Peter Grand

Assessing Mode Effects in SC Approaches – Comparing F2F and Push-to-Web

1 Introduction

Survey research has always been under pressure, mainly because of trying to reach two mutually exclusive goals simultaneously – the ideals of sampling theory and cost-effectiveness. In between these two general requirements is the choice of the most efficient mode (in connection with the sampling design) to reach these goals, and the modes available always depended on technological innovation and certain network effects, i.e., using telephone interviews makes sense only when a sufficient number of persons or households are part of this network. Certainly, the access to, e.g., telephone networks has had to reach a certain tipping point to make the register of telephone numbers a useful and practicable tool for sampling purposes, thus in Edison's times it would have made no sense to use the telephone register (if something like that already existed) as a sampling frame. Similarly, it now makes less and less sense to use the register of landline telephone numbers to draw a representative sample. The dominant survey mode in the second half of the 20th century was interviewer-driven either as a face-to-face interview (CAPI/CAWI) or in form of a telephone interview (CATI). The golden standard of survey research from a methodological point of view in these times has been until recently/today computer-assisted in-person interviews. This data collection mode has become ever more costly while response rates are continuously decreasing.

Consequently, there is always some dispute about the appropriate mode and the pros and cons of choosing a specific mode and the possible consequences in terms of, e.g., representativeness, non-response bias, response rates, unit- & item-nonresponse, etc. – the so-called mode-effects. Those questions also arise when one has to change the data collection mode because of one reason or another. While continuously decreasing response rates and steadily increasing costs in face-to-face (f2f)-survey research may lead to a mode change in the medium term, the COVID-19-pandemic forced some survey endeavours to change their mode immediately because of COVID-19 restrictions and social distancing rules. The COVID-19 pandemic made problems connected with f2f-interviewing not only more visible and more severe but also motivated some research projects to experiment with different modes and generally with strategies of conducting surveys that avoid f2f interviews.

The European Social Survey European Research Infrastructure Consortium (ESS ERIC) has also been hit hard by the COVID-19 pandemic and its consequences and confronted with experimenting with alternatives to the hitherto required mode of f2f-interviews. Thus, ESS ERIC conducted experiments with so-called self-completion approaches, namely a "push-to-web"-strategy in three countries: Hungary, Austria and Serbia. This contribution will focus on the Austrian part of this push-to-web experiment and compare the data to the Austrian ESS round 9 data and administrative data, mainly with the relevant EU-SILC data, to assess possible mode effects between the push-to-web experiment and ESS round 9 data.

2 Survey Research at a crossroads – again?

The only allowed mode of data collection within ESS fieldwork until ESS Round 10 has been a f2f- interview either as a paper-assisted personal interview (PAPI, not possible anymore since ESS round 8) or as a computer-assisted personal interview (CAPI). ESS round 8 and round 9 in Austria have been conducted using CAPI. Although in both rounds approximately a response rate of 52 % could be achieved it became more and more difficult in terms of refusal conversion, i.e., convincing a respondent to conduct an interview although this individual has refused to do so at least once.

In the turmoil of the COVID-19 pandemic starting in late 2019 and the following worldwide restrictions to combat the Corona disease in early 2020, it soon became clear that following the traditional patterns of data collection in an international survey research project was becoming highly problematic and difficult. This insight is not solely attributable to the COVID-19 pandemic but has only been accelerated. ESS ERIC already recognized for several years that their workhorse for data collection has become increasingly costly and national teams have had to put even more effort into reaching acceptable response rates in terms of interviewer training, refusal conversion, fieldwork monitoring, etc.

Big-scale international survey projects will also face another obstacle if they want to go ahead with personal interviews: During the COVID-19 pandemic the interviewer staff of many survey agencies shrank and it will be very difficult to return to pre-COVID19 levels if not impossible. For example, Switzerland has had problems finding interviewers during ESS round 10 and in Austria interviewer staff size has been cut in half at some survey agencies.

3 The characteristics of the "self-completion"-approach

Consequently, ESS ERIC conducted a "push-to-web"-experiment in three ESS countries – Austria, Hungary, and Serbia at the end of 2020. This experiment aimed to test if a self-completion approach produces response rates and data quality comparable to the ESS golden standard f2f CAPI interviews. The design of the experiment followed a "web first self-completion"-strategy, i.e., each respondent received an invitation to interview via mail. This aviso letter also contained a link to a landing page designed by the respective national team and a serial number linked to one specific respondent. The landing page forwarded the respondent to a respective Qualtrics page where the respondents had to fill in their serial number to get access to the interview. In Austria, we sampled from a household register (sampling frame), the same register as we used in ESS8–ESS10, and we chose the "correct" respondent using the next-birthday-method. Generally, we used the same sample design as in the conventional ESS waves except for the clustering step since we had no restrictions regarding travel expenses.

Along with the aviso letter we also sent a 5 Euro voucher as an unconditional incentive and motivation to conduct the interview and offered a 10 Euro voucher as a conditional incentive for the complete interview. We also "allowed" a maximum of two persons to conduct the interview in case the first interview has not been finished by the "correct" respondent.

In case we received no valid interview from a household we sent a reminder two weeks after the invitation letter and a second reminder two weeks thereafter. The second reminder also contained a paper questionnaire – including a pre-paid/stamped return envelope to send the paper questionnaire back free of charge – leaving the respondent with the choice between interviewing via web or via the paper questionnaire.

We also used neutral envelopes without any logos which only differed in colour and size between the different mailing waves.[1] The aim was that our letters should not be mistaken with common advertisements and that it should look as neutral as possible to increase the possibility of it being opened.

The questionnaire consisted of around 80 questions which led to a median interview length of 20 minutes.

1 Dillman, Don A./Smyth, Joelene D./Christian, Leah Melani: Internet, Phone, Mail, and Mixed-Mode Surveys: the Tailored Design Method, 4th edition, Hoboken 2014, p. 384 ff.)

4 Mode effects and their consequences

The chosen survey mode can affect responses in multiple ways and in different directions. In an interviewer-driven setting not only can survey responses be influenced by the mere presence of an interviewer but also by characteristics of the specific interviewer because of dynamic interactions between respondent and interviewer.[2] F2F interviews may lead to less item non-response because of the presence of an interviewer and the fact that interviewers may probe "don't know"-responses. An interviewer can also motivate a respondent to increase their mental effort in answering survey questions. In a self-completion environment respondents may answer questions using less energy and/or effort (also called 'satisficing') because they are not only concentrating on the survey but also doing other things and an encouraging interviewer is missing. Self-completion surveys may also lead to more nondifferentiation, i.e., respondents don't use the range of options offered but stick to specific options, e.g., the middle one. Krosnick[3] classifies nondifferentiation as satisficing, and describes satisficing in a dramatic way: "Respondents are likely to satisfy whatever desires motivate them to participate just a short way into an interview, and they are likely to become increasingly fatigued, disinterested, impatient, and distracted as the interview progresses. […] Their motivation to work hard has evaporated, and the cognitive costs of hard work have become increasingly burdensome. Nonetheless, the interviewer continues to ask a seemingly unending stream of questions and to carefully record responses, which suggests that the interviewer expects the respondent to devote the effort necessary to generate high-quality responses".[4]

Social interactions also entail a phenomenon called 'social desirability' which potentially biases survey responses. The respondent does not answer 'honestly' but anticipates what the interviewer may expect and what is socially desirable, thus the answer is biased. The larger the social distance is between two or more individuals the less the consequences of 'social desirability' may be, thus the social distance in a self-comple-

2 Cf. Ansolabehere, Stephen/Schaffner, BrianF.: Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison, in: Politcal Analysis 22 (2014), pp. 285–303; Bowyer, Benjamin T./Rogowski, Jon C.: Mode Matters. Evaluating Response Comparability in a Mixed-Mode Survey, in: Political Science Research and Methods 5 (2017), pp. 295–313; Heerwegh, Dirk: Mode Differences Between Face-to-Face and Web Surveys. An Experimental Investigation of Data Quality and Social Desirability Effects, in: International Journal of Public Opinion Research 21 (2009), pp. 111–121; Heerwegh, Dirk/Loosveldt, Geert: Face-To-Face versus Web Surveying in a High-Internet-Coverage Population. Differences in Response Quality, in: Public Opinion Quarterly 72 (2008), pp. 836–846.

3 Krosnick, JonA.: Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys, in: Applied Cognitive Psychology 5 (1991), pp. 213–236.

4 Krosnick 1991, p. 214.

tion mode is larger compared to a f2f-situation.[5] Empirical testing of this 'candour-hypothesis' yields mixed results.[6] Generally, a bias stemming from social desirability induced behaviour only occurs during questions touching sensitive issues, e.g. attitudes towards racism or sexual behaviour.

5 Comparing the push-to-web experiment and the ESS

Comparing different surveys and assessing possible mode effects is a tricky endeavour.[7] We will start with a comparison to a validated baseline, which in our case refers, first, only to socioeconomic characteristics stemming from the official Austrian Statistics, and second, offers only rough evidence of how representative the sample in question really is, i.e., using this strategy we can only speak about the composition of the sample and not about the validity of the collected data itself, e.g. the amount of bias affecting the mean of a specific question. In a second step we will evaluate the means and standard errors of selected variables between web and paper and compare them to regular ESS round data. Finally, we analyse the distributions of those same variables, again web to paper and push-to-web to ESS round 9 data. This method analyses two empirical cumulative distribution functions (ECDF) and assesses the equality of these two distributions for each value options.[8]

5.1 Comparisons against validated official statistics

The first analysis comprises several comparisons of socio-economic characteristics against official data from Statistics Austria. Table 1 shows the figures for sex broken down by mode and data source. The first two columns refer to the results regarding those respondents who conducted the interview via paper or web respectively. The third column "PtW" comprises all respondents, i.e. those who answered the questionnaire via web or paper, and the fourth column shows the respective figures for ESS9.

5 Cf. Holbrook, Allyson L./Green, Melanie C./Krosnick, Jon A.: Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias, in: Public Opinion Quarterly 67 (2003), pp. 79–125.

6 Cf. Gnambs, Timo/Kaspar, Kai: Socially Desirable Responding in Web-Based Questionnaires. A Meta-Analytic Review of the Candor Hypothesis, in: Assessment 24 (2017), pp. 746–762.

7 Ansolabehere/Schaffner 2014; Bowyer/Rogowski 2017; Heerwegh 2009.

8 Goldman, Matt/Kaplan, David M.: Comparing Distributions by Multiple Testing across Quantiles or CDF Values, in: Journal of Econometrics 206 (2018), pp. 143–166; Kaplan, David M.: distcomp: Comparing Distributions, in: Stata Journal 19 (2019), pp. 832–848.

Table 1: Sample composition by sex (weighted)

Sex	Paper (n=99)	Web (n=226)	PtW (n=325)	ESS R9 (n=2,499)	Population statistics
	%	%	%	%	%
Male	40.8	52.6	49.3	48.6	48.9
Female	59.2	47.4	50.7	51.4	51.1

The share of men and women answering the push-to-web survey using the paper questionnaire are evenly distributed while in the group of those respondents who choose to use the web, there are slightly more men than women (52.6 % to 47.4 %). There are some differences between the push-to-web experiment compared to the ESS9 in terms of sex, however, they amount to less than 0.7 percentage points. Compared to the official population statistics the differences are even smaller with less than 0.5 percentage points.

Table 2: Sample composition by Age (weighted)

Age	Paper (n=103)	Web (n=229)	PtW (n=332)	ESS R9 (n=2,499)	Population statistics
	%	%	%	%	%
15–29	4.0	19.9	15.3	20.6	19.1
30–49	11.2	31.8	25.9	32.2	32.8
50–64	28.9	24.8	26.0	24.8	24.5
65+	55.9	23.5	32.8	22.4	23.8

We find stark differences when breaking down the data by age between the two self-completion modes – although to some extent this is to be expected. The older a respondent the more likely they will use the paper option to complete the survey, thus only 4 % of the 15–29y old choose paper compared to 20 % who conducted the survey online. The difference becomes somewhat smaller in the next age group, the 30–49y old, 32 % of them used the web compared to roughly one third using the paper questionnaire sent with the final reminder. With respect to the two older age groups the picture changes, whereas the percentages of the 50–64y groups are of fairly similar size, in the case of the 65+y old 56 % choose to conduct the survey using the paper questionnaire and 24 % the web. Also, regressing mode choice on age reveals that age is a strong predictor for the mode whereas sex is not. Thus, there are to some extent selection effects, i.e., the

choice of the mode does not happen by chance only. It is also important to mention that after testing the age groups distribution for equality over mode as well as over the data collection process the distributions are statistically the same.

Table 3: Sample composition by citizenship (weighted)

Citizenship	Paper (n=101)	Web (n=225)	PtW (n=326)	ESS R9 (n=2,498)	Population statistics
	%	%	%	%	%
Yes	89.8	94.5	91.1	90.9	91.9
No	10.3	5.5	8.9	9.1	8.9

A further possibility is to analyse the data alongside citizenship, i.e., if the respondent is an Austrian citizen regardless of her country of birth. Table 3 shows the respective figureS. 90 % of those who choose paper against the web option indicated their Austrian citizenship compared to 10 % non-Austrian citizens. Among the online respondents 95 % have the Austrian citizenship whereas 5 % denied the citizenship. The ratio between paper and web, roughly 2:1, should not be overexaggerated because the sample sizes in the respective cells are very low. The figures for the overall push-to-web data and the ESS9 as well as the official data from the population statistics are similar, thus the data from the push-to-web experiment closely resembles the data from ESS round 9 and from Statistics Austria.

Summarizing the comparisons to official population statistics results in a very positive picture regarding the representativity of the self-completion data. Even the age distribution – often problematic when it comes to online surveys – is statistically not different from the distributions generated by the face-to-face data collection mode used in ESS round 9.

In the next step we will have a closer look at the means and standard errors of selected variables and compare them against the ESS round 9 data. This change in focus concentrates hereafter on the answering behaviour of the respondents and not on the sample composition.

5.2 Comparisons of means and standard deviations

The analysis of central tendencies and standard errors and thus confidence intervals offers a further possibility to assess possible mode effects. In the following we are restricted to variables also available in the main questionnaire of the ESS round 9. Although most of these variables are categorical variables and calculating the mean

usually makes no sense because it is generally not interpretable, for comparing distributions this strategy offers intuitive and easily interpretable results because we do not assess the central tendency of the question itself with its manifest meaning but rather compare several distributions against each other. The first set of variables touches different dimensions of social connectedness, included are the following questions:

- (ppltrst) Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? (11-point scale ranging from "You can't be too careful" to "Most people can be trusted")
- (pplfair) Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair? (11-point scale ranging from "Most people would try to take advantage of me" to "Most people would try to be fair")
- (pplhlp) Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves? (11-point scale ranging from "People mostly look out for themselves" to "People mostly try to be helpful")

Table 4: Means and SE – social connectedness (weighted)

Variable	Paper (n=99)		Web (n=226)		PtW (n=325)		ESS R9 (n=2,486)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
ppltrst	4.58	0.28	5.10	0.18	4.97	0.14	5.55	0.05
pplfair	6.07	0.23	6.15	0.16	6.12	0.12	6.29	0.04
pplhlp	5.58	0.25	5.69	0.16	5.65	0.13	5.69	0.05

Table 4 shows the means and standard errors for these variables tapping social connectedness. These three variables with their 11-point scale can easily be treated as quasi-metric variables and thus using the mean and standard errors poses no problems.

Figure 1 visualizes the data from Table 4 thereby showing the mean as point estimation and two confidence intervals, the 90 % (thicker horizontal line) and the 95 % CI (thinner horizontal line). In the case of interpersonal trust, we see that there is no considerable difference between paper and web, and by definition there is no systematic difference to the whole sample of the push-to-web experiment because the mean is the weighted average from the web and paper subsample. But compared to the ESS round 9 data, there is a statistically significant higher level of trust compared to the PtW-data ($p=0.00$, $F(1,2824)=18.38$)[9]. Regarding the other two variables we cannot identify any substantive mean differences.

9 An adjusted Wald test has been used instead of a t-test to assess mean differences of weighted data. The

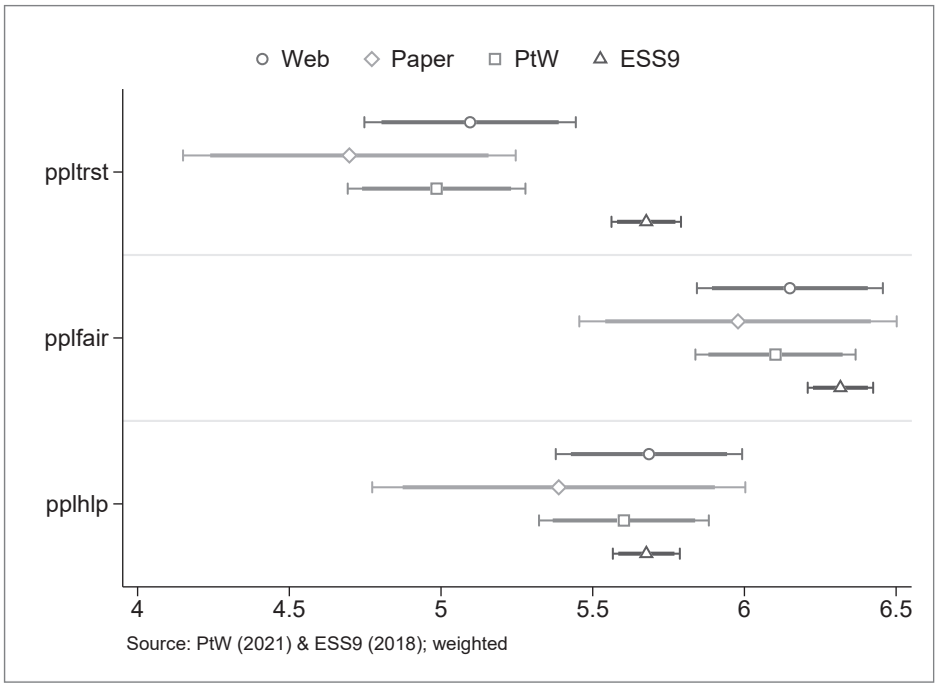

Figure 1 : Social connectedness

The next bundle of variables ask after political interest and political efficacy, i.e., the subjective evaluation of one's own ability and the possibility to be politically active. The respective question wording:

- (polintr) How interested would you say you are in politics? (4-point scale ranging from "Very interested" to "Not at all interested")
- (psppsgva) How much would you say the political system in [Country] allows people like you to have a say in what the government does? (5-point scale ranging from "Not at all" to "A great deal")
- (actrolga) How able do you think you are to take an active role in a group involved with political issues? (5-point scale ranging from "Not at all able" to "Completely able")
- (psppipla) How much would you say that the political system in [Country] allows people like you to have an influence on politics? (5-point scale ranging from "Not at all" to "A great deal")

results are the same.

– (cptppola) How confident are you in your own ability to participate in politics? (5-point scale ranging from "Not at all confident" to "Completely confident")

Table 5: Means and SE – political interest & efficacy (weighted)

Variable	Paper (n=91)		Web (n=225)		PtW (n=316)		ESS R9 (n=2,385)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
polintr	2.10	0.08	2.10	0.06	2.10	0.05	2.39	0.02
psppsgva	2.32	0.10	2.45	0.06	2.42	0.05	2.37	0.02
actrolga	2.44	0.14	2.66	0.08	2.60	0.07	2.39	0.03
psppipla	2.21	0.11	2.49	0.06	2.41	0.05	2.38	0.02
cptppola	2.78	0.13	2.91	0.08	2.88	0.07	2.60	0.03

Again, Figure 2 visualizes the data and shows the point estimates and respective confidence intervals. Three variables are of considerable importance: political interest (polintr) and two variables tapping the subjective evaluation of one's own political efficacy (actrolga, cptppola). ESS round 9 data show a lower level of political interest (the scale is reversed) and lower levels of subjective political efficacy compared to the PtW data.

Visually the confidence intervals for the ESS9 data and those for the group which used the paper questionnaire overlap, the large confidence interval of the latter data is mainly due to the small sample size. However, after conducting an adjusted Wald test for equality of variable means we are confident that in all three cases the means of the ESS9 data and the PtW data are statistically significant different.[10]

The concept "institutional trust" unites the next battery of questions, whereas each item has the same 11-point answer scale ranging from "No trust at all" to "Complete trust". The battery starts with the introduction: "On a scale of 0–10 how much do you personally trust each of the following institutions?" and it includes the following items:

– (trstprl) [Country]'s parliament?
– (trstlgl) The legal system?
– (trstplc) The police?
– (trstplt) Politicians?
– (trstprt) Political parties?
– (trstep) The European Parliament?
– (trstun) The United Nations?

10 polintr (p=0.00, F(1, 2825)=35,5); actrolga (p=0.0037, F(1, 2755)=8.43); cptppola (p=0.0001, F(1, 2781)=15.93).

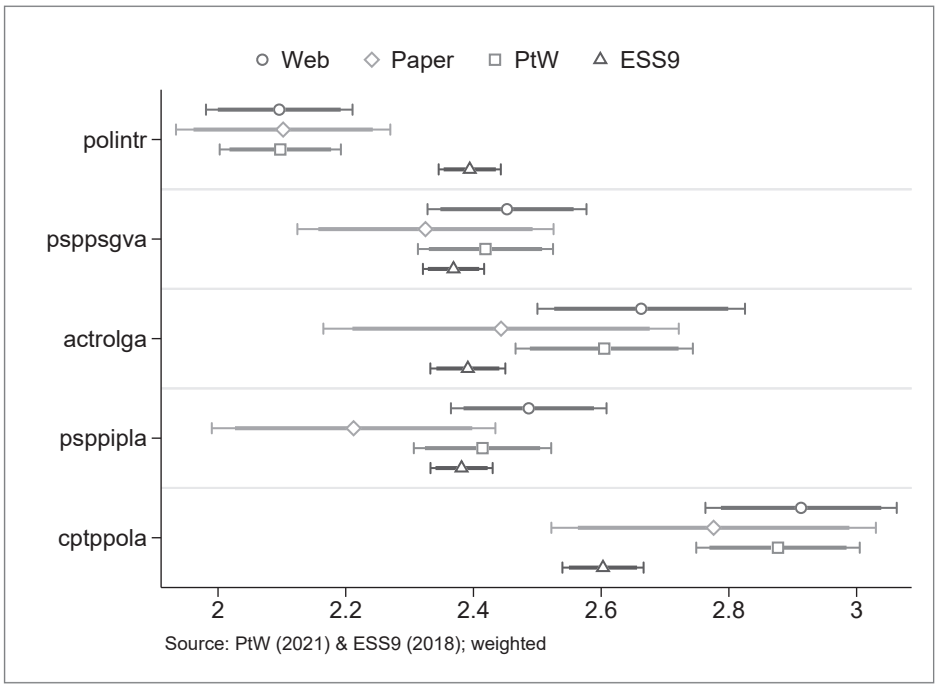

Figure 2: Political interest & efficacy

Table 6: Means and SE – institutional trust (weighted)

Variable	Paper (n=87)		Web (n=216)		PtW (n=303)		ESS R9 (n=2,246)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
trstprl	5.06	0.36	5.68	0.18	5.52	0.17	5.51	0.07
trstlgl	5.57	0.39	6.35	0.17	6.14	0.17	6.64	0.07
trstplc	6.77	0.26	7.19	0.17	7.08	0.14	7.26	0.06
trstplt	3.83	0.26	4.11	0.17	4.03	0.14	4.30	0.06
trstprt	3.61	0.25	4.09	0.16	3.96	0.14	4.23	0.06
trstep	3.40	0.31	4.25	0.19	4.03	0.17	4.62	0.07
trstun	3.99	0.32	4.61	0.18	4.45	0.16	4.96	0.07

Before looking at possible differences we must keep in mind – the same refers to previous questions analysed – that the data from ESS9 and PtW have been collected at different points in time, ESS round 9 data collection took place from September to December 2018 and the PtW experiment took place in February 2021 in the midst of

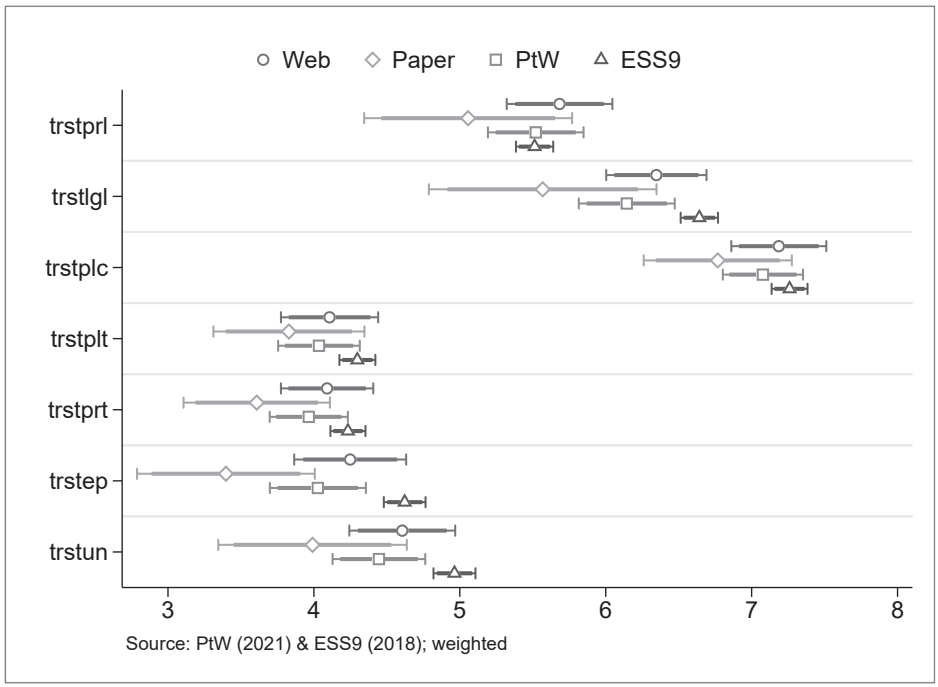

Figure 3: Institutional trust

the COVID-19 pandemic. On the 8th of February a "hard" lock-down ended in Austria and the first cases of mutation B.1.351 "South-African" led to travel restrictions to and from Tyrol on 12th of February. Thus, it is very plausible that attitude changes have occurred. Keeping that in mind, we can again identify several variables in Figure 3 with significant differences between PtW and ESS9 data.

Visually we can identify that trust in the legal system, the European parliament and the United Nations are higher in the ESS9 data compared to the PtW data. Those suspicions are confirmed after using again an adjusted Wald test.[11] Whatever the reasons might be for the lowered trust in the legal system, the European parliament and the United Nations, there is one important point to mention and that is the constant rank order of trust levels in the ESS9 as well as the PtW data. The lowest institutional trust on average is ascribed to political parties, followed by politicians, the European parliament, the United Nations, the national parliament, the legal system and the winner in the "trust race" in both datasets – the police.

11 (trstlgl (p=0.0087, F(1, 2801)=6.89); trstep (p=0.002, F(1, 2717)=9.53); trstun (p=0.0029, F(1, 2643) z=8.88).

General satisfaction with one's own life and with different institutions is the thematic context for the next six variables under consideration. The respective question wordings are:

- (stflife) All things considered, how satisfied are you with your life as a whole nowadays? Please answer using this card, where 0 means extremely dissatisfied and 10 means extremely satisfied.
- (stfeco) On the whole how satisfied are you with the present state of the economy in [country]?
- (stfgov) Now thinking about the [country] government, how satisfied are you with the way it is doing its job?
- (stfdem) And on the whole, how satisfied are you with the way democracy works in [country]?
- For the next two questions the end points of the 11-point answer scale change ranging from "Extremely bad" to "Extremely good".
- (stfedu) […] please say what you think overall about the state of education in [country] nowadays?
- (stfhlth) […] please say what you think overall about the state of health services in [country] nowadays?

Table 7: Means and SE – satisfaction (weighted)

Variable	Paper (n=92)		Web (n=216)		PtW (n=308)		ESS R9 (n=2,258)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
stflife	7.31	0.25	7.25	0.17	7.26	0.14	7.96	0.05
stfeco	4.82	0.35	5.31	0.18	5.18	0.17	7.00	0.05
stfgov	5.15	0.37	5.36	0.18	5.30	0.17	5.22	0.07
stfdem	5.86	0.37	6.48	0.17	6.31	0.16	6.39	0.06
stfedu	5.16	0.33	5.35	0.17	5.30	0.15	6.27	0.06
stfhlth	6.93	0.26	7.20	0.16	7.13	0.14	7.17	0.06

Table 7 provides the means and standard errors of the items at hand. The analysis process stays the same, thus we will have a look at the graphical visualizations of this data. Figure 4 provides the means and respective confidence intervals for the items under discussion.

At first glance we see stark differences between Web/Paper/PtW and ESS9 especially regarding three items, the satisfaction with the economy (stfeco), the subjective evaluation of the state of education in Austria (stfedu), but also with the general satisfaction

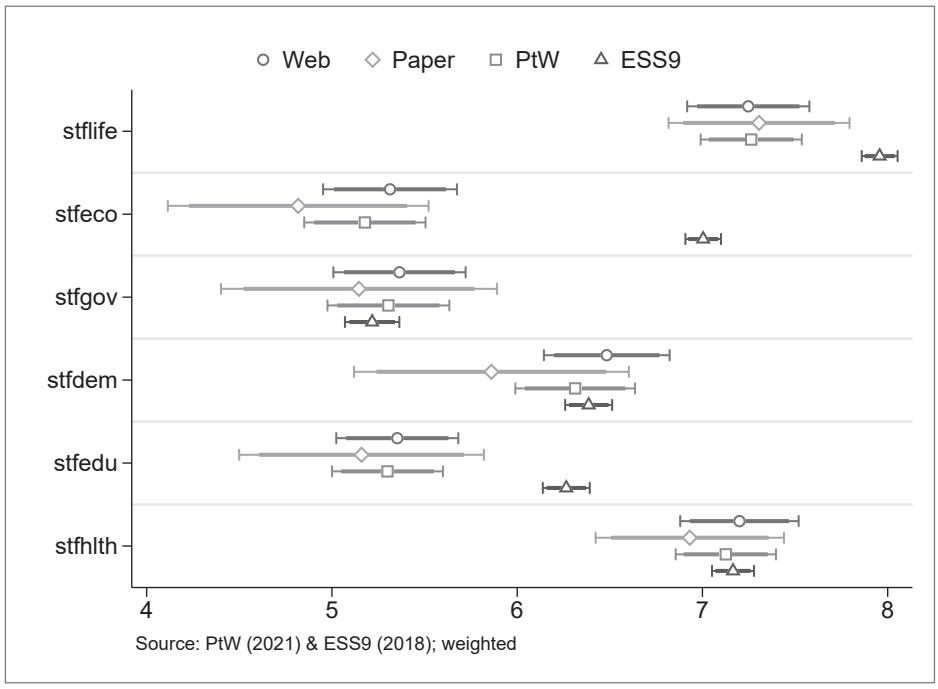

Figure 4: Satisfaction

with one's own life (stflife). Generally, there is more divergence between ESS9 and PtW data within this set of items compared to the previous examinations. The drop in satisfaction with the economy from ESS9 to PtW is considerable, amounting to roughly 1.8 scale points. However, the substance of the items and their respective changes are not at the core of this analysis. What makes this set of variables so interesting is that also the rank order changed between these two datasets, the average level of satisfaction with the economy and the mean evaluation of the state of education in Austria changed their position in the rank order.[12] Again, we have no possibility to assess whether these changes are caused by an attitudinal change or by mode effects.

The next set of variables is of special interest, all items touch different issues regarding homosexuality. If mode effects can be caused by social desirability because of the absence of an interviewer in self-completion approaches then attitudes should on average be more dismissive in the PtW data. All items are part of a battery with an answer scale ranging from "Agree strongly" to "Disagree strongly". The following

12 The differences for these three variables are again statistically significant. stflife ($p=0.00$, $F(1, 2823)=22.66$); stfeco ($p=0.00$, $F(1, 2764)=111.78$); stfedu ($p=0.00$, $F(1, 36.05)$).

question wordings have been fielded in ESS round 9 and the PtW experiment after the introduction "[…] please say to what extent you agree or disagree with each of the following statements.":

- (freehms) Gay men and lesbians should be free to live their own life as they wish.
- (hmsfmlsh) If a close family member was a gay man or a lesbian, I would feel ashamed.
- (hmsacld) Gay male and lesbian couples should have the same rights to adopt children as straight couples.

A first look at the data in Table 8 reveals a clear pattern. Two means are lower in the PtW data (freehms & hmsacld) which implies that they have more positive attitudes towards gay/lesbian rights and one mean is higher which implies that on average respondents in the PtW data disagree more with the statement that they would feel ashamed if a family member is gay/a lesbian.

Table 8: Means and SE – gay/lesbian related issues (weighted)

Variable	Paper (n=88)		Web (n=220)		PtW (n=309)		ESS R9 (n=2,311)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
freehms	1.96	0.13	1.55	0.06	1.65	0.06	1.87	0.03
hmsfmlsh	3.75	0.13	4.12	0.08	4.02	0.07	3.97	0.03
hmsacld	2.72	0.15	2.18	0.09	2.32	0.08	2.60	0.04

The graphs show a further pattern, namely that there may also be differences within the PtW data between the mode web and paper.

After conducting several hypotheses tests we can say that the mean attitudes towards the free life of gay man and lesbians as well as the level of agreement that gay male and lesbians should have the same right to adopt children are statistically different in the ESS9 and the PtW data. Furthermore, the mean differences between web and paper (freehms, hmsfmlsh & hmsacld) are also statistically significant. Responses collected online are more liberal, i.e. more in favour of gay men and lesbians, whereas respondents who choose the paper questionnaire are on average more conservative. As previously mentioned age is a strong predictor of mode choice, thus it is very likely that not the mode is the cause for those differences in attitudes but age. Generally, we can find no signs or provisional evidence for any effects caused by social desirability.

The last set of items deals with attitudes towards migrants with different characteristics and the consequences of migration on society. The answer scale for the first three

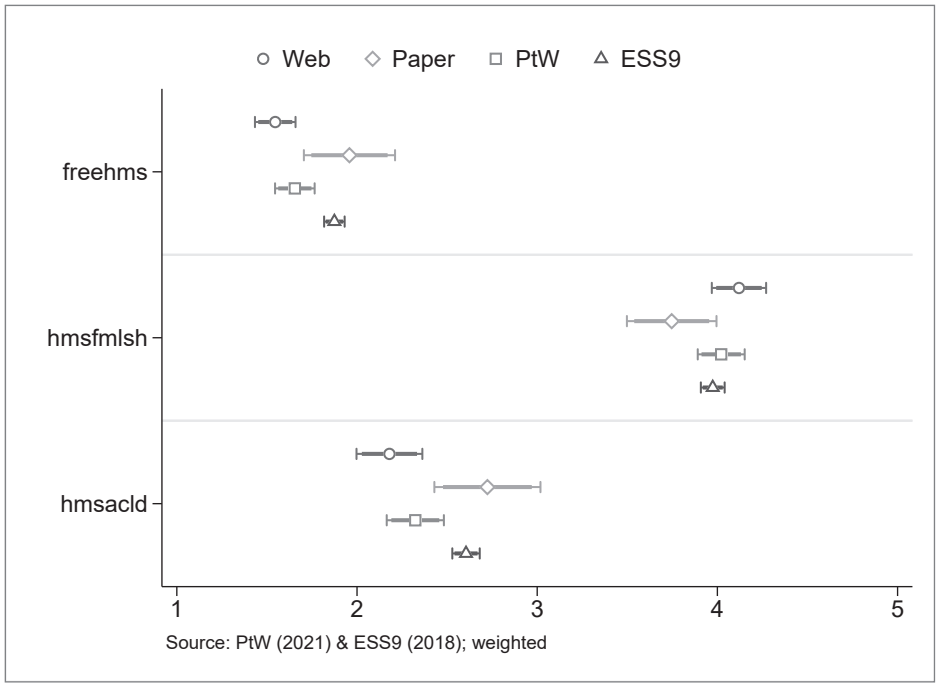

Figure 5: Gay/lesbian related issues

items comprises four values and ranges from "Allow many to come and live here" to "Allow none". The other response scales are below. The question wordings are:

- (imsmetn) […] to what extent do you think [country] should allow people of the same race or ethnic group as most [country]'s people to come and live here?
- (Imdfetn) How about people of a different race or ethnic group from most [country] people?
- (impcntr) How about people from the poorer countries outside Europe?
- (imbgeco) Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries? (11-point scale ranging from "Bad for the economy" to "Good for the economy")
- (imueclt) […] would you say that [country]'s cultural life is generally undermined or enriched by people coming to live here from other countries? (11-point scale ranging from "Cultural life undermined" to "Cultural life enriched")
- (imwbcnt) Is [country] made a worse or a better place to live by people coming to live here from other countries? (11-point scale ranging from "Worse place to live" to "Better place to live")

From the data presented in Table 9 we immediately see that the average attitudes in the ESS round 9 data are consistently higher compared to the PtW data.

Table 9: Means and SE – immigration (weighted)

Variable	Paper (n=83)		Web (n=209)		PtW (n=293)		ESS R9 (n=2,225)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
imsmetn	2.12	0.08	1.89	0.06	1.95	0.05	2.09	0.02
imdfetn	2.52	0.11	2.32	0.07	2.37	0.06	2.52	0.02
impcntr	2.52	0.08	2.36	0.07	2.40	0.05	2.60	0.03
imbgeco	4.71	0.33	5.38	0.20	5.20	0.17	5.54	0.07
imueclt	3.97	0.31	4.84	0.23	4.61	0.19	5.18	0.07
imwbcnt	3.89	0.31	4.50	0.18	4.34	0.16	4.67	0.07

From Figure 6 we can observe that there are mean differences regarding the attitudes about the openness to the migrants with different characteristics and with respect to the impact of migration on cultural life. Indeed, we find statistically significant differences between ESS round 9 and PtW data for all of the six items.[13]

What is striking is that for the first three items attitudes in the PtW data are consistently (and significantly) lower than in the ESS round 9 data implying that on average respondents are more open to migration. Simultaneously, regarding the consequences of migration respondents from the PtW experiment evaluate the outcome of migration in Austria on average more negatively. Of course that could be an artefact of aggregation and we must not interpret this findings as contradictory individual attitudes because of the possibility of an ecological fallacy. What is again important to note is that the rank order from the ESS round 9 data is also preserved in the PtW data, i.e. respondents are on average more open to migrants from the same ethic group similar to Austrians and more eager to reject migrants who are poor. Similarly, respondents evaluate the consequences of migration for the economy more positively compared to the consequences for the cultural life and least favourable for Austria as a place to live.

The analysis so far offers some evidence for the equivalence of PtW and ESS9 data not being severely plagued by mode effects. However, looking only at means and their confidence intervals might be misleading because we miss the functional form and equivalence of the distributions. Assessing these characteristics will be the next step.

13 imsmetn ($p=0.0051$, $F(1, 2764)=7.87$); imdfetn ($p=0.161$, $F(1, 2756)=5.8$); impcntr ($p=0.0005$, $F(1, 2740)=12.02$); imbgeco ($p=0.0332$, $F(1, 2724)=4.54$); imueclt ($p=0.0018$, $F(1, 2745)=9.75$); imwbcnt ($p=0.0424$, $F(1, 2735)=4.12$)

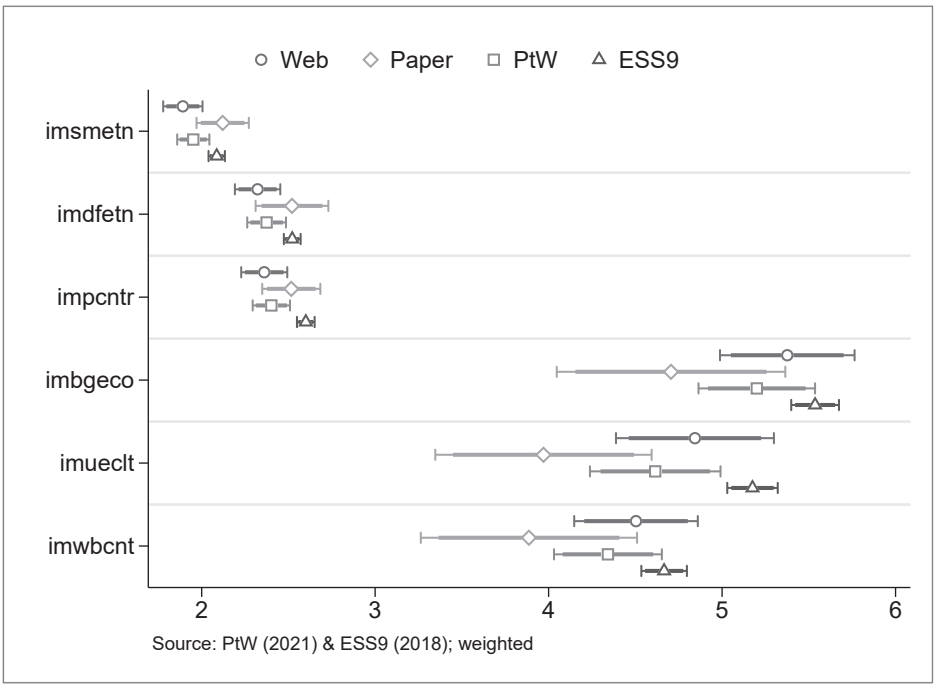

Figure 6: Immigration-related issues

5.3 Comparing empirical cumulative distributions

There are several possibilities of comparing distributions and testing if they come from the same underlying population. One can use a simple t-test, comparing two means, a Kolmogorov-Smirnov-Test, testing distributional equality, or a X2-test, comparing expected and empirical frequencies. Another approach is to test two distributions value by value, which also offers the possibility of assessing at which values distributions differ instead of being only able to estimate whether the distributions are equal as a whole or not.[14]

We have tested 46 items using the Kolmogorov-Smirnov-test and tested each value of these distributions using the distcomp command provided by David Kaplan.[15] The results of these two tests for all the variables under consideration are presented in the annex. Summarizing the results, the Kolmogorov-Smirnov-test rejected the distributional equality in 7 out of 46 cases and the evaluation using distcomp rejected distributional

14 Goldman/Kaplan 2018; Kaplan 2019.

15 Kaplan 2019.

equality in 2 out of 46 cases. Overall, those results are very promising, and it seems that the PtW data generating process resembles the one from the ESS9 f2f-mode very closely.

A last step of analysis remains and referring to the phenomenon of satisficing, i.e., respondents rushing through the questionnaire always choosing the same answer, especially through an item battery. This response pattern is also referred to as non-differentiation or straightlining.

5.4 Assessing Satisficing

The theory of survey satisficing assumes that certain respondents try to avoid the cognitive workload to process each item separately and form an adequate attitude, but under certain conditions choose the same response value for all items. We assume that this effect should be even more pronounced if the item bundle is also visually presented to respondents as in the PtW paper questionnaire, whereas in the web version each single trust question has been presented on a separate screen. That should be the case because the respondent is immediately aware of a bunch of items with the same response scales.

To assess the amount of satisficing we calculate the coefficient of variation for each respondent, i.e., the standard deviation divided by the mean using the user-written command respdiff in Stata. In the following we will focus on the trust item battery because this bundle offers interesting characteristics for testing satisficing effects:

- A sufficient number of items (usually 4–5 as a minimum)
- A visually different presentation between modes (web vs paper)
- Testing possible between a fully self-completion approach and interviewer-led data collection
- A rather high number of answer options

In a first step we will have a look at the distributions of the differentiation index for the two modes in the PtW data and then between the PtW and the ESS9 data. Figure 8 shows the graphs of these four distributions. What is striking is that there are low to medium levels of satisficing, the plethora of values are under 0.5 of the differentiation index scale, the lower the value of the index the more respondents differentiate between various response options and the lower are effects of satisficing. Focusing on the differences of satisficing between web and paper our hypothesis that in the paper version satisficing effects are higher compared to the web version are corroborated although only with using a one-sided t-test ($p(T<t)=0.038$, $t=-1.76$, $df=329$). Thus, the satisficing effects are indeed higher in the paper version compared to the web survey. We find no empirical evidence that there are statistically significant differences in the average satisficing effects between the PtW and ESS9 data.

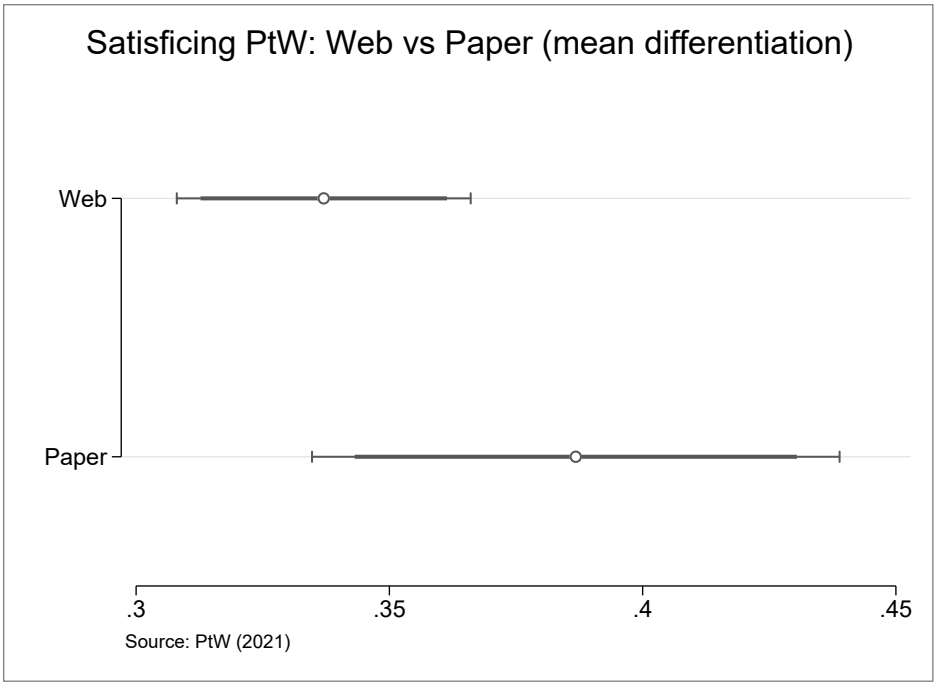

Figure 7: Comparing differentiation index

6 Conclusion

The switch to a self-completion push-to-web mode in round 10 in Austria has been driven by the dynamics of the COVID-19 pandemic. As often mentioned in other situations, the pandemic did not create these problems but rather brought them to surface or made the problem more severe. Comparably, survey practitioner as well as survey researcher knew that interviewer-led surveys become more and more problematic, mainly because of the high costs. ESS ERIC consequently decided to switch completely to a self-completion approach in the short to medium term. Thus, it is of importance to know which consequences such a mode switch might have and how to encounter possible drawbacks. Consequently, the push-to-web experiment conducted in Austria, Hungary and Serbia is a first step in assessing what the mode differences might be.

The empirical investigation in Austria so far yielded promising results that a switch to a self-completion push-to-web mode is a viable strategy to cope with most of the problems the "classical" F2F-approach is plagued with. The sample composition of the PtW data is pretty much in line with the ESS9 data except the overrepresentation of older cohorts which may vanish as the "new" technologies become a common tool for

all age groups. Age is still a significant driver of selection effects in a self-completion mode. In combination with satisficing effects this circumstance is of course an important area for further investigations and refinement. There is obviously a need for sophisticated analysis to disentangle trends as well as selection and mode effects to be able to evaluate them separately, because when we found changes between two points in time, we have otherwise no possibilities to evaluate what the causes for change are. The best solution would have been to randomly assign mode to respondents, which obviously is not the case because respondents chose by themselves which mode to use.

The comparison of means and their respective confidence intervals between the two datasets also offers promising results regarding the adequacy of a self-completion approach in the ESS. As already mentioned, we are unfortunately not able to differentiate between trend effects and mode effects explicitly. However, the mean differences between the PtW and ESS round 9 data are usually very small and even in those cases when we find statistically significant differences the rank order between related variables are preserved. Deviating from this pattern is only the measurement of satisfaction with the economy, the difference approximately amounts to 1.8 scale points, which is a huge difference on a 11-point scale. The satisfaction with the economy is much lower in the PtW data, which may be the case because of the COVID-19 situation and the massive economic consequences the pandemic created.

Comparing distributional equality has been the next step after checking central tendencies. Using two methods – the Kolmogorov-Smirnov-Test and the value-wise comparison of discrete cumulative distributions – yields very strong support for the similarity or equality of the data-generating processes of the f2f-approach of ESS round 9 and the self-completion approach of the push-to-web survey.

Finally, we investigated the PtW data for satisficing effects – a strategy to avoid cognitive efforts by choosing the same response options throughout a battery of items with the same response scale – which should be more present in situations without an interviewer. This analysis also yields empirical support for only small mode effects between f2f and self-completion.

This push-to-web experiment has been the first tryout to find a suitable solution within the ESS survey infrastructure to switch the mode of data collection. As already mentioned, the COVID-19 pandemic did not really create the need for a mode switch but accelerated the process. Further research – as always – is needed to assess possible mode effects, especially creating a research design which allows to disentangle possible trend effects and selection effects from mode effects. Nevertheless, the empirical results at hand are promising that self-completion may be indeed a reliable and trustworthy alternative to the hitherto gold standard within the ESS survey infrastructure and beyond.

Literature

Ansolabehere, Stephen/Schaffner, Brian F.: Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison, in: Politcal Analysis 22 (2014), pp. 285–303.

Bowyer, Benjamin T./Rogowski, Jon C.: Mode Matters. Evaluating Response Comparability in a Mixed-Mode Survey, in: Political Science Research and Methods 5 (2017), pp. 295–313.

Dillman, Don A./Smyth, Jolene D./Christian, L eah Melani: Internet, Phone, Mail, and Mixed-Mode Surveys: the Tailored Design Method, 4th edition, Hoboken 2014.

Gnambs, Timo/Kaspar, Kai: Socially Desirable Responding in Web-Based Questionnaires. A Meta-Analytic Review of the Candor Hypothesis, in: Assessment 24 (2017), pp. 746–762.

Goldman, Matt/Kaplan, David M.: Comparing Distributions by Multiple Testing across Quantiles or CDF Values, in: Journal of Econometrics 206 (2018), pp. 143–166.

Heerwegh, Dirk: Mode Differences Between Face-to-Face and Web Surveys. An Experimental Investigation of Data Quality and Social Desirability Effects, in: International Journal of Public Opinion Research 21 (2009), pp. 111–121.

Heerwegh, Dirk/Loosveldt, Geert: Face-To-Face versus Web Surveying in a High-Internet-Coverage Population. Differences in Response Quality, in: Public Opinion Quarterly 72 (2008), pp. 836–846.

Holbrook, Allyson L./Green, Melanie C./Krosnick, Jon A.: Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias, in: Public Opinion Quarterly 67 (2003), pp. 79–125.

Kaplan, David M.: distcomp: Comparing Distributions, in: Stata Journal 19 (2019), pp. 832–848.

Krosnick, Jon A.: Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys, in: Applied Cognitive Psychology 5 (1991), pp. 213–236.

Annex

	Kolmogorv-Smirnov-Test			Results from distcomp
variable	P-value	exact P-value	Sig.	Equality of ECDFs
netusoft	0,00	-0,00	non sig.	Reject
hmsacld	0,01	0,01	non sig.	Do not reject
aesfdrk	0,01	0,01	non sig.	Do not reject
freehms	0,03	0,03	non sig.	Do not reject
health	0,04	0,04	non sig.	Do not reject
imsmetn	0,04	0,04	non sig.	Do not reject
psppipla	0,05	0,04	non sig.	Do not reject
hlthhmp	0,06	0,05	sig.	Do not reject
hmsfmlsh	0,09	0,08	sig.	Do not reject
actrolga	0,09	0,08	sig.	Do not reject

	Kolmogorv-Smirnov-Test			Results from distcomp
imwbcnt	0,11	0,09	sig.	Do not reject
imueclt	0,14	0,12	sig.	Do not reject
trstep	0,16	0,15	sig.	Do not reject
imbgeco	0,21	0,19	sig.	Do not reject
psppsgva	0,22	0,20	sig.	Do not reject
lrscale	0,23	0,21	sig.	Do not reject
cptppola	0,26	0,24	sig.	Do not reject
gincdif	0,28	0,26	sig.	Do not reject
trstun	0,32	0,29	sig.	Do not reject
ppltrst	0,33	0,30	sig.	Do not reject
sclmeet	0,48	0,45	sig.	Do not reject
pstplonl	0,58	0,54	sig.	Do not reject
eufttf	0,59	0,55	sig.	Do not reject
trstprt	0,62	0,58	sig.	Do not reject
imdfetn	0,71	0,67	sig.	Do not reject
trstlgl	0,72	0,68	sig.	Do not reject
stfeco	0,75	0,71	sig.	Reject
stfhlth	0,76	0,72	sig.	Do not reject
trstprl	0,80	0,77	sig.	Do not reject
trstplc	0,81	0,77	sig.	Do not reject
stfdem	0,81	0,77	sig.	Do not reject
sclact	0,82	0,79	sig.	Do not reject
pplhlp	0,83	0,79	sig.	Do not reject
impcntr	0,89	0,85	sig.	Do not reject
sgnptit	0,96	0,95	sig.	Do not reject
stfgov	0,97	0,95	sig.	Do not reject
trstplt	0,97	0,95	sig.	Do not reject
stflife	0,98	0,97	sig.	Do not reject
pplfair	0,98	0,98	sig.	Do not reject
stfedu	0,99	0,98	sig.	Do not reject
happy	1,00	1,00	sig.	Do not reject
polintr	1,00	1,00	sig.	Do not reject
badge	1,00	1,00	sig.	Do not reject
contplt	1,00	1,00	sig.	Do not reject
crmvct	1,00	1,00	sig.	Do not reject
bctprd	1,00	1,00	sig.	Do not reject