., Laurie, H., Jäckle, A., Sala, E. (2006). "Sampling and Data Collection Strategies for the Proposed UK Longitudinal Household Survey" (draft version)

- Mayda, A.M (2006). "Who Is Against Immigration? A Cross-Country Investigation of Individual Attitudes toward Immigrants". *The review of Economics and Statistics*. August 2006, Vol. 88, No. 3, Pages 510-530

- Meredith, W. (1993). "Measurement invariance, factor analysis and factorial invariance". *Psychometrika* 58: 525-543.

- Newman, J.C, Des Jarlais, D.C., Turner, C.F, Gribble, J., Cooley, P., and Paone, D. (2002). "The Differential Effects of Face-to-Face and Computer Interview Modes". *American journal of Public Health*. February 2002, Vol. 92, No. 2.

- Newton, K. (2007). "Social and Political Trust". *The Oxford Handbook of Political Behavior* Chapter 18: 342-359.

- Perlis, T.E., Des Jarlais, D.C., Friedman, S.R., Arasteh, K., Turner, C.F. (2004). "Audiocomputerized self interviewing versus face-to-face interviewing for data collection at drug abuse treatment programs". *Addiction*, 99:885-896.

- Saris, W.E. and Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Paul P. Biemer, Robert M. Groves, Lars Lyberg, Nancy Mathiowetz, and Seymour Sudman (Eds.), *Measurement errors in surveys.* Pp. 575-597. New York: Wiley.

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

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

- Saris, W.E, Satorra, A., Van der Veld, W.M. (2009). "Testing Structural Equation Models or Detection of Misspecifications?" *Structural equation models.*

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

- 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.

- Schonlau, M., B. J. Asch, and C. Du (2003). "Web Surveys as Part of a Mixed-Mode Strategy for Populations that Cannot be Contacted by e-Mail." *Social Science Computer Review* 21 (2):218.

- Schonlau M., Zapert K., Payne Simon L., Sanstad K., Marcus S., Adams J., Spranca M., Kan H.J., Turner R., Berry S. (2004) "A comparison between a propensity weighted Web survey and an identical RDD Survey". *Social Science Computer Review*, 22 (1)

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

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

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