FORS
explore.understand.share.

Caroline Vandenplas and Oliver Lipps

Robustness of items within and across surveys

Lausanne, October 2014

FORS Working Papers 2014-03

How to cite this document:
Vandenplas, C. & Lipps, O. (2014). Robustness of items within and across surveys. *FORS Working Paper Series,* paper 2014-3. Lausanne: FORS.

Acknowledgements
The authors would like to thank Boris Wernli and Ivett Szalma for their suggestions that greatly helped to improve the paper.

ISSN 1663-523x (online)

FORS
c/o University of Lausanne, Géopolis
1015 Lausanne
Switzerland
E-mail: paperseries@fors.unil.ch

Summary

Sociologists often draw conclusion based on one single survey. Estimates from survey items can however be affected by different errors such as nonobservation bias, interviewer effects, question designs or measurement effects more generally. The sources of these errors heavily depend on the survey design and the budget allocated to it (mode, contact procedure, refusal conversion etc.). In addition, weights may correlate little with substantive variables and using weights in the analysis model may not be sufficient to correct for nonobservation bias.

In this paper we compare the mean values of three often analysed items (political interest, satisfaction with democracy and health) and try to find the possible sources of error explaining found differences. We analyse discrepancies of means within (using or not using post-stratification weights) and between six different surveys run at the same time: The Swiss part of the European Social Survey in 2010, the Swiss part of the International Social Survey Programme (ISSP) in 2011, the Swiss Household Panel in 2011, the Swiss electoral study (CATI and web) in 2011 and the Swiss Labor Force Survey in 2010.

Results show that while there are small differences within surveys, large differences may occur between surveys. The differences in means can probably be explained by selection bias that coverage and nonresponse weight adjustments fail to correct for, or measurement bias due to question wording, different answer categories, or different modes. It should be kept in mind that it is very difficult, in a non-experimental set-up, to identify and disentangle the different sources of error. The paper raise however awareness about drawing conclusions based on a single survey and the survey errors that should be taken into account.

Keywords: weighting, comparing surveys, satisfaction scores across surveys, selection effects, survey errors.

Robustness of items within and across surveys

Caroline Vandenplas[1]
Oliver Lipps[2]

1. Introduction

Estimates from survey data suffer from different types of errors due to sampling, data collection, and data processing. Hence different survey designs may give rise to different estimates that could eventually lead to different conclusions if the errors are not adequately considered (e.g. Heath et al. 2009). The total survey error (TSE) framework (Biemer 2010, Groves & Lyberg 2010) describes the different sources of errors affecting survey data and how to minimize the total error taking cost restrictions into account. In TSE, two types of errors are identified, errors of nonobservation and measurement errors.

In this paper, we will compare the mean value of survey items that are included in different social surveys conducted in Switzerland at the same time (2010-2011). We selected three items – political interest, satisfaction with democracy and self-rated health – because of their popularity in recent research (e.g., Bopp et al. 2012, Hadjar & Schlapbach 2009, Krieger et al. 2011, Stadelmann-Steffen & Vatter 2012, Voorpostel & Coffé 2012).

We start by giving a brief review of the different survey error sources that can be affected by different survey designs.

Sampling frames from which sample members are drawn are the first source of error. Even a population register excludes people like recent immigrants or those living illegally in the country, or may not be up-to-date. Depending on survey modes, the population coverage is biased on standard socio-demographic variables: for landline telephone surveys, young people, singles, and non-Swiss citizens are less likely to have a listed landline number available (Lipps and Kissau 2012, Lutz et al. 2012, Mohorko et al. 2013). Web users are known to be rather young, higher educated and more often male (Bethlehem 2010, Callegaro 2013, Lipps & Pekari 2013, Lynn 2013, Scherpenzeel 2001). More severely, other (unobserved) sample characteristics may also be biased. Sampling frames also influence the sampling design and the sampling error associate with it. If an individual is part of the survey population but no list of the

[1] Katholieke Universiteit Leuven, Caroline.Vandenplas@soc.kuleuven.be
[2] FORS, Lausanne, Switzerland, oliver.lipps@fors.unil.ch

target population exists, one has in general to draw in a first step towns, addresses or households, and in the second step individuals within these first step groups. This extra sampling step increases the sampling error.

During the data collection period, a main source of bias emerges from nonresponse: some sample members cannot be contacted or refuse to participate (for example, Bethlehem et al. 2011, Groves & Couper 1998, Stoop et al. 2010). Like undercoverage, nonresponse generally produces selection effects, on both socio-demographic and substantive variables. Young or old people, the lower educated and the socially disadvantaged, non-owners, those living in urban areas, the unmarried and foreigner have a higher likelihood to not respond in social surveys. This is generally more pronounced in CATI than in CAPI surveys (Holbrook et al. 2003, Béland & St-Pierre 2008). Web respondents, in turn, tend to be younger, male, and higher educated than telephone respondents (Braunsberger et al. 2007, Lipps & Pekari 2013, Nagelhout et al. 2010, Roster et al. 2004, Vandenplas et al. 2013). There are different approaches to reducing nonresponse bias, e.g.; targeted/responsive fieldwork (Groves & Heeringa 2006, Luiten & Schouten 2013, Peytchev et al. 2010, Wagner 2008), multiple imputation, or weighting. To account for problems arising from undercoverage and nonresponse, different weights can be designed to adjust the final respondent sample to the population under consideration. Non-observation adjustment weights attempt to reduce bias by correcting for under-coverage and non-response (for example, Bethlehem 2002, Deville & Särndal 1992, Heckman 1976, Kass 1980, Kalton & Flores-Cervantes 2003). Weights affect survey (point) estimates, hopefully reducing bias, but also often increase the standard errors.

Another common source of error is measurement. To design questions such that they measure the intended construct can be challenging (Saris & Gallhofer 2007). Next, attention must be given to choosing a scale, that should cover the entire measurement continuum, be truly ordinal and such that the meaning of different points does not overlap. Finally all respondents should have the same interpretations of each point (Krosnick & Presser 2010). Moreover, including a middle point or not can strongly influence results. On one hand, a middle point offers an answer option for respondents who truly have no opinion but on the other, some respondents that could easily give a substantive answer if there was no middle point, choose the middle point to reduce cognitive effort. It is however believed that adding a middle point increases reliability and validity of the measurement (O'Muircheartaigh et al. 1999). The number of points on the scale is also important; a dichotomous or a trichotomous scale (agree, neutral, disagree) may be easy to interpret but some respondents may not be able to map their opinion on such a scale. Studies have shown that reliability increase from 2 to 7-point scales and to a lesser extend to 11-point scales (Birkett, 1986, Givon & Shapira, 1984, Masters, 1974). Similarly the validity of the items grows with the length, but less with longer scales (Green & Rao 1970, Lehmann & Hulbert 1972, Lissitz & Green 1975, Martin 1973, 1978, Ramsay 1973). Labelling all points of the scale or only some (usually endpoints only) also influences the measurement effect; small scales can easily be fully labelled but doing so for longer scales can be time-consuming to read. At the same time, if a scale is not fully labelled the respondents has to work out the meaning of the unlabeled points which asks for some cognitive effort (e.g., Dickinson &

Zellinger 1980). Research has shown that reliability is higher when the scale is fully labelled (e.g., Krosnick & Berent 1993). Needless to add that the number of scaling points and labelling depends on the survey mode, with for example fully labelled long scales being simply impossible in telephone surveys.

In addition, other factors play a role depending on the mode of data collection. Examples are interviewer effects such as social desirability especially in interviewer-based surveys, or straight-lining in web or paper questionnaires (Biemer et al. 2004). We thus expect higher social desirable answers in CATI and CAPI surveys than in web surveys (Kreuter et al. 2009, Sakshaug et al. 2010). Even though one may expect a higher impact of social desirability in face-to-face surveys due to the physical presence of an interviewer, there are no systematic differences between measurement effects in CATI and CAPI (Jäckle et al. 2010, Ye et al. 2011). De Leeuw & van de Zouwen (1998) even found more social desirability in CATI surveys.

The literature on comparing results from different surveys is quite sparse. Some test for example two modes in an experimental design to separate measurement and sample selection effects (Duffy et al. 2005, Jäckle et al. 2006, Scherpenzeel 2001, Vandenplas et al. 2013, Vannieuwenhuyze & Loosveldt 2014). Brown et al. (2005) compared learning achievement and functional literacy for international large-scale assessment studies (IALS[3], PISA[4], PIRLS[5] and TIMSS[6]), and found that there is generally a reasonable degree of agreement between the surveys. In 2013, Dahlberg and Persson compared the (mainly telephone) European Election study (EES) and the face-to-face Swedish National European Parliament study to assess the bias in the European Election study. As a reference they used official register data. They found that the EES overestimates turn-out levels and has a large overrepresentation of highly educated citizens as well as more extreme positioning of parties and of their political opinion. Our last example comes from the medical sector: In 2008, Bethell et al. compared three national surveys in the US in terms of the prevalence of children with special health care needs. They considered two telephone surveys (the 2001 National Survey of Children with Special Health Care Needs (NS-CSHCN), and the 2003 National Survey of Children’s Health (NSCH)), and the 2001-2004 face-to-face Medical Expenditures Panel Surveys (MEPS). They found that the prevalence was 4.8 percentage points higher in the NSCH and 6.0-6.5 percentage points higher in the MEPS than in the NS-CSHCN (12.8%).

In this paper, our main objective is to evaluate differences in mean values of three items included in a number of Swiss large social surveys that were conducted in in 2010 or 2011: political interest, satisfaction with democracy, and subjective health status. We first describe the considered surveys and their characteristics. Next, we analyze selection bias on socio-demographic variables in the different surveys, before we compare the estimated mean values, including or not nonresponse adjustment

[3] International Adult Literacy Survey
[4] Program for International Student Assessment
[5] Progress in Reading and Literacy Study
[6] Trends in International Mathematics and Science Study

weights and interviewer design effects within a survey and across surveys. Then, we discuss the different possible error sources that cause differences. Finally, we summarize and discuss our results. We conclude that researchers in social science should take the different survey design features that could bias their results into account and compare their outcomes with existing research to validate them.

2. Data and method

The Swiss Household Panel (SHP) 2011

The Swiss Household Panel (SHP; http://www.swisspanel.ch/) is an annually, centrally conducted CATI panel survey, based on a stratified random sample of the Swiss residential population. Samples were drawn at random from the telephone register, excluding unlisted telephone numbers. The SHP is an academically driven multi-topic survey run by the Swiss Centre of Expertise in the Social Sciences (FORS). It started in 1999 with more than 5,000 households and added a refreshment sample with more than 2,500 households in 2004 (wave 6), also randomly selected from the telephone register. All household members aged 14 or over are interviewed. Fieldwork is conducted each year between September and January using about 100 interviewers.

The cross-sectional weighting scheme of the SHP is rather complex (Voorpostel et al. 2013). All weights are based on those of the first wave. The adjustment for nonresponse is done by segmentation (Kass 1980), based on the language area, the seven big regions of Switzerland, the type of commune, the household size, children in the household, age, marital status, education, gender, working status, nationality and type of permit (Garriguet & Latouche 2004). The combination of the original and the refreshment sample is done using the method of Merkouris (2001), which allocates a relative importance to each sample according to its size. Finally, cross-sectional calibration match population totals on age*gender, the big regions of Switzerland, Swiss nationality and being married.

European Social Survey (ESS) 2010 / MOSAiCH 2011

The ESS (http://www.europeansocialsurvey.org/) is an academically-driven cross-national survey that has been conducted every other year in some 25 European countries since 2002, the Swiss part also being run by FORS. The survey measures attitudes, beliefs and behavior patterns. The Swiss part of the 2010 ESS includes around 1,500 CAPI surveyed respondents, sampled by a Simple Random Sampling (SRS) procedure of individuals aged 15 or over using the national individual population register, and interviewed by more than 50 interviewers. The ESS provides post-stratification weights using information on age-group, gender, education, and region.

The MOSAiCH (Measures and Sociological Observation of Attitudes in Switzerland) is the Swiss part of the International Social Survey Programme (ISSP), and has similar properties as the ESS with however a somewhat smaller number of respondents of about 1,000, and is run in the uneven (non-ESS) years. No nonresponse adjustment weights are provided for the MOSAiCH.

Selects: CATI and web 2011

Selects (www.selects.ch) is a CATI post-election survey designed to study voting behavior and is conducted every four years after the Swiss federal elections. In 2011, an online experimental survey was run alongside the usual CATI survey using the same questionnaire content. The samples for both surveys were drawn from the national individual population register, representative of Swiss citizens aged 18 years or older. SRS stratified by the cantons is used. Smaller cantons are overrepresented in Selects CATI such that the number of respondents in the smallest cantons amounts to about 100. Telephone numbers for the CATI were matched using the telephone register owned by the Swiss Federal Statistical Office and listed matched telephone numbers were delivered to the survey agency. To match lacking addresses, commercial sources like getstone.ch were used resulting in an overall matching rate of about 86% (Lipps et al. 2013). The CATI has about 2,350 respondents, the web survey about 440 respondents.
Design weights for Selects (CATI) correct for the overrepresentation of some cantons, and nonresponse adjustment weights for biased voting participation and party choice. No weights were designed for the web survey.

Swiss Labor Force Survey (SLFS) 2010 (first quarter)
The main purpose of the SLFS is to provide information on the structure of the labor force and employment behavior patterns in Switzerland and is conducted each quarter since 2010. The SLFS is a centralized CATI survey based on a sample of some 60,000 respondents selected at random from the telephone register owned by the Swiss Federal Statistical Office, including unlisted telephone numbers. The SLFS has a panel component; in the sample considered here, 56% of the respondents are in their first, and 44% in their third survey wave. The SLFS is representative of the permanent resident population aged 15 or older.

The weights for the Swiss Labor Survey are threefold, in a first step, design weights are calculated, in a second step, weights that correct for possible future attrition based on sex, age and participation in previous waves are calculated. Finally, post-stratification weights based on sex, age, nationality, civil status and geographic location are added. Unfortunately in the SLFS design weights and post-stratification weights are not separated and only the combined weight is delivered.

To make comparisons across surveys possible, we restrict the analysis sample on Swiss citizens, who are 18 years or older. Because in the SLFS only one (random) person per household is selected to be interviewed, we (design-)weight the sample members of the household sampled SLFS by the number of household members aged 18 or older. [7]

Coverage and Response Rates
The surveys considered differ in their implementations and in their purpose. They also differ in the amount of under-coverage (for telephone and web surveys), and nonresponse. In the post-election survey selects 86.2% of the gross sample could be

[7] The underlying assumption of one telephone number per household is only approximate since it is likely that households with more adults have a higher probability of having more telephone numbers.

matched to a telephone number (Lipps et al. 2013). Of the individuals for whom a telephone number was matched 29% participated in the survey (Lipps & Pekari 2013). The SLFS has an estimated under-coverage rate of 15% and reports a 22.0% nonresponse rate for the data considered here (SFSO 2012). The ESS 2010 and MOSAiCH 2011 had (similar) response rates of respectively 53.3% and 53.2% (AAPOR RR1) (ESS 2010 Documentation, Vandenplas et al. 2014). In the web experiment of Selects, 29.6% of the gross sample completed the survey online. Finally, the first wave of the SHP had a household response rate of 64%; within participating households the individual-level of participation was 85%. In the 2004 refreshment sample, these figures amount to 65% and 76%, respectively. Of the participating household in the first wave, 58% participated in 2011 and of the original responding households of the refreshment sample in 2004, 59% participated in 2011 (Voorpostel et al. 2013).

Weights

The surveys in this article are drawn with a SRS design[8] and the sampling error depends on the sample size only. No design weights are thus needed. Exceptions are the telephone surveys Selects CATI, where design weights correct for the overrepresentation of small cantons, and the SLFS, where design weights correct for the unequal inclusion probability of sample members in households of different sizes. The auxiliary variables used to construct the post-stratification weights are very different. While in ESS, SHP and the SLFS socio-demographic variables are used, in Selects CATI weights that adjust for bias in voting turnout and party choice (the substantive variables of interest) are used. Selects web and MOSAiCH do not contain adjustment weights.

Expectation for the items

We expect political interest to be higher in the election surveys due to the topic. Probably the same will hold for the other telephone surveys given coverage bias (the less interested are typically overrepresented among those without a landline phone). While the online survey could involve less social desirability with lower mean estimates in general, selecting more educated and younger people would in turn predict a higher political interest. Similar considerations hold for satisfaction with democracy. Health is expected to be worse in the web component of the election survey both due to selection and measurement effects and better in the telephone surveys. We expect that the face-to-face surveys ESS and MOSAiCH measure the lowest scores because they do not suffer from undercoverage, and have generally higher response rates, although social desirability could boost the score compared to the web survey.

[8] We do not consider stratification by the seven large regions in Switzerland which is used in some surveys. Its effect is negligible.

3. Results

Selection bias

Surveys can have different coverage and/or response rates but if the response is random between surveys, there are no selection effects. In the following table, we compare the distribution of gender, age (6 categories: 18-27, 28-37,…, 68 and more), education (2 categories: 1=university, applied science or pedagogical university)[9,] employment status (1=full-time working), marital status (1=married or living in a registered partnership), household composition (3 categories: 1, 2, and 3 or more persons) and language in which the interview has been conducted (German, French, Italian, English (for the SLFS)). We use the person-weighted distributions from the pooled 2010/2011 Swiss structural surveys[10] as a reference. The socio-demographic distributions of the samples across our surveys are depicted in For a discussion of the reason of selection bias, we refer the interested readers to other literature (for example, Voorpostel et al. 2013 for the SHP, Lipps et al. 2013 for representation for Selects, and Pollien & Joye 2014 for the ESS and MOSAiCH).

In **Error! Not a valid bookmark self-reference.**, we see that men are underrepresented in the SHP and the SLFS, which are both longitudinal telephone surveys. Generally in telephone surveys (SHP, Selects CATI, SLFS), women are known to be overrepresented (e.g., Schneiderat & Schlinzig 2012). Attrition does however not play a role concerning gender selection in the SHP and the SLFS (results not shown). Men are overrepresented in all other surveys with the highest percentage in the web survey (50%).

The percentages of respondents in the younger age category are relatively close to the target statistics (15.0%) with the exception of Selects web (too high), and Selects CATI and especially the SLFS (too low). This can be expected from the survey mode: while young people have a high internet availability, they suffer from under-coverage in (fixed line) telephone surveys. The 28 to 37 years old are underrepresented (Groves & Couper 1998, Stoop 2005) in all surveys but again the web survey. The age category 38 to 47 is mostly well represented, but again overrepresented in the web survey. Respondents between 48 and 57 years old are slightly overrepresented, whilst the 58 to 67 are overrepresented in all but Selects web. Finally, the oldest (68 and above) are well represented, but strongly underrepresented in the web and overrepresented in the SLFS. This also meets our expectations from the literature.

[9] Reasons for only two education categories are the difficulty encountered when comparing the categories in the different surveys other than the highest education levels.

[10] http://www.bfs.admin.ch/bfs/portal/en/index/infothek/erhebungen__quellen/blank/blank/rs/01.html

Table 1.

For a discussion of the reason of selection bias, we refer the interested readers to other literature (for example, Voorpostel et al. 2013 for the SHP, Lipps et al. 2013 for representation for Selects, and Pollien & Joye 2014 for the ESS and MOSAiCH).

In **Error! Not a valid bookmark self-reference.**, we see that men are underrepresented in the SHP and the SLFS, which are both longitudinal telephone surveys. Generally in telephone surveys (SHP, Selects CATI, SLFS), women are known to be overrepresented (e.g., Schneiderat & Schlinzig 2012). Attrition does however not play a role concerning gender selection in the SHP and the SLFS (results not shown). Men are overrepresented in all other surveys with the highest percentage in the web survey (50%).

The percentages of respondents in the younger age category are relatively close to the target statistics (15.0%) with the exception of Selects web (too high), and Selects CATI and especially the SLFS (too low). This can be expected from the survey mode: while young people have a high internet availability, they suffer from under-coverage in (fixed line) telephone surveys. The 28 to 37 years old are underrepresented (Groves & Couper 1998, Stoop 2005) in all surveys but again the web survey. The age category 38 to 47 is mostly well represented, but again overrepresented in the web survey. Respondents between 48 and 57 years old are slightly overrepresented, whilst the 58 to 67 are overrepresented in all but Selects web. Finally, the oldest (68 and above) are well represented, but strongly underrepresented in the web and overrepresented in the SLFS. This also meets our expectations from the literature.

Table 1: Mean values of socio-demographic variables in each survey (standard errors in brackets)

	Population[a]	SHP	Selects CATI[g]	Selects web	ESS	MOSAiCH	SLFS[c]
	2010/11 (%)	2011 (%)	2011 (%)	2011 (%)	2010 (%)	2011 (%)	2011 (%)
Male	47.7	44.0	48.6	50.1	49.0	49.9	45.8
	(.0009)	(.0062)	(.0103)	(.0239)	(.0144)	(.0158)	(.0025)
Age							
18-27	15.0	14.8	13.2	18.0	14.2	14.8	11.4
	(.0007)	(.0044)	(.0080)	(.0184)	(.0100)	(.0112)	(.0019)
28-37	13.8	10.2	10.7	15.5	10.8	11.8	11.7
	(.0006)	(.0038)	(.0075)	(.0173)	(.0089)	(.0102)	(.0016)
38-47	18.4	18.9	18.6	23.5	18.9	17.4	18.4
	(.0007)	(.0049)	(.0092)	(.0202)	(.0112)	(.0120)	(.0019)
48-57	18.3	21.0	20.3	19.6	20.5	19.7	18.4
	(.0007)	(.0051)	(.0096)	(.0190)	(.0116)	(.0126)	(.0020)
58-67	15.6	17.5	18.1	15.0	17.8	17.5	18.3
	(.0006)	(.0047)	(.0091)	(.0171)	(.0110)	(.0120)	(.0019)
68+	18.7	17.7	19.0	8.4	17.8	19.0	21.8
	(.0007)	(.0048)	(.0094)	(.0133)	(.0110)	(.0124)	(.0019)
Full-time	38.5	33.6	34.6	42.6	44.7	42.1	37.0
	(.0009)	(.0059)	(.0113)	(.0236)	(.0143)	(.0156)	(.0025)
Married	53.1	56.9	59.3	55.6	57.7	57.7	61.0
	(.0009)	(.0062)	(.0117)	(.0237)	(.0142)	(.0156)	(.0024)
HHld size							
1	20.6	16.4	16.0	16.6	18.2	18.4	16.5
	(.0007)	(.0046)	(.0087)	(.0179)	(.0111)	(.0122)	(.0014)
2	35.3	38.7	37.4	37.1	39.6	40.6	41.6
	(.0008)	(.0061)	(.0115)	(.0232)	(.0141)	(.0155)	(.0024)
3+	44.1	44.9	46.7	46.3	42.2	41.0	41.9
	(.0009)	(.0062)	(.0119)	(.0240)	(.0142)	(.0155)	(.0026)
Language	[b]				[f]		
German	71.9	71.5	76.8	71.8	77.4	75.4	72.9
	(.0008)	(.0056)	(.0096)	(.0215)	(.0120)	(.0136)	(.0022)
French	21.2	25.3	19.6	21.6	19.9	21.4	21.6
	(.0007)	(.0054)	(.0090)	(.0197)	(.0115)	(.0130)	(.0021)
Italian	4.4	3.2	3.6	6.6[d]	2.7	3.2	5.4[d]
	(.0003)	(.0022)	(.0039)	(.0119)	(.0047)	(.0056)	(.0011)
English	.3	-	-	-	-	-	.2
	(.0001)						(.0002)
University	12.5	16.5	20.4	23.5	12.8	13.9	14.6
	(.0006)	(.0046)	(.0095)	(.0202)	(.0096)	(.0109)	(.0018)
N	403,182	6,408	2,354	439	1,212	1,002	46,656

[a] Pooled Swiss structural surveys 2010 and 2011, person weighted.

[b] Main language of respondent (not necessarily language of interview). 2.2% have a main language other than German, French, Italian or English.

[c] design-weight which only controls for the unequal selection of the number of adult persons in household.

[d] not comparable, the Ticino was oversampled (target proportion in the SLFS: 6%; SFSO 2012).

[f] Language region (not necessarily language of interview).

[g] design-weighted to control for overrepresentation of sample members from small cantons.

The representation of the full-time working people varies in a strong way which is also due to the different definition of this variable. It seems though that telephone surveys get a lower estimate (underrepresentation) than face-to-face or web surveys (overrepresentation). It is possible that full-time working people are less likely to answer the telephone and find easier means to avoid being interviewed in telephone surveys, e.g. refusal by proxy.

People who are married or live in a legal partnership are overrepresented in all surveys, with the highest percentages in Selects CATI and the SLFS. Although attrition was slightly higher among the married in the refreshment sample in the SHP between 2004 and 2011 (figures not shown), married people are still overrepresented in the SHP 2011.

People living alone represent between 15% to slightly more than 18% of the total population in all surveys, and are thus underrepresented (target 20.6%). To the contrary, people living in larger households are overrepresented, with the exception of large households with three or more persons in the face-to-face surveys and the SLFS.

Between 72 to 77% of the respondents completed the interview in German, with too high numbers in the Selects CATI, ESS and MOSAiCH. French is overrepresented in the SHP.

Finally, the percentage of respondents having completed university is overestimated in the telephone surveys and especially in the Selects web. This last finding is probably due to the correlation between political interest and telephone coverage and especially internet coverage.

In total, Selects web stands out from the other surveys with more men, more young (less than 47 years old), more full-time working and higher educated people. These biases were to be expected based on previous research on coverage and selection effects in web surveys and probably emphasized by the political topic of the survey.

Selects CATI has an overrepresentation of married, highly educated and non-single households, while single households and full-time workers are underrepresented. This can be due both to the topic of the survey and the telephone data collection mode.

In the SHP, men, single living people, and full-time working people are underrepresented and a higher percentage of interviews were completed in French. Both could be a consequence of the survey mode and the type (longitudinal) of the survey. The overrepresentation of the French speaking population could be due to a "home effect"[11].

In the SLFS, young adults and those living alone are underrepresented, while there are too many old or married people, or people living as a couple. There might be an authority effect (note that the SLFS is the only survey which is run by a federal statistical office). Also this could be a consequence of the design weight (number of adults in the household) which is possibly overweighting larger households.

[11] Both the headquarter of the agency conducting the interviews and the SHP group running the survey are located in the French speaking area of Switzerland.

ESS and MOSAiCH have, not surprisingly, very similar respondent compositions with quite a high percentage of people working full-time and living alone compared to other surveys as well as a higher percentage of interviews completed in German. These effects are likely due to the face-to-face survey mode, and when comparing to the other survey modes, to the better population coverage.

Means of three survey items

For each of the three items, we calculate the mean values and the standard errors using the design-weighted data, the post-stratification-weighted data, and finally with interviewer clustering (the extra estimation variance due to the interviewer variance, if applicable) taken into account. We compare the results and discuss the possible error sources which may influence the estimate.

As it has appeared to be crucial, we list the question wording and the response categories of the items considered in English (German and French translation is also available[12]) in a table for each variable. This gives an idea of the scales and wording used in the different surveys and their possible (different) impact on the measurement error.

Political Interest

Table 2 displays the different question designs in the different surveys. The SHP uses an 11-point scale with a midpoint with only the endpoints labelled, whilst the other surveys use a fully labelled 4-point scale. The wording between ESS/MOSAiCH and SELECTS studies differ however a little.

Table 2: Question wording and answer categories: political interest

	SHP[a]	ESS[b]/MOSAiCH[b]	Selects CATI[c]/Web[c]
Question wording in English	Generally, how interested are you in politics, if 0 means "not at all interested" and 10 "very interested"?(endpoint labelled)	How interested would you say you are in politics –are you 3=Very interested, 2=Quite interested, 1= Hardly interested, 0=Not interested at all (fully labelled, with showcards).	Generally, how interested are you in politics, are you3=Very interested, 2=Rather interested, 1= Rather not interested 0=Not interested at all (fully labelled)

[a] SHP: Question in German: Wie stark interessieren Sie sich ganz allgemein für Politik, wenn 0 "gar nicht" und 10 "sehr stark" bedeutet?
Question in French: “De manière générale, quel intérêt portez-vous à la politique, si 0 signifie pas du tout intéressé et 10 très intéressé"?

[b] ESS and MOSAiCH: Question in German: Wie stark interessieren Sie sich für Politik? Würden Sie sagen, Sie sind... 3=Sehr interessiert 2=Ziemlich interessiert 1=Kaum interessiert 0=Überhaupt nicht interessiert.
Question in French: “Quel intérêt avez-vous pour la politique? Êtes-vous... 3= très intéressé 2=assez intéressé 1=peu intéressé 0=pas du tout intéressé.

[12] The Italian wording is not shown due to the small portion of the Italian-language sample.

[c] Selects CATI and web: Question in German: Wie interessiert sind Sie eigentlich im Allgemeinen an der Politik? Sind Sie da… 3=sehr interessiert, 2=eher interessiert, 1=eher nicht interessiert oder 0=überhaupt nicht interessiert.
Question in French: De manière générale, quel intérêt portez-vous à la politique? Etes-vous… 3=très intéressé, 2=plutôt intéressé, 1=plutôt pas intéressé ou 0=pas du tout intéressé par la politique.

The following table displays the mean values of political interest in the five surveys together with their standard deviations. Specifically, the raw or design weighted means are reported, as well as the means taking the nonresponse adjustment weights into account when such weights are available. We also report the additional effect on the standard error due to accounting for interviewer variance. Finally, sample sizes are displayed.

Table 3: Mean values and standard errors of political interest scores

Political interest	SHP	ESS	MOSAiCH	Selects CATI	Selects web
Sample size	6,297	1,211	999	2,332	433
Design weighted / raw	5.365 (.036)	1.711 (.024)	1.746 (.028)	1.878 (.020)	1.871 (.037)
Post-strat. weighted	5.255 (.041)	1.706 (.025)		1.878 (.020)	
Post-strat. weighted + I'wer cluster	5.255 (.050)	1.706 (.028)	1.746 (.041)	1.878 (.021)	

We can only compare the means between surveys which use the same scales.

In the SHP, there is a statistically insignificant difference between the raw and the weighted means on the 5% level (the 95% confidence intervals x_i±2se_i overlap). In Selects, this difference is almost zero. The difference between Selects CATI and Selects web and between ESS and MOSAiCH is also not significant

Using the post-stratification weights and taking into account interviewer clustering effects increases the standard errors by a small amount only.

Differences between the two Selects surveys and the face-to-face surveys are however significant with the latter having lower mean values. This could be due to the political topic of the Selects surveys that attracts more politically interested respondents. The selection bias could also explain the difference: Selects has an overrepresentation of highly educated people who are known to have a higher political interest (Dudley & Gitelson 2002, Hadjar & Schlapbach 2009, Verba et al. 2005). Due to social desirability effects from the presence of an interviewer, we would also expect to have higher score in a telephone or face-to-face survey compared to web. That could explain why the web component of Selects has a slightly lower mean.

An explanation for the lower values in ESS/MOSAiCH could be the different wording of the answer categories between Selects and ESS/MOSAiCH: “Quite interested” (ESS/MOSAiCH) tends to be closer to “Very interested” than “Rather interested” (Selects), and “Rather not interested” (Selects) further away from “Not interested at all” than “Hardly interested” (ESS/MOSAiCH). The translations in German and French (and Italian) reflect the same differences.

Interestingly, the SHP raw mean value of 5.365 (rescaled 1.610) is even lower than that of the face-to-face surveys. This is surprising as we would expect an opposite effect from panel attrition. The lower scores in the SHP may result from the different scale used (11-point scale vs 4-point scale, no middle point vs middle point, fully labelled vs endpoint labelled). In addition, the question wordings are different.

Satisfaction with democracy

Table 4 displays the different question designs in the different surveys. The SHP and ESS uses an 11-point scale including a midpoint with only the endpoints labelled, whilst the other surveys use a fully-labelled 4-point scale.

Table 4: Question wording and answer categories: satisfaction with democracy

	SHP[a]	ESS[b]	MOSAiCH[c]/Selects[c] CATI/Web
Question wording in English	Generally, what is your level of satisfaction with the way democracy works in our country if 0 means “not at all satisfied” and 10 “Completely satisfied” (endpoint labelled)	And on the whole, how satisfied are you with the way democracy works in Switzerland? 0 Extremely dissatisfied 10 Extremely satisfied (endpoint labelled, with showcards).	In General, are you 3=very satisfied, 2= rather satisfied, 1=rather not satisfied, 0=not at all satisfied with the way democracy works in our country? (fully labelled, with showcards (MOSAiCH))

[a] SHP: Question in German: Wie sind Sie im allgemeinen mit dem Funktionieren von der Demokratie in unserem Land zufrieden, wenn 0 "gar nicht zufrieden" und 10 "vollumfänglich zufrieden" bedeutet?
SHP: Question in French: Globalement, quel est votre degré de satisfaction du fonctionnement de la démocratie dans notre pays, si 0 signifie "pas du tout satisfait", et 10 "tout à fait satisfait"?

[b] ESS: Question in German: Und wie sehr sind Sie mit der Art und Weise, wie die Demokratie in der Schweiz funktioniert, zufrieden? 0=Äusserst unzufrieden … 10=Äusserst zufrieden
ESS: Question in French: Et dans l'ensemble, dans quelle mesure êtes-vous satisfait/e de la manière dont la démocratie fonctionne en Suisse? 0=Très insatisfait … 10=Très satisfait

[c] MOSAiCH and Selects CATI and web: Question in German: Sind Sie mit der Art und Weise, wie die Demokratie in der Schweiz funktioniert, alles in allem gesehen, 3=sehr zufrieden, 2=ziemlich zufrieden, 1=nicht sehr zufrieden oder 0=überhaupt nicht zufrieden?
MOSAiCH and Selects CATI and web: Question in French: Dans l’ensemble, êtes-vous 3=très satisfait, 2=plutôt satisfait, 1=plutôt pas satisfait 0=pas du tout satisfait du fonctionnement de la démocratie en Suisse?

Table 5 displays means with and without nonresponse adjustment weights and interviewer variance (for CATI and CAPI surveys) and the sample size for each survey. Once again, the effect of the post-stratification weights within surveys is insignificant and the use of post-stratification weights as well as taking the interviewer clustering into account has a limited impact on the standard error.

Table 5: Mean values and standard errors of satisfaction with democracy scores

Satisfaction with democracy	SHP[a]	ESS[a]	MOSAiCH [b]	Selects CATI[b]	Selects web[b]
Sample size	6,217	1,197	988	2,325	422
Design weighted / raw	6.171 (.024)	7.022 (.056)	2.019 (.020)	2.004 (.015)	1.820 (.031)
Post-strat. weighted	6.149 (.028)	7.008 (.058)		2.004 (.015)	
Post-strat. weighted + I'wer cluster	6.149 (.040)	7.008 (.070)	2.019 (.025)	2.004 (.014)	

The ESS shows a higher mean than the SHP, which is surprising as we expected a positive effect from attrition.[13] However, if we look at the sample composition, more men, full-time employed and married people as well as more German speakers participated in the ESS, see table 1. These factors are positively related to satisfaction with democracy (Halla et al. 2008, Stadelmann-Steffen & Vatter 2012). A lower portion of highly educated people in the ESS is probably not enough to offset this. We would however expect that the post-stratification weights would at least partially correct for this selection effect as there are based on (some) of these socio-demographical variables. The effect of the nonresponse adjustment goes in the right direction for ESS (smaller mean) but not significantly. The difference is even exaggerated by the lower weighted mean for the SHP. We can thus observe that nonresponse-adjustment weighting does not correct for the surveys specific selection effect. There could however be a lower social desirability effect or primacy effect in the SHP (CATI survey) then in ESS (CAPI survey).

Selects web score is lower than both MOSAiCH and Selects CATI. This could partly be explained by the respondent sample composition: less married, but at the same time more men, full-time working and high educated in the web survey. However it is more likely that this difference comes from mode-related measurement effects, the web mode being less influenced by social desirability.

Rescaling shows that the rescaled value of the SHP (6.171, rescaled 1.851) is closer to Selects web, while the rescaled ESS value (7.022, rescaled 2.107) is closer to MOSAiCH and Selects CATI. Comparability between SHP and ESS on one hand, and MOSAiCH, Selects CATI, and Selects web on the other is further limited by the different question wording and different number of categories.

Self-rated Health

The last variable we consider is self-reported health, and the following table displays the question and answer category wording for each variables. Note that the scale for health is negative.

[13] While participants of both years 2004 and 2011 had a mean of 5.95 in 2004, attritors in 2011 had a mean of only 5.78 in 2004.

Table 6:: Question wording and answer categories: negative health

	SHP[a]	ESS[b]	MOSAiCH[c]	SFLS[d]
Question wording in English	How do you feel right now? : 1=very well, 2=well, 3=so, so (average), 4=not very well, 5=not well at all (fully labelled).	How is your health in general? Would you say it is ...1=Very good, 2-Good, 3=Fair, 4=Bad, 5=or, Very bad? (fully labelled, with showcards).	In general, would you say your health is ... 1=excellent, 2=very well, 3=well, 4=acceptable, 5=bad (fully labelled, with showcards).	How is your health in general? Is it 1=Very good, 2=good, 3=Fairly good, 4=Bad, 5=Very bad (fully labelled).

[a] SHP: Question in German: Wie geht es Ihnen zur Zeit gesundheitlich? 1=sehr gut, 2=gut, 3=es geht so/mittelmässig, 4=schlecht, 5=sehr schlecht
SHP: Question in French: Comment est votre santé en général? 1=très bonne, 2= bonne, 3=moyenne, 4=mauvaise, 5= très mauvaise

[b] ESS: Question in German: Wie würden Sie Ihren allgemeinen Gesundheitszustand einstufen? Halten Sie Ihren Gesundheitszustand für ... 1=sehr gut, 2=gut, 3=mittelmässig, 4=schlecht, 5=sehr schlecht?
ESS: Question in French: Quel est votre état de santé en général? Diriez-vous qu'il est ... 1=très bon, 2= bon, 3=passable, 4=mauvais, 5= très mauvais

[c] MOSAiCH: Question in German: Wie würden Sie Ihren allgemeinen Gesundheitszustand bezeichnen? 1=Ausgezeichnet 2= sehr gut 3=gut 4=Akzeptabel 5=Schlecht
MOSAiCH: Question in French: Dans l'ensemble, vous diriez que votre santé est … 1=Excellente, 2= Très bonne, 3=Bonne, 4=Correcte, 5=Mauvaise

[d] SLFS: Question in German: Wie ist Ihr Gesundheitszustand im Allgemeinen? Ist er … 1=sehr gut, 2=gut, 3=mittelmässig, 4=schlecht, 5=sehr schlecht
SLFS: Question in French: Comment est votre état de santé en général? Est-il … 1=très bon, 2= bon, 3=assez bon, 4=mauvais, 5= très mauvais

Table 7 displays the mean values of negative health which is measured in four of the five surveys considered. Again, we list means without and with accounting for nonresponse adjustment weights, and finally the effect of interviewer variance taken into account.

Table 7: Mean values and standard errors of negative health scores

Negative health	SHP[a]	ESS[b]	MOSAiCH[c]	SLFS[d]
Sample size	6,299	1,212	1,000	46,468
Design weighted raw	1.987 (.008)	1.880 (.022)	2.503 (.027)	1.833 (.004)
Post-strat. weighted	2.002 (.009)	1.893 (.023)		1.805 (.004)
Post-strat. weighted + I’wer cluster.	2.002 (.010)	1.893 (.027)	2.503 (.033)	1.805 (.008)

Again, the means and standard errors do almost not change once post-stratification weights and/or interviewer clustering are taken into account. An exception to the latter is the SLFS, suggesting a strong interviewer effect.

The differences between the surveys are significant. The highest score is obtained in MOSAiCH (worse health), which probably results from the asymmetric (positively skewed) scale. SHP has the second highest score for negative health. This is again surprising since we expected a positive selection effect from attrition. Nevertheless, more women, less full-time workers (Mansyur et al. 2008), and less highly educated (Bobak et al. 2000) people suggest a composition effect in the SHP. Note also, that the weighted mean leads to an even worse self-rated health. In addition, the negative tendency in the SHP could be a seasonal effect (SHP is fielded between September and January), because the question is – unlike in the other surveys – related to the time of the survey.

ESS and SFLS have almost exactly the same wording and the same answer categories but the ESS results display worse health. This result is probably a sample composition effect since the SLFS has less men, but more old or married people (Fylkesnes & Førde 1991) and more French and Italian speakers. Again, the post-stratification weights only exacerbate this difference, leading to an even worse health in ESS and better health in SLFS. Also the mode of data collection could be a reason, with people being more negative in a face-to-face interview. In addition, coverage tends to be better in face-to-face surveys with a higher probability to include "bad" risks. It could also be hypothesized that attrition in the SLFS causes this difference. However there is only a slightly better health of those who are in their third wave, compared to those in their first wave.

4. Conclusion and discussion

The goal of this research was to compare three often used items in social research (political interest, satisfaction with democracy, and subjective health) across six different Swiss surveys from 2010/11 (Swiss Household Panel (SHP), MOSAiCH, Swiss,ESS, Swiss Labor Force Survey (SLFS), Selects CATI, and Selects web), considering sample selection adjustment weights and interviewer effects or not. We aimed to assess the robustness of univariate means against different factors influencing the total survey error. In this paper, we mainly considered coverage and nonresponse bias, measurement error due to the presence or not of an interviewer, question wording, the number and the labeling of categories, and the data collection mode.

We found differences between the surveys on both socio-demographic variables and substantive survey items.

Selection bias can be due to many factors such as the topic of the survey (Selects) or the survey mode with differential under-coverage or different nonresponse mechanisms. Finally attrition in longitudinal studies also has an influence on the composition of the respondents.

Concerning substantive items within surveys, our first and clearest finding is that – although differences between surveys are sometimes substantial – raw (or only design weighted) mean values of our three items are not significantly different from those using post-stratification weights and that the standard errors do mostly not increase much when interviewer variances are taken into account. This shows that the weights, which are generally based of socio-demographic variables, do not manage to correct bias arising from nonresponse and under-coverage.

Second, we find that across the surveys considered, the means are similar within survey blocks and different between these blocks. The blocks (separated by ",") are SHP/ESS, MOSAiCH/Selects CATI, Selects web for political interest; SHP, Selects web/Selects CATI/ESS, MOSAiCH for satisfaction with democracy, and MOSAiCH/SHP/ESS, SLFS for self-rated health. Especially the differences between longitudinal (SHP) and cross-sectional surveys (ESS) seem to be large. Although we do not have an experimental set-up and the error-source causing differences between surveys cannot really be disentangled, we nevertheless manage to relate some selection and measurement effects to different causes. First, we identify selection effects due to topic, undercoverage, and nonresponse. Second, we found measurement effects due to different modes and mainly different question designs as possible causes for the observed differences.

The biggest differences are found for 'self-rated health' where answer category differences are the largest. This highlights the importance of the question design. Coverage and nonresponse bias can also have an effect that is difficult to capture based on only socio-demographic variables. We saw that in most cases socio-demographic differences only explain a very small part of the non-observation bias, and attrition makes it even more complex. Post-stratification weights do not decrease the differences between surveys. Also survey modes with different measurement effects (such as less socially desired answers in the absence of an interviewer) are plausible, in particular for Selects web. Our attempts to explain all differences however failed. Unmeasured factors may influence the outcome, like the special situational context the interview took place or the place of the question in the questionnaire/interview.

We believe that our aim to highlight the limitations of basing conclusions on a single survey and relying on the weights usually delivered with the survey data is met to some extent. Studies should always be cross-validated with existing research and error factors taken into account. Different outcomes do no always mean wrong outcomes and can often be explained by differences in survey designs.

To summarize, great care should be taken when interpreting survey results. Researchers must not forget that survey and instrument characteristics can largely influence the results and need to take this into account when drawing conclusions. We hope that this paper is able to raise awareness among substantive researchers of the possible impact of survey errors on their results. Finally, we only considered means and their standard errors here. Model estimation differences between surveys and influences of different survey designs should be considered in further research.

References

AAPOR (2011), Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. American Association for Public Opinion Research, 7th edition.

Béland, Y. & St-Pierre, M. (2008), Mode Effects in the Canadian Community Health Survey: A comparison of CATI and CAPI, In J.M. Lepkowski, C; Tucker, J. M; Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link & R. L. Sangster (Eds): Advances in Telephone Survey Methodology, New York, Wiley & sons, Inc.

Bethell, C. D., Read, D., Blumberg, S.J. & Newachek, P.W. (2008), What is the Prevalence of Children with special Health Care Needs? Toward an Understanding of Variations in Findings and Methods across three National Surveys, Matern Child Health J 12, 1-14.

Bethlehem, J. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information, In: Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little, R. J .A.(eds). Survey Nonresponse, New York:Wiley.

Bethlehem, J. (2010), Selection Bias in Web Survey, International Statistical Review, 78 (2), 161–188 doi:10.1111/j.1751-5823.2010.00112.x.

Bethlehem, J., Cobben, F. & Schouten, B.(2011), Handbook of Nonresponse in Household Surveys, New Jersey: Wiley.

Biemer, P. (2010), Total Survey Error: Design, Implementation, and Evaluation, Public Opinion Quarterly 74 (5), 817-848.

Biemer, P., Groves, R. M., Lyberg L. E., Nancy A. M., Sudman, S. (2004), Measurement Errors in Survey, New Jersey: Wiley.

Birkett, N. J. (1986). Selecting the number of response categories for a Likert-type scale. Proceedings of the American Statistical Association, 488–492.

Bobak, M., Pikhart, H., Hertzman, C., Rose, R. & Marmot, M. (2000), Socioeconomic factors, perceived control and self-reported health in Russia. A cross-sectional survey, Social Science & Medicine 47(2), 269 – 279.

Bopp M., Braun J., Gutzwiller F., & Faeh D.,(2012) Health Risk or Resource? Gradual and independent Association between Self-Rated Health and Mortality Persists Over 30 Years, PLoS ONE 7(2): e30795. doi:10.1371/journal.pone.0030795.

Braunsberger, K., Wybenga, H. & Gates, R. (2007), A comparison of reliability between telephone and web-based survey, Journal of Business Research, 60(7) 758–764.

Brick, J., R. McGuinness, S. Lapham, M. Cahalan, D. Owends & Gray, L. (1995), Interviewer variance in two telephone surveys. Proceedings of the Section on Survey Research Methods (pp. 447–452). Orlando, FL: American Statistical Association.

Brown, G., Micklewright, J., Schnepf S.V. & Waldma, R. (2005), Cross-National Surveys of Learning Achievement: How Robust are the Findings? IZA Discussion Papers, No. 1652.

Burns, N., Schlozman, K. L. & Verba, S. (1997), The Public Consequences of Private Inequality: Family Life and Citizen Participation. The American Political Science Review 91(2), 373-389.

Callegaro, M. (2013), Web coverage in the UK and its potential impact on general population web surveys, presented at the conference "Web surveys for the general population: How, why and when?"

Dahlberg, S. & Persson, M. (2013), Different Surveys, Different Results? A Comparison of two Surveys on the 2009 European Parliamentary Election, West European Politics, DOI: 10.1080/01402382.2013.814961.

Dalton, R. J. (2006), The two faces of citizenship. Democracy & Society, 3, 21-29.

De Leeuw, E. & van de Zouwen, J. (1998), Data Quality in Telephone and Face-to-Face Surevys: a meta-analysis. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T.

Massey, W.L. Nicholls & J. Waksberg (EDs), Telephone Survey Methodology (pp 283-299). New York: Wiley & Sons, Inc.

Deville, J.-C. & Särndal, C.-E. (1992), Calibration estimators in survey sampling. Journal of the American Statistical Association 87(418), 376–382.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating mixed standard scale formats. Journal of Applied Psychology 65, 147–154.

Dudley, R. L. & Gitelson, A. R. (2002), Political literacy, civic education, and civic engagement: A return to political socialization? Applied Developmental Science, 6(4), 175-182.

Duffy, B., Smith, K., Terhanian, G. & Bremer, J. (2005), Comparing Data from Online and Face-to-Face Surveys. International Journal of Market Research, 47(6), 615-639.

ESS, ESS 2010 – Documentation report third edition, Retrieved from: http://www.europeansocialsurvey.org/docs/round5/survey/ESS5_data_documentation_report_e03.pdf.

Fylkesnes, K. & Førde, O.H. (1991), The Tromsø Study: predictors of self-evaluated health--has society adopted the expanded health concept? Soc Sci Med. 32(2), 141–146.

Fitzgerald, J., Gottschalk, P. & Moffitt, R. (1998), An Analysis of Sample Attrition in Panel Data. The Michigan Panel Study of Income Dynamics, Journal of Human Resources 33(2), 251-299.

Garriguet, D. & Latouche, M. (2004), Weighting of the Swiss Household Panel Wave 5 Detailed Description.

Gerrits, M. H., Van den Oord E. J. C. G. & Voogd R. (2001), An Evaluation of Nonresponse Bias in Peer, Self, and Teacher Ratings of Children’s Psychosocial Adjustment, J. Child Psychol. Psychiat. 42 (5), 593–602.

Givon, M. M., & Shapira, Z. (1984). Response to rating scales: A theoretical model and its application to the number of categories problem. *Journal of Marketing Research* 21, 410–419.

Goudy, W.J. (1976), Nonresponse Effects on Relationships Between Variables, Public Opinion Quarterly 40(3), 360-369.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery – How many scales and response categories to use? Journal of Marketing 34, 33–39.

Groves, R.M. & Couper M.P. (1998), Nonresponse in Household Interview Surveys, New York: Wiley.

Groves, R. & Heeringa, S. (2006), Responsive Design for Household Surveys: Tools for actively controlling Survey Errors and Costs. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169 (3), 439-457.

Groves, R.M. & Lyberg, L. (2010), Total Survey Error: Past, Present and Future, Public Opinion Quarterly 74 (5), 849 -879.

Hadjar, A. & Schlapbach, F. (2009), Educational Expansion and Interest in Politics in Temporal and Cross-cultural Perspective: A Comparison of West Germany and Switzerland, European Sociological Review 25 (3), 271-286.

Halla, M., Schneider, G. F. & Wagner, F. A. (2008), Satisfaction with Democracy and Collective Action Problem: the Case of the Environment, IZA Discussion Paper, Discussion Paper No. 3613.

Heath, A., Martin, J.Spreckelsen, T.,Cross national comparability of survey attitudes measures, Journal of Public Opinion Research, 21 (3).

Heckman, J.J. (1976), The common structure of statistical models of truncation, sample selection and limited dependent variable, Ann. Econ. Soc. Measure 47, 153-161.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone vs. 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.

Jäckle, A., Roberts, C. & 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. (2010), Assessing the effect of data Collection Mode, International Statistical Review 78 (1), 3- 20.

Kalton, G. & Flores-Cervantes, I. (2003), Weighting Methods, Journal of Official Statistics 19 (1), 81-97.

Kass, G. V.(1980), An exploratory technique for investigating large quantities of categorical data, Journal of the Royal Statistical Society Series C 29(2), 119-127.

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

Krieger N., Kosheleva A., Waterman P. D., Chen J. T., & Koenen K. (2011) Racial Discrimination, Psychological Distress, and Self-Rated Health Among US-Born and Foreign-Born Black Americans. American Journal of Public Health 101(9), 1704-1713.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. American Journal of Political Science 37, 941–964.

Krosnick, J.A., & Presser, S. (2010) 'Question and Questionnaire Design' in: Handbook of Survey Research. 2nd Edn. Emerald. pp.263-313.

Lehmann, D. R., & Hulbert, J. (1972). Are three-point scales always good enough? Journal of Marketing Research 9, 444–446.

Lipps, O., & Kissau. K. (2012). Nonresponse in an Individual Register Sample Telephone Survey in Lucerne (Switzerland). In Telephone Surveys in Europe: Research and practice, M. Häder, S. Häder and M. Kühne (eds.). Springer, 187-208.

Lipps, O. & Pekari, N. (2013), Mode and incentive effects in an individual register frame based Swiss election study. FORS working paper 3-2013, Lausanne

Lipps, O., Pekari, N. & Roberts, C. (2013), Coverage and nonresponse errors in an individual register frame based Swiss telephone election study. FORS working paper 2-2013, Lausanne.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology 60, 10–13.

Lynn, P. (2013), Issues of Coverage and Sampling in Web Surveys for the General Population: An overview. Synthesis paper for NCRM Web Survey Network opening conference.

Luiten, A. & Schouten, B. (2013), Tailored fieldwork design to increase representative household survey response rate: an experiment in the Survey of Consumer Satisfaction, J.R. Statist. Soc A 176(1), 169-189.

Lutz, G., Borrat-Besson, C., Ernst –Stähli, M., Werner, B. (2012), Wer ist denn noch am Festnetzanschluss erreichbar? Jahrbuch 2012: Markt und Sozialforschung, 24- 26.

Mansyur, C., Amick, B. C., Harrist, R. C. & Franzini, L. (2008), Social capital, income inequality, and self-rated health in 45 countries, Social Science & Medicine 66, 43–56.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. Journal of Educational Measurement 11, 49–53.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research* 10 , 316–318.

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations, Journal of Marketing Research 15, 304–308.

Merkouris, T. (2001), Cross-sectional estimation in multiple-panel household surveys. Survey Methodology, 27(2), 171–181.

Mohorko, A., de Leeuw, E. & Hox, J. (2013), Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time. Survey Methods: Insights from the Field. Retrieved from http://surveyinsights.org/?p=828.

Norris, F. H. (1987), Effects of Attrition on Relationships Between Variables in Surveys of Older Adults, Journal of Gerontology 42(6), 597-605.

Nagelhout, G.E., Willemsen, M.C., Thompson, M.E., Fong G.T., van den Putte, B. & de Vrie, H. (2010), Is web interviewing a good alternative to telephone interviewing? Findings from the International Tobacco Control (ITC) Netherlands Survey, *BMC Public Health* 10:351 doi:10.1186/1471-2458-10-351.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the American Association for Public Opinion Research annual meeting, St. Petersburg, FL

Peytchev, A., Riley,, Rosen, J. and Lindblad, M. (2010), Reduction of nonresponse bias in survey through case prioritization, *Survey Research Methods*, 4 (1)

Pollien A. & Joye, D. (2014), Patterns of Contact Attempts in Surveys. In P. Blanchard, F. Bühlmann & J.-A. Gauthier (Eds.), Advances in Sequence Analysis: Theory, Method, Applications, chapter 15. London: Springer (in press).

Ramsay, J. O. (1973). The effect of number categories in rating scales on precision of estimation of scale values. Psychometrika 38, 513–532.

Roster, C.A., Rogers, R.D., Albaum, G. & Klein, D. (2004), A comparison of response characteristics from web and telephone surveys, Int J Mark Res 46 (3) , 359–374.

Sakshaug, J.W., Yang, T. & Tourangeau R. (2010), Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and Non-sensitive Items, Public Opinion Quarterly 74(5), 907-933.

Sapiro, V. (1982), Private Costs of Public Commitments or Public Costs of Private Commitments? Family Roles versus Political Ambition. American Journal of Political Science 26(2), 265-279.

Saris, W.E. & Gallhofer, I.N. (2007) Estimation of the effect of measurement characteristics on the quality of survey questions. Survey Research Methods 1, 31-46.

Scherpenzeel, A. (2001), Mode effects in panel surveys: A comparison of CAPI and CATI. Actualités OFS, October 2001, order number: 448-0100, Neuchâtel, Swiss Federal Statistical Office.

Schneiderat, G. & Schlinzig, T. (2012), Mobile-and Landline-Onlys in Dual-Frame-Approaches: Effects on Sample Quality. In (S. Häder, M. Häder & M. Kühne (Eds.), Telephone Surveys in Europe. Springer Berlin Heidelberg, 121-143.

SFSO (Swiss Federal Statistical Office) (2012), Die Schweizerische Arbeitskräfteerhebung ab 2010. Konzepte – Methodische Grundlagen – Praktische Ausführung. Neuchâtel November 2012.

Solon G., Haider S. J. & Wooldridge J. (2013), What are we weighting for? NBER Working Paper No. 18859, Retrieved from: http://www.nber.org/papers/w18859.

Stadelmann-Steffen, I. & Vatter, A. (2012), Does Satisfaction with Democracy Really Increase Happiness? Direct Democracy and Individual Satisfaction in Switzerland, Political Behaviour 34, 535-559.

Stoop, I. (2005), The Hunt for the Last Respondent. Nonresponse In Sample Surveys, Ph.D. Thesis, University of Utrecht.

Stoop, I, Billiet, J., Koch A. & Fitzerald R. (2010), Improving Survey response: Lessons learned from the European Social Survey, Wile& Sons: West Sussex.

Vandenplas, C., Joye, D., Ernst Staehli, M. & Pollien, A. (2014), Usability of non-response survey in the Swiss context, in preparation.

Vandenplas, C., Roberts, C., Joye D. & Ernst Staehli, M. (2013), A Results of a Mixed Mode Experiment conducted in Switzerland, paper presented at the Priority program for Survey Methodology conference, September 12-13.

Vannieuwenhuyze, J. & Loosveldt, G. (2014), Evaluating Mode Effects in Mixed-Mode Survey Data using covariate adjustment models. Journal of Official Statistics 30, (1), 1–21.

Verba, S., Burns, N. & Schlozman, K. L. (1997), Knowing and caring about politics: Gender and political engagement. The Journal of Politics 59(4), 1051-1072.

Voorpostel, M. & Coffé, H. (2012), Transitions in partnership and parental status, gender, and political and civic participation. European Sociological Review 28(1), 28-42.

Voorpostel, M., Tillmann, R, Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., Schmid, F., Rothenbühler, M. & Wernli, B. (2013), Swiss Household Panel Userguide (1999-2012), Wave 14, November 2013. Lausanne: FORS.

Winship, C. & Rabil, L. (1994), Sampling Weights and Regression Analysis, Sociological Methods & Research 23(2), 230-257.

Wagner, J. R. (2008), Adaptive Survey Design to Reduce Nonresponse Bias, PhD dissertation, Retrieve from: http://deepblue.lib.umich.edu/bitstream/handle/2027.42/60831/jameswag_1.pdf?sequence=1.

Ye, C., Fulton, J. & Tourangeau R. (2011), More Positive or More Extreme? A meta-analysis of mode difference in response choice, Public Opinion Quarterly 75(2), 349-365.