Lugtig, 2024a
2024-10-16

Comparison of ESS rounds 9 and 10: mode differences

Peter Lugtig

Why estimating mode measurement effects is difficult

One of the major problems in estimating the mode measurement effect in mixed-mode surveys is that isolation of the causal effect of mode on measurement is difficult. Using face-to-face or postal recruitment for the survey will result in different types of respondents that participate in the survey, and differences in selection effects that are hard to separate from the fact that respondents also react differently to the mode of the interview. Some studies have tried to eliminate selection effects by for example re-interviewing face-to-face respondents in a self-interviewing mode shortly after the original interview (Klausch et al., 2014). Or they have randomized respondents into a survey mode only after successfully recruiting a respondent into the survey (Heerwegh, 2009). Both these designs are a bit artificial, and are in practice always complex to explain and administer to respondents. The European Social Survey (ESS) did not use such a complex design in their mixed-mode experiments in round 10.

Mode experiments in Round 10 of the ESS and ways to estimate mode measurement effects

ESS round 10 was conducted with a self-completion instrument (web and paper) in 9 countries, with the 22 other countries using face-to-face interviewing. In Finland and Great Britain, a within country experiment was conducted where the samples were randomly assigned to one of two modes: one part of the sample was recruited and interviewed face-to-face, while the second part of the sample was assigned to postal recruitment and self-interviewing (push-to-web). These differences in interview modes both within countries and between countries, along with comparisons to data collected in earlier rounds provide the ingredients for estimating the mode measurement effects.

Several kinds of comparisons can be made:

1. For the 9 countries using self-completion interviews, we can estimate the change in data associated with the switch from face-to-face (before round 10) to self-interviewing at round 10 for variables that have been collected at every wave. Apart from the problem that the change in mode between rounds 9 and 10 will affect both measurement and nonresponse, a third problem is that aggregate change between rounds 9 (collected before Covid) and round 10 (collected during Covid) is likely to occur as well for several variables. For the 22 countries that used face-to-face interviewing, we can estimate the same type of change however, and comparing both sets of change (countries switching towards self-interviewing vs staying with face-to-face interviewing) we can still learn quite a lot about how statistics change by moving to a self-interviewing mode. What will be easy to establish is whether mode effects occur, how large they are on average, and for what variables we find mode effects. It will be more difficult to establish why mode effects occur. Is this because of the progression of time and Covid, because of the fact that self-interviewing surveys attract different types of respondents compared to face-to-face surveys, or because of a true measurement effect?

2. For Great Britain and Finland, we can compare respondents from the within country experiments on all variables collected in ESS round 10 common to both modes. In this comparison, we still have the problem that answers may differ due to nonresponse and measurement, but there is no "Covid-effect". This paper will concentrate on the change from face-to-face interviewing to self-interviewing between round 9 and 10. The analysis of the experiments conducted within Finland and Great Britain are reported on in a second paper (Lugtig, 2024b). Findings in the companion report partly build on findings that are reported in this paper.

Table 1: Effects that can lead to a difference in interview modes in ESS

Comparisons	Time and Covid-effect	Mode selection effect	Mode measurement effect
8 countries: round 10 vs round 9 - face-to-face vs self-interviewing	xxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx
20 countries: round 10 vs round 9 - face-to-face vs. face-to-face	xxxxxxxxxxxxxx		
2 countries: round 10 within-country experiment - face-to-face vs self-interviewing		xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx

From Table 1 it is clear that there is no comparison within or between countries, or between different rounds of the ESS that makes it easy to isolate the mode measurement effect. The two designs (over-time comparisons and within country experiments) ask for a different analysis plan to take account of the unique problems in isolating and estimating the mode measurement effect. In this report, we focus on a comparison of the first two rows of Table 1. A separate analysis will dig deeper into the experiments conducted in Finland and Great Britain.

Setup and common variables between ESS round 9 and 10

The analyses reported in this document are focused on the effects of moving from face-to-face interviewing in round 9 of the ESS to self-interviewing in round 10 for 8 countries: Austria, Cyprus, Germany, Latvia, Poland, Serbia, Spain and Sweden. Israel also used self-interviewing in round 10, but is not included in the analyses because in round 9 no data were collected. In 20 countries, face-to-face interviewing was used in both round 9 and 10 to interview respondents. This is the case for Belgium, Bulgaria, Switzerland, Czechia, Estonia, Finland, France, the United Kingdom, Croatia, Hungary, Ireland, Iceland, Italy, Lithuania, Montenegro, the Netherlands, Norway, Portugal, Slovenia and Slovakia. North-Macedonia and Greece were not included in the analyses because they did not participate in round 9. Finally, cases experimentally assigned to self-interviewing in Great Britain and Finland are not included in these analysis.

Countries collecting data via self-completion in practice used a combination of web and paper-and-pencil self-interviewing. For brevity, we will use 'self-interviewing' or 'self-completion' as shorthand for 'web and paper-and-pencil interviewing'. Additionally, in some countries some face-to-face interviews were conducted using live video calls. When we refer to 'face-to-face' interviewing, we imply interviews conducted via CAPI or video-CAPI.

The ESS includes questions that are repeated over waves, and questions that are unique to a particular wave. In this document we will focus on variables that ESS round 9 and ESS round 10 have in common. These are 380 variables. Of these variables, 13 variables are character variables (indicating for example the respondent id or PSU). Further, 227 variables contain large amounts of missing data, mostly because these are questions asked only to a small subset of respondents. For example, some include details about up to 15 household members that are answered by only a few respondents, are follow-up questions, or questions asked uniquely in the self-interviewing questionnaires or one of the ESS rounds 9 or 10, but not both. Finally, the dataset contains 8 variables containing information about weights and the ESS round. This leaves us with 108 numerical variables that we can use in the analysis.

Mode effects for numerical variables

To explain the approach for studying mode effects, we will first concentrate on one question, labeled 'ppltrst' in the ESS datasets. This question asks respondents:

"Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted".

The plots on the next pages show the results for three groups of respondents:

1. The top figure shows responses in ESS round 9, for countries using self-interviewing in round 10
2. The second figure shows responses in ESS round 9, for countries using face-to-face interviewing in round 10
3. The third figure shows responses in ESS round 10 using self-interviewing
4. The bottom figure shows responses in ESS round 10 using face-to-face interviewing

The distribution of answers in the 2 top figures shows for this particular question that there is little difference in the general trust for round 9 between countries that in round 10 switch to self-interviewing, or stay using face-to-face interviewing. The mean level of general trust in other people per group is indicated by the red line, and the lines show very small differences when comparing the top 2 figures. The mean level of trust is about 5.0 in both. In round 9 we see no large differences between countries that keep using face-to-face interviewing in round 10, or switch to self-completion.

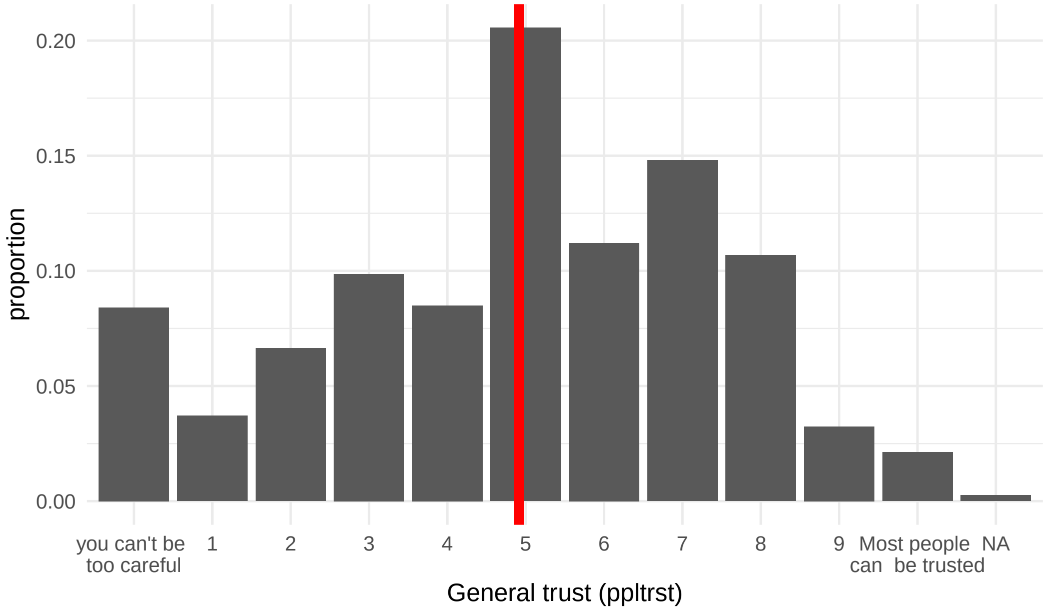
The results for general trust in round 10 are shown in the bottom two figures. For countries using self-interviewing, we can compare figure 1 with figure 3 to find quite large changes in the distribution of answers between round 9 and 10 for countries that switch to self-interviewing. The mean declines from about 5.0 in round 9 (face-to-face) to about 4.5 in round 10 (self-interviewing). The interquartile range also becomes wider in round 10 when interviews are conducted via self-interviewing, and we see a substantial increase in the number of people that answer '0' - no trust at all in self-interviewing. For countries that use face-to-face interviewing in round 10 (bottom figure), we find little or no changes between rounds 9 and 10, implying that the decline in general trust is not seen when countries use the same mode of interviewing. By comparing the bottom two figures, we see further evidence that changing the mode of interviewing from face-to-face to self-interviewing leads to more negative answers, more '0' answers, and a decrease in the mean level of general trust by about 0.5.

These findings are in line with the theory that self-interviewing interviewing is less susceptible to social desirability, leading to more negative attitudes in self-interviewing in comparison to face-to-face interviewing. Because we do not see any change between rounds 9 and 10 in the face-to-face interviewing groups, the change we see in round 10 interviews by self-interviewing is most likely caused by the mode switch. A potential competing explanation for these changes we see with the transition to self-interviewing is that the correlates of nonresponse are different in face-to-face and self-interviewing. It could be that those respondents with lower social trust are more likely to participate on the web or paper surveys than in face-to-face surveys, leading to an observed decline in social trust when interviews are conducted via self-interviewing. Or, that countries in which the mode-effect or time-effect was particularly large were more likely to switch to self-interviewing. Although the difference we find seems to be too large to be potentially explained by such nonresponse effects, we cannot exclude this as a potential cause. The difference we find can be caused by mode selection differences, by mode measurement differences, or both.

The analyses of the experiments conducted in Great Britain and Finland will allow us to disentangle mode selection and mode measurement effects in these two countries. For now we just study whether change occurs that is associated with the change from face-to-face to self-interviewing. A break in the time-series for the mean caused by a mode-switch risks making it hard to conduct longitudinal analyses using ESS data. Such breaks in the time series can perhaps be reduced by changing fieldwork procedures, questionnaire design or introducing specific adjustment methods, but the first step is to find out how large mode effects are, and for what variables they exist. The question is of course whether our finding for the single political trust item extends to other variables, and whether we find similar mode effects for other statistics, such as variances and covariances. To investigate how large the problem of mode effects is across the entire ESS, we now turn to a systematic analysis of mode effects for all variables in ESS rounds 9 and 10.

Figure 1: Answers for 'ppltrst' in ESS9, mode−switch countries

Among countries that used face−to−face at round 9 and self−completion at round 10



The mean value for 'ppltrst' is shown by the red line.

Figure 2: Answers for 'ppltrst' in ESS9, face−to−face only countries

Among countries that used face−to−face at round 9 and round 10

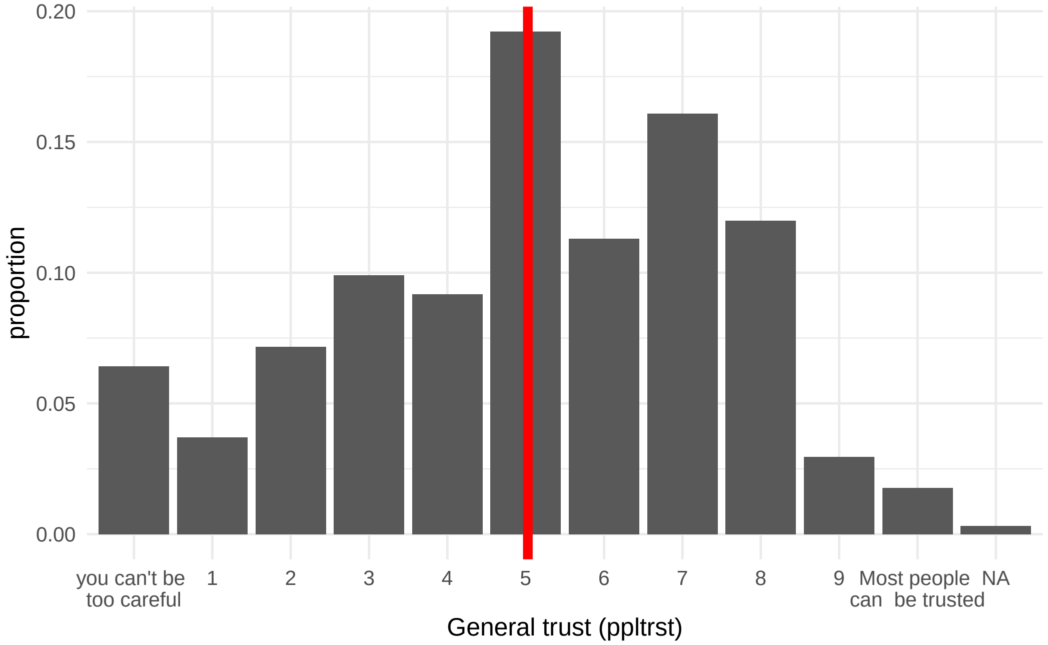


Figure 3: Answers for 'ppltrst' in ESS10 self−completion

Among countries that used face−to−face at round 9 and self−completion at round 10

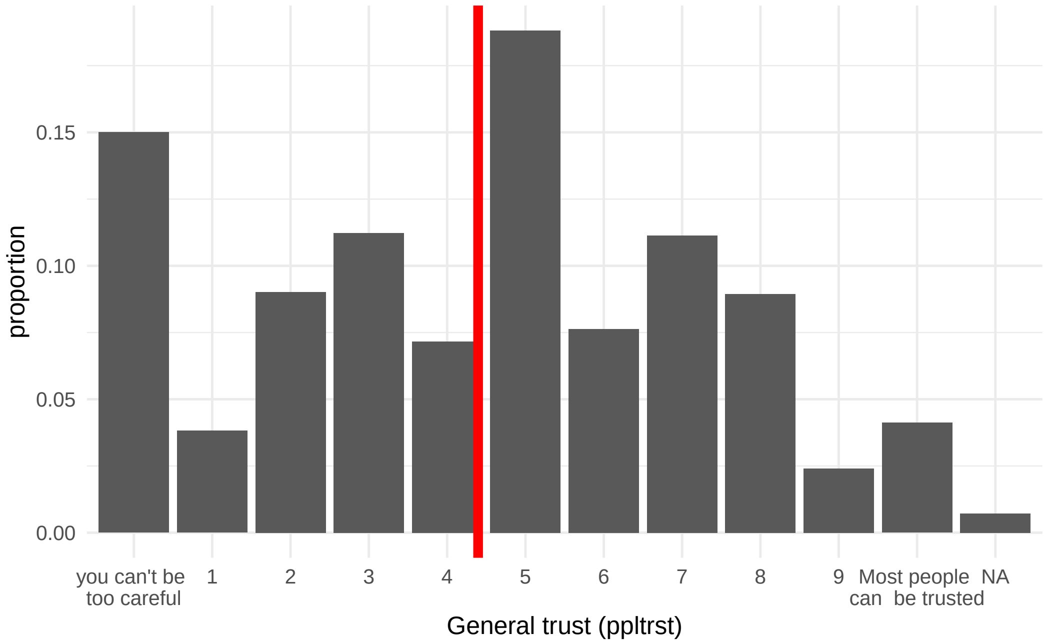
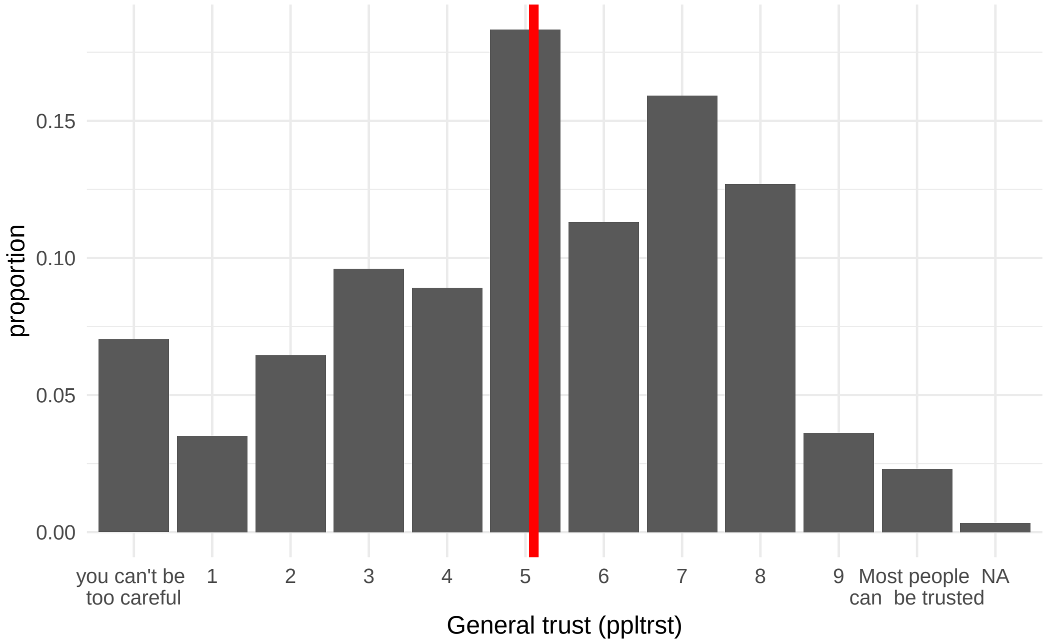


Figure 4: Answers for 'ppltrst' in ESS10 face−to−face

Among countries that used face−to−face at round 9 and round 10



In order to facilitate the comparison across variables, we will not look at changes in the mean, but rather use a standardized difference measure to facilitate the interpretation of the size of the difference. As we are interested in the difference in means, we are using the 'Standardized Mean Difference' (SMD), which is defined as the difference between groups divided by the (pooled) standard deviation from the two groups. This measure is also known as Hedges g-value (Hedges, 1982).

For the example of the question of the variable 'ppltrst', we are comparing 4 groups, and could therefore theoretically make 6 different comparisons. We will concentrate however on comparing the following groups:

1. The Standardized Mean Difference for the comparison between round 10 face-to-face and round 9 face-to-face for countries using face-to-face interviewing in both rounds. This indicates an effect of 'time'. In our example, for the variable 'ppltrst' the observed difference in the mean between figures 2 and 4 is -0.06, which is translated into a Hedges g-value of -0.03, which is a very small effect (Kelley & Preacher, 2012).
2. The Standardized Mean Difference for the comparison between round 10 self-interviewing and round 9 face-to-face interviewing for countries that switched modes. This effect includes both the effect of 'time' as in the first comparison, but also the switch of modes. In our example the observed difference between the means in figures 1 and 3 is -0.51. This translates into a Hedges g-value of -0.19. This is still a small, but in most cases relevant, difference (Kelley & Preacher, 2012).

Comparison across variables

Now, we systematically compute Hedges g-values for both comparison groups across all numeric variables in the ESS.

The results of this analysis are shown in table 2. The first two rows in this table show the distribution of Hedges g across the 108 variables for which we tested the standardized differences. Higher Hedges g-values imply larger (standardized) differences.

Table 2: Distribution of Hedges g for time (round 9-10) and mode-effect (f2f-self) across ESS variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Hedges g face-to-face round 9 vs 10	-0.12	-0.02	0.01	0.01	0.03	0.13
Hedges g mode-switch round 9 vs 10	-0.44	-0.12	0.00	-0.01	0.08	0.37
absolute Hedges g face-to-face rounds 9 vs 10	0.00	0.01	0.03	0.04	0.06	0.13
absolute Hedges g mode-switch round 9 vs 10	0.00	0.04	0.09	0.11	0.17	0.44

For the difference in means between face-to-face interviews in round 9 and 10, we find the median Hedges g to be 0.00, indicating that over all variables, respondents have reported almost the same means across time. The interquartile range is between values of -0.02 and +0.03, indicating that for most variables we see very small levels of change. There are some outlier variables for for which changes are larger; these outliers are sometimes positive and sometimes negative, and on the whole cancel each other out. In short, we find that among face-to-face interviews conducted in rounds 9 and 10, there is some level of aggregate change over time in some variables, but these aggregate changes are mostly small or very small. On the one hand, this may not be so surprising, as societal change does not occur rapidly. On the other hand, it does appear there is no overall effect of Covid on the time-series.

When comparing the differences for countries using self-interviewing in round 10 and face-to-face interviewing in round 9, the Hedges g-values are more dispersed. The median Hedges g-value between self-interviews in round 10 and face-to-face interviews in round 9 is again 0.00, which implies no aggregate change because of the switch in modes and the effect of time. The interquartile range however is much larger, and lies between -0.12 and +0.09. Such differences are still small, but they do show that for some variables we find larger changes between waves when there is also a mode switch. This is also illustrated by the maximum Hedges g-value for change between rounds 9 and 10 of countries that switched modes being much larger, at around 0.5.

One reason why we find both positive and negative differences is that the polarity of scales may differ from variable to variable. For this reason, we also report absolute Hedges g in rows 3 and 4 of table 2. Here we see the larger differences in the comparison of self-interviewing vs face-to-face as compared to face-to-face round 9 vs face-to-face round 10 more clearly. The average absolute Hedges g for means between rounds 9 and 10 when face-to-face interviewing is used in both waves is about 0.04, which again shows there is no, or very little, overall change between waves. This average difference is about 3 times higher (0.11) in absolute terms when a switch is made from face-to-face to self-interviewing. This is an important finding: although shifts in means associated with the change of modes are overall small, the change to self-interviewing does on average lead to small changes in means. In order to understand this better, it is important to look at the specific variables that change.

Figure 5 below shows the absolute Hedges g-values, shown for countries using face-to-face interviewing in both round 9 & 10 on the x-axis, with Hedges g for the countries switching to self-interviewing shown on the y-axis. Observations represent the Hedges g on both comparisons per variable. When observations are on or close to the reference line (in red), this indicates a shift in means that is about the same over time (rounds 9 and 10) as it is across interviewing methods and time (self-interviewing vs face-to-face). Observations below the red line show variables where the change in means over time is larger than the change due the composite effect of mode and time. Observations above the red line indicate variables for which the mode and time effect is larger than the time effect alone.

What we see is that for most variables, the mode and time effect is much larger than the time effect. Although there is a fairly large proportion of variables for which the mode effect is small, there is a small, but still substantial group of variables for which the mode effect is large. Again, it is important that we bear in mind that this mode-effect can be caused both by mode selection effects and mode measurement effects. Nonetheless, these findings show that, without action, a change of the ESS from face-to-face to self interviewing would be associated with large changes in the time series of means of some variables.

Figure 5: Hedges g for face−to−face and self−completion mode differences

For 108 common variables in ESS rounds 9 and 10

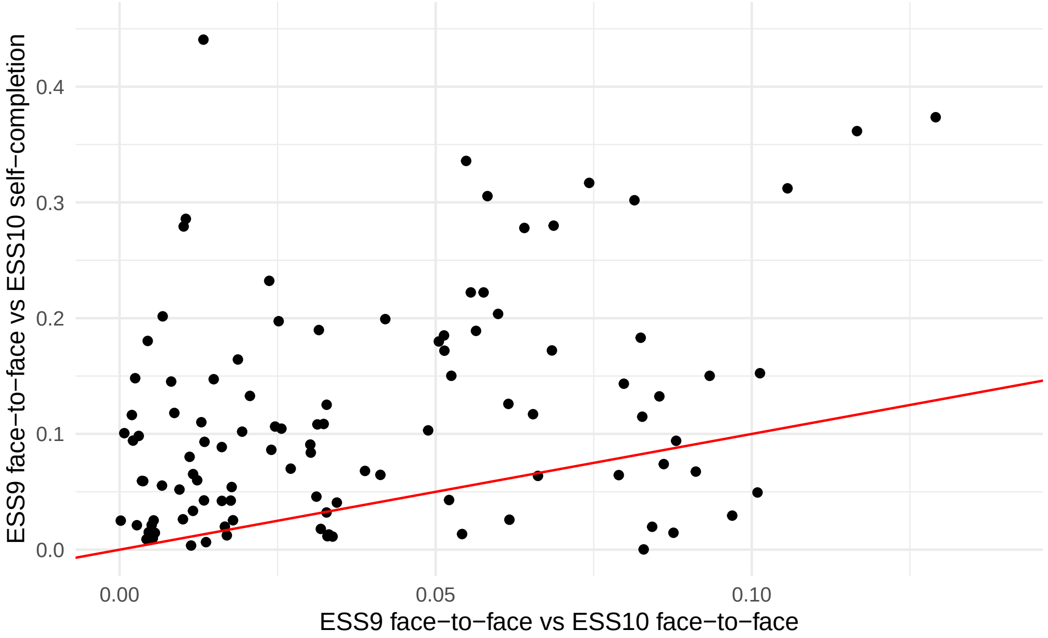


Table 3 below displays the 25 variables where we find the largest mode effect. The absolute Hedges g for all variables listed below is 0.18 or higher (small effect size according to Kelley and Preacher, 2012). We see that most of the variables where we find large mode effects are attitudinal. However, there are also several more factual variables where we find large mode effects.

Table 3: Top 25 variables with largest mode effects in means

Variable name	Variable label	Hedges g for time + mode
stfedu	State of education in country nowadays	-0.44
netusoft	Internet use, how often	0.37
hmsacld	Gay and lesbian couples right to adopt children	-0.36
freehms	Gays and lesbians free to live life as they wish	-0.34
edulvlb	Highest level of education	0.32
impcntr	Allow many/few immigrants from poorer countries outside Europe	-0.31
rlgatnd	How often attend religious services apart from special occasions	0.31
eisced	Highest level of education, ES - ISCED	0.30
happy	How happy are you	-0.29
imdfetn	Allow many/few immigrants of different race/ethnic group from majority	-0.28
rlgdgr	How religious are you	-0.28
aesfdrk	Feeling of safety of walking alone in local area after dark	0.28
polintr	How interested in politics	-0.23
imsmetn	Allow many/few immigrants of same race/ethnic group as majority	-0.22
dngna	Doing last 7 days: no answer	0.22
stfeco	How satisfied with present state of economy in country	-0.20
sclact	Take part in social activities compared to others of same age	-0.20
hmsfmlsh	Ashamed if close family member gay or lesbian	0.20
pray	How often pray apart from at religious services	0.20
ppltrst	Most people can be trusted or you can't be too careful	-0.19
sclmeet	How often socially meet with friends, relatives or colleagues	-0.19
atncrse	Improve knowledge/skills: course/lecture/conference, last 12 months	-0.19
stflife	How satisfied with life as a whole	-0.18
dngoth	Doing last 7 days: other	0.18
trstplt	Trust in politicians	-0.18

Because we find mode effects for means to be problematic, we will illustrate the problems across modes for three variables with some of the strongest mode effects. We chose not to show the variable 'dngna', as that variable indicates which respondents give no answer to the question asking what they have been doing in the last 7 days. In self-interviewing, skips are coded into this category, whereas in face-to-face interviewing, such responses are coded by the interviewer as 'don't know' or 'refusal.

- "stfedu", which is an attitudinal variable asking respondents how they view the state of education in their country.
- "netusoft", which is a behavioral variable asking respondents how often they use the Internet.
- "freehms", which is an attitudinal variable asking respondents whether gay men and lesbians are free to live as they wish.

The first variable we explore in more detail is the variable "stfedu". This asks respondents about the state of education in the country with answers categories:

0 - Extremely bad
1
2
3

4
5
6
7
8
9
10 - Extremely good

Strictly speaking, the variable is not continuous, and the median would be a more appropriate measure of central tendency than the mean. Still, changes in the mean do reflect changes in the answer distribution of this variable. Table 4 below shows that in self-interviewing, respondents appear to be much more negative about the state of education than when respondents are interviewed face-to-face. This finding may be caused partly by selection effects; respondents who are more negative about education are more likely to participate in a web survey.

Table 4: Proportions of answers for ‘state of education’ in round 9 (face-to-face) and 10 (self-interviewing)

	R10	R9
0 - Extremely bad	0.10	0.04
1	0.05	0.02
2	0.11	0.06
3	0.15	0.11
4	0.12	0.11
5	0.16	0.17
6	0.11	0.14
7	0.10	0.16
8	0.07	0.12
9	0.02	0.04
10 - Extremely good	0.02	0.04

As a second example, we look at the variable “netusoft”. This variable asks respondents how often they use the Internet, with answer options:

1 - Never
2 - Only occasionally
3 - A few times a week
4 - Most days
5 - Every day

Although this again is a variable that is ordinal, and the mean is not a proper central tendency measure, it is still useful as an indicator for detecting shifts in the distribution that occur with the change from face-to-face to self-interviewing. Table 5 shows that in self-interviewing in round 10, 72% report to use the Internet everyday, while in the same countries in round 9, this was 61%. At the other end of the distribution, we find that 7% of respondents in round 10 (self-interviewing) indicate never to use to Internet, while this is 20% in round 9 (face-to-face). Although a measurement effect could here also be at play, it is also conceivable that the difference we find here is caused mainly by selection differences between modes as we are more likely to see higher internet use among respondents in a web survey.

Table 5: Proportions of answers for ‘Frequency Internet use’ in round 9 (face-to-face) and 10 (self-interviewing)

	R10	R9
1 - Never	0.07	0.20
2 - Only occasionally	0.06	0.06

	R10	R9
3 - A few times a week	0.05	0.05
4 - Most days	0.10	0.07
5 - Every day	0.72	0.61

The third example we show is the variable ‘freehms’. This questions asks respondents whether they believe gay men and lesbians are free to live as they wish, with response options:

1 - Agree strongly
2 - Agree
3 - Neither agree nor disagree
4 - Disagree
5 - Disagree strongly

Table 6 shows that respondents in self-interviewing are more likely to respond positively to this question. In face-to-face interviewing, 72% of all respondents interviewed face-to-face are on the positive end of the scale, whereas in self-interviewing this is 81%. Although we cannot exclude the possibility that this difference is caused by selection effects, it is also possible that respondents respond more positively to this item when asked in self-interviewing, which would indicate a mode measurement effect.

Table 6: Proportions for ‘Gay men and Lesbians are free to live as they wish’ in round 9 (face-to-face) and 10 (self-interviewing)

	R10	R9
1 - Agree strongly	0.49	0.36
2 - Agree	0.32	0.36
3 - Neither agree not disagree	0.12	0.13
4 - Disagree	0.04	0.08
5 - Disagree strongly	0.03	0.07

The three examples just shown illustrate that there can be several reasons for why the distribution of answers to a survey question may change when the mode of interviewing is changed from face-to-face to self-interviewing web and paper. Mode measurement effects can occur due to differences in questionnaire presentation (possible to change with question redesigns) or social desirability (harder to do something about). Mode selection effects can also lead to shifts in means; these selection effects are in part preferable since they can be (partly) corrected for if these selection effects are correlated to variables used in the ESS weighting procedures. For the ‘Internet use’ question, variables like age, income and level of education are probably correlated with both the frequency of Internet use, and selection differences between the modes. It is important to keep in mind that there are multiple causes for the mode effects we find, and only experimental data can shed light on this.

Effects on variances

The descriptive results for the variable ‘ppltrst’ showed that apart from a shift in the mean that can be caused by a switch in survey modes, there may also be an effect on variances and covariances. For example, self-interviewing respondents can sometimes be more likely to choose a middle category when responding to questions, reducing variances. In contrast, for some questions, a reduction in social desirability may lead respondents to choose a more extreme response option. Including possible selection effects, it is unclear whether a switch to self-interviewing would on the whole lead to a change in variances, but this we will empirically investigate.

Similar to how we studied the effect on means, we will study the effect on variances for 108 variables that are asked in both ESS round 9 and round 10. The comparison we will make is to simply compare the variances across the 3 groups of interest:

1. responses in ESS round 9, for countries using self-interviewing in round 10
2. responses in ESS round 9, for countries using face-to-face interviewing in round 10
3. responses in ESS round 10 using self-interviewing
4. responses in ESS round 10 using face-to-face interviewing

We will again compare two groups: the self-completion web and paper responses in ESS round 10 to the face-to-face responses in ESS round 9, and compare them on the ratio of the variances from both groups. A ratio higher than 1 implies that variances are larger than in the face-to-face responses for round 10, and a ratio smaller than 1 implies that variances are smaller.

The ratios shown indicate that the effects of moving from face-to-face to self-interviewing web and paper on variances are relatively small. The ratios of the variances for the variables derived from face-to-face interviewing in rounds 9 and 10 center around or are slightly larger than 1, with the median being 1.02. Variances are on average about 2% larger in round 9 as compared to round 10. For some variables variances are slightly higher than 1, and others slightly lower than 1, but such fluctuations are to be expected due to chance.

In contrast to the rather large effects on means, we see that a transition to self-interviewing does not affect the variances very much. The median ratio of the variance is slightly higher than 1, with the median being 1.09. Variances in self-interviewing are on average 9% higher than variances in face-to-face interviewing. These differences are comparatively smaller than the differences we found in means.

Figure 6: Ratio of variances, ESS9−10 face−to−face

Among countries that used face−to−face in round 9 and round 10

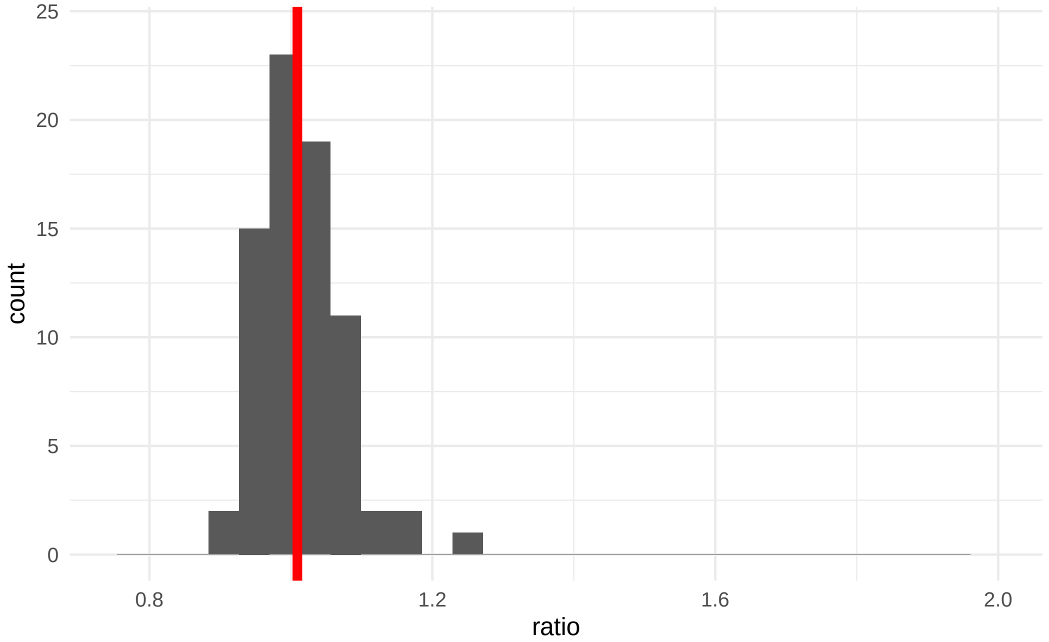
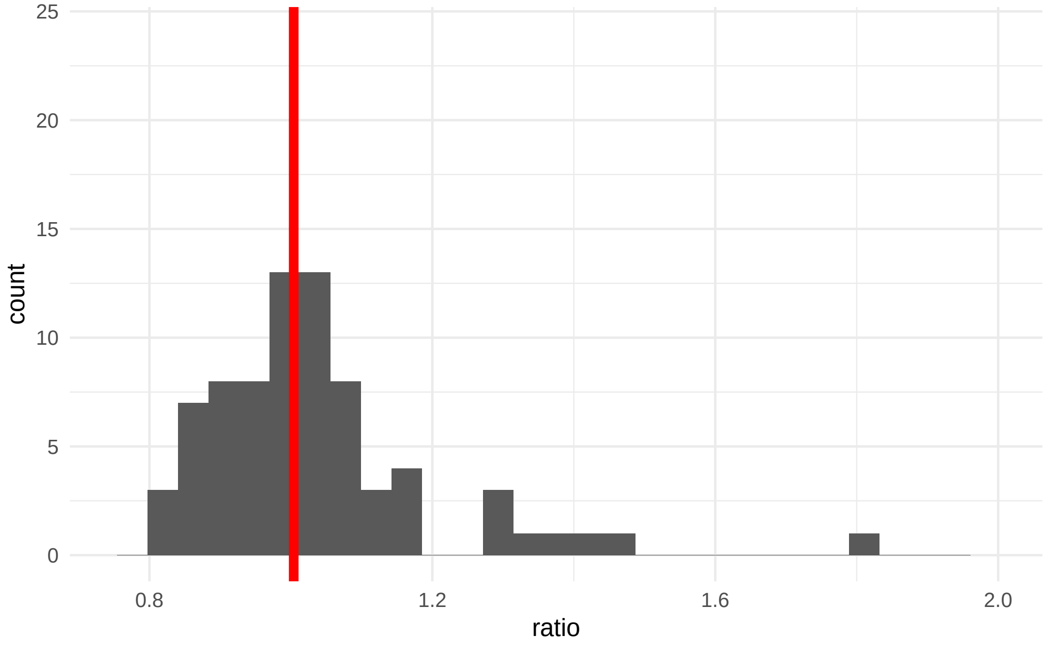


Figure 7: Ratio of variances, ESS9−10 mode−switch

Among countries that used face−to−face in round 9 and self−completion in round 10



Analysis on covariances

As a final step, we will analyse the effect on covariances. Here, we simply evaluate the size of the Spearman correlations between the 111 variables that we have been using so far, and compare the size of the correlations between the four groups:

1. responses in ESS round 9, for countries using self-interviewing in round 10,
2. responses in ESS round 9, for countries using face-to-face interviewing in round 10
3. responses in ESS round 10 using self-interviewing
4. responses in ESS round 10 using face-to-face interviewing

We will compare these four groups in two ways:

1. The difference in spearman correlations between round 10 face-to-face and round 9 face-to-face

2. The difference in spearman correlations between round 10 self-interviewing and round 9 face-to-face interviewing.

Our findings are similar to what we found for variances. Because we are now analysing nearly 15,000 correlations between variables, the plots showing differences in correlations contain more data, and therefore reflect smoother distributions. In both figures below, we see that the differences in correlations due to time (top figure), and the mode-switch (bottom figure) are very small. The average difference in correlations for face-to-face interviews between rounds 9 and 10 across all variable-pairs is exactly 0. The interquartile range of differences is about .02, implying that the vast majority of changes in correlations are very small, with a few exceptions.

For the mode effect, we similarly find that the average difference in correlations comparing face-to-face interviews in round 9 and self- interviews in round 10 is exactly 0. The interquartile range amounts to .04, and we find slightly more outliers. Closer inspection of the variables showing large differences shows that some variables in the ESS have very skewed distributions. Only small shifts in the distributions across small cell sizes can result in relatively strong differences in correlations. So even when differences in correlations are apparently large, in practice there seems to be little change in the size of correlations due to a switch to self-interviewing.

Figure 8: Difference in spearman r, ESS9−10 face−to−face

Among countries that used face−to−face in round 9 and round 10

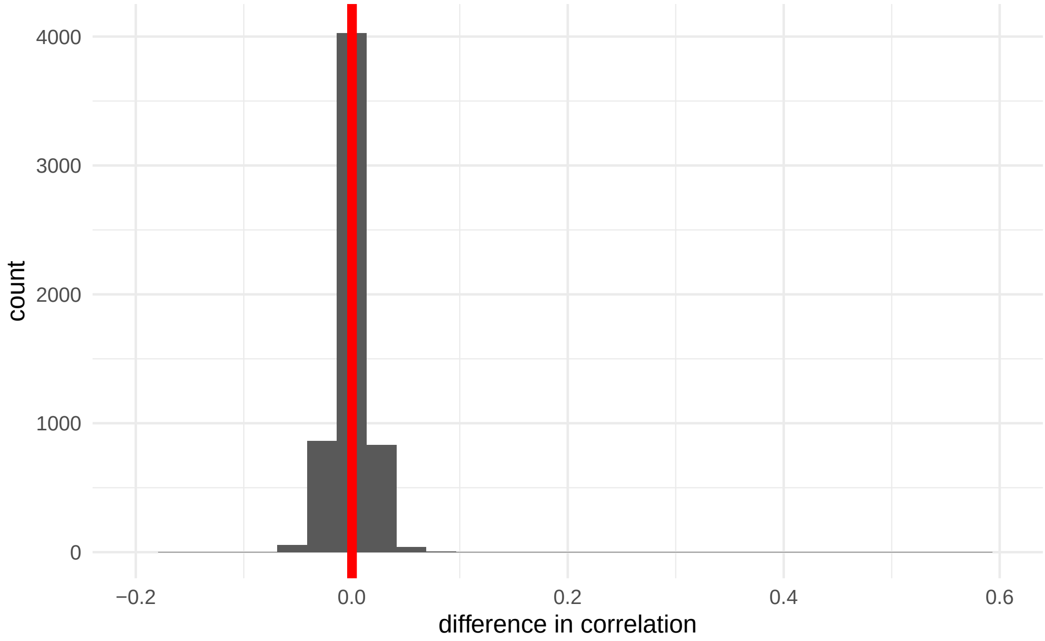
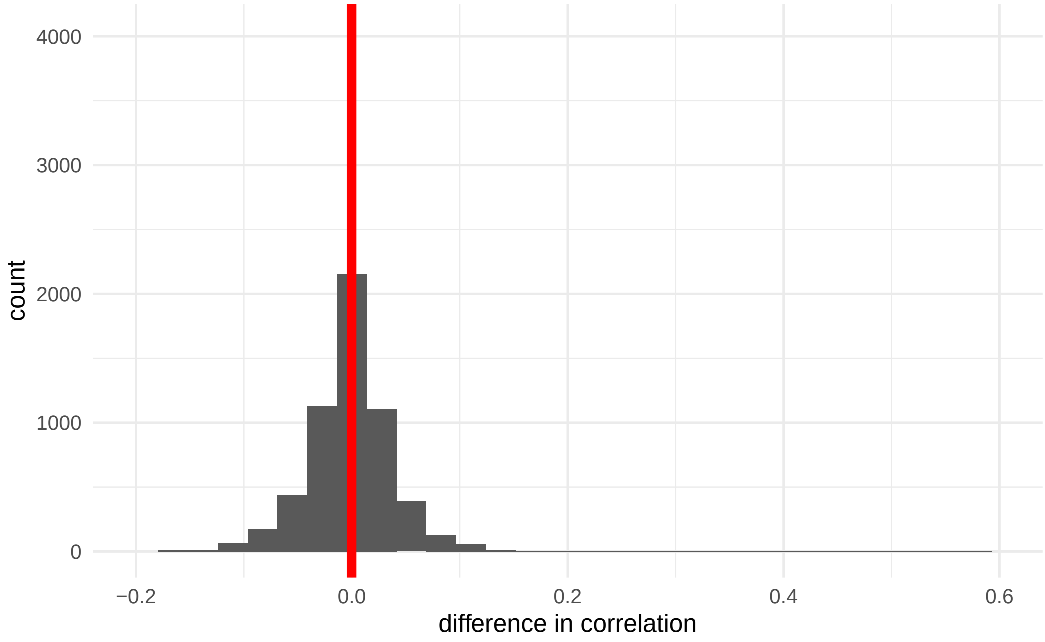


Figure 9: Difference in spearman r, ESS9−10 mode−switch

Among countries that used face−to−face in round 9 and self−completion in round 10



Zooming in on the reasons for mode effects in means

Our analysis so far showed that the switch towards self-interviewing that happened in round 10 of the ESS resulted in a shift in statistics that can be attributed to the change in survey mode. Mode effects are strongest for means, where we find that the average absolute shift in means is about 3 times larger than the change in means over time. Still, for most variables, the size of the mode effect is very small. It is only for about 20% of all variables that we find mode effects in means to be larger than 0.2 standardized means. Further, the mode effect is much weaker, or almost non-existent - for variances and correlations.

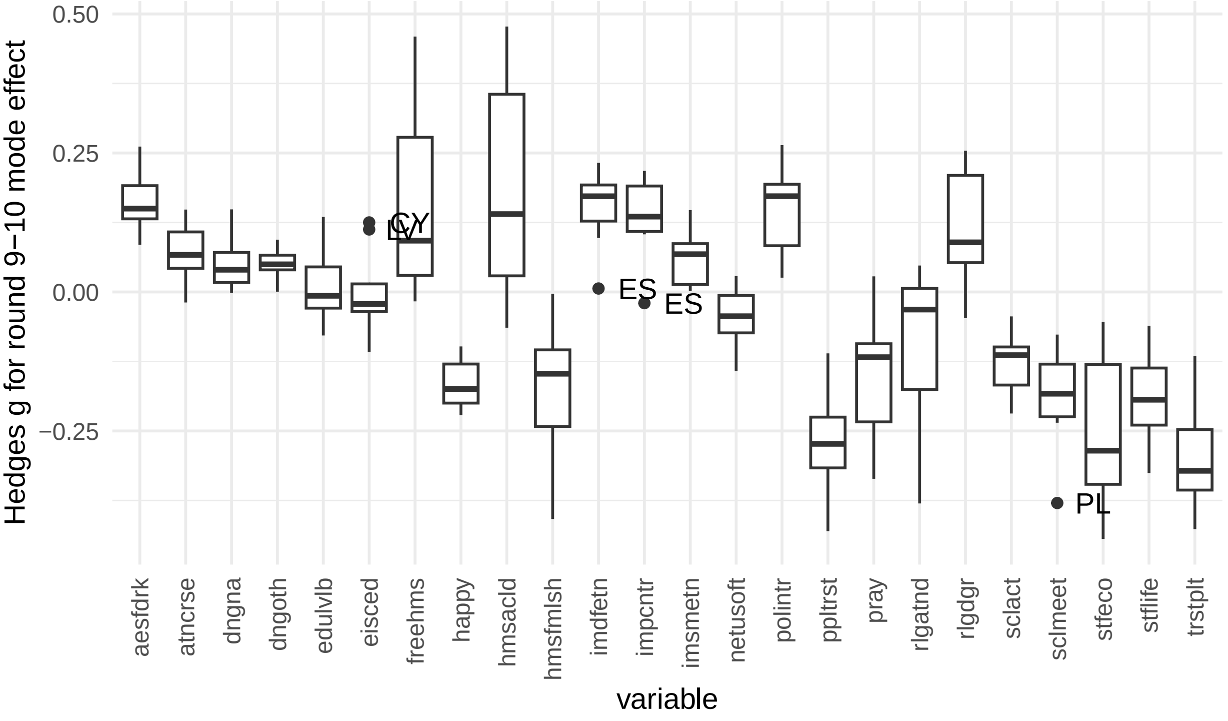
Now we will zoom in to understand whether mode effects in means are consistent across countries, whether the actual survey mode (web or paper) can explain parts of the effects that we find, and zoom in on other causes. We will start with looking at mode effects across countries.

One reason why it is interesting to look at the consistency of mode effects is from a survey design perspective. If mode effects are consistent, it is very likely that some feature of the survey design (be it the questionnaire design or fieldwork procedures) causes consistent mode selection or mode measurement effects. If mode effects are inconsistent across countries, it may mean that effects go in different directions, and will be much harder to control. Also, should there be a consistent break in time series, such a break can perhaps be corrected for. When breaks are inconsistent across countries, they are much harder to correct for.

We will concentrate on the difference for countries using self-interviewing in round 10, and face-to-face in round 9.

Figure 10: Country variation in ESS9−10 mode effects on means

Variables with the 25 largest mode effects between face−to−face and self−completion across countries



What we see in figure 10 above is that the largest shifts in means that we observe are quite consistent across countries. For some variables we find considerable between-country variation in the size of the mode effect, but overall, we see that differences across variables are much larger than differences between countries. This implies that solutions to deal with the mode-effects we find can be generic for the entire ESS, and no country-specific solutions are probably necessary. It is important to note that the analyses here are based on just the 9 countries switching towards self-interviewing.

Do we find differences within the self-interviewing mode?

Until here, we have treated the self-completion mode as being one mode. In practice, we know that respondents can complete the survey via web or on paper. In much of the literature, people have found that mode measurement effects between paper and web self-completion are small. However, there may be some particular questions that are an exception to this. Questions with long answer scales are typically presented horizontally on paper, but vertically on web to accommodate mobile survey completion which could cause a measurement effect. We will in these analyses only use data from round 10, and only use countries that used self-completion.

Figure 11: ESS9−10 mode effects vs within self−completion mode effects

Hedges g for self−completion & face−to−face vs Hedges g for self−completon web & paper

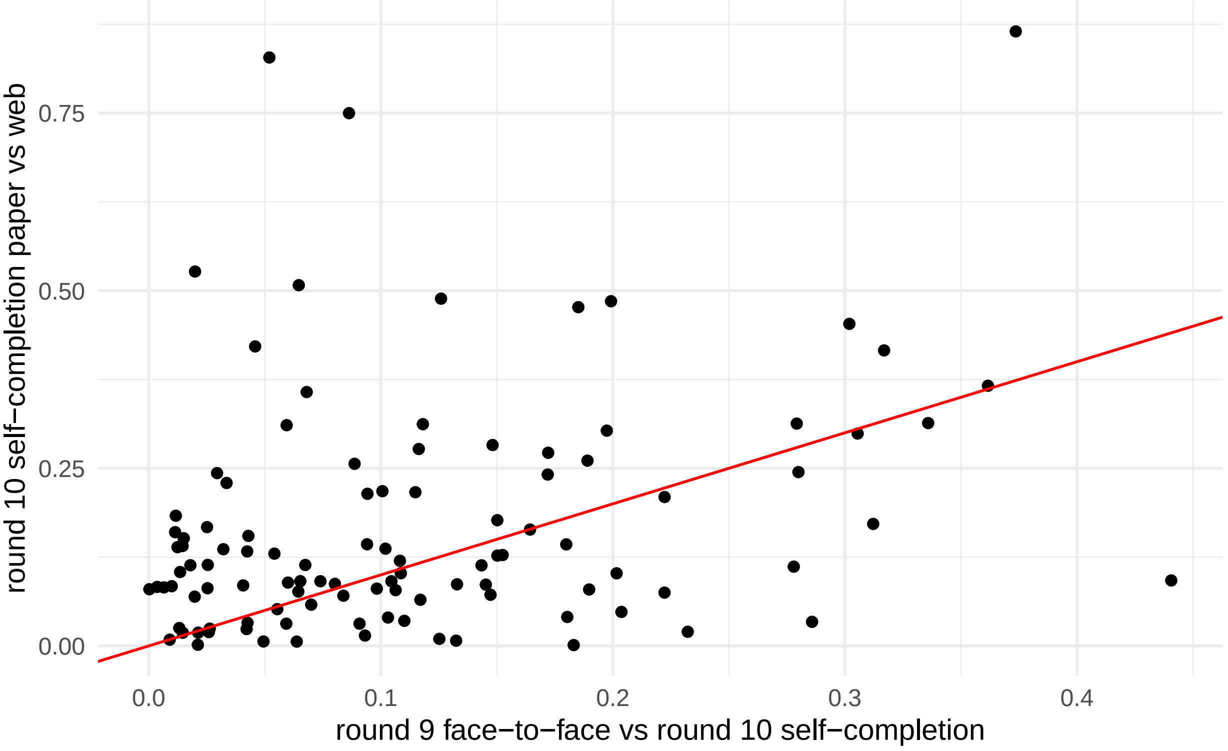


Figure 11 shows that we find large differences between paper and web responses. On the y-axis we show standardized mean difference between respondents completing the ESS by self-completion paper vs. web across the 108 numeric variables we also used in earlier analyses. The red line in the graph illustrates situations where the mode effect of paper vs. web is of the same size as the mode-plus-time effect that we earlier investigated when comparing self-completion responses in round 10 to face-to-face interviewing in round 9. The figure shows that the size of the differences between web and paper are often as big as the differences between face-to-face and self-interviewing. This implies that mode effects potentially also occur within self-completion modes. Whether these effects are caused by true measurement differences, or by selection effects into paper and web is unclear. It is to some degree expected that the web and paper modes attract different kinds of respondents. This is often even desirable, as earlier research has shown that a combination of web and paper modes leads to less nonresponse bias than using the web mode alone.

To investigate whether the differences we observe between the paper and web modes are caused by selection differences, we use statistical matching as a technique to create a balanced set of respondents that have responded in both modes. We expect that after matching, respondents in the matched web subsample and the matched paper subsample do not only resemble each other on the variables used in matching, but — assuming the matching variables are correlated with out outcome variable — also resemble each other more on variables not used in the matching variables. We therefore expect the differences that we find between the web and paper modes in ESS round 10 to decrease because of matching. The size of the decrease tells us something about the relative size of selection effects conditional on the matching covariates present in these data.

As covariates we use the same variables as in the parallel report that focused on the experimental results of Finland and the United Kingdom. These variables are:

- gender of the respondent (variable name: 'gndr')
- educational level (variable name: 'edulvlb')
- age (variable name: 'agea')
- reporting no internet access (variable name: 'accnone')
- Whether the respondent already had covid at the time of the survey (variable name: 'respc19')
- Whether someone in the household already had covid at time of the survey (variable name: 'resphh19')

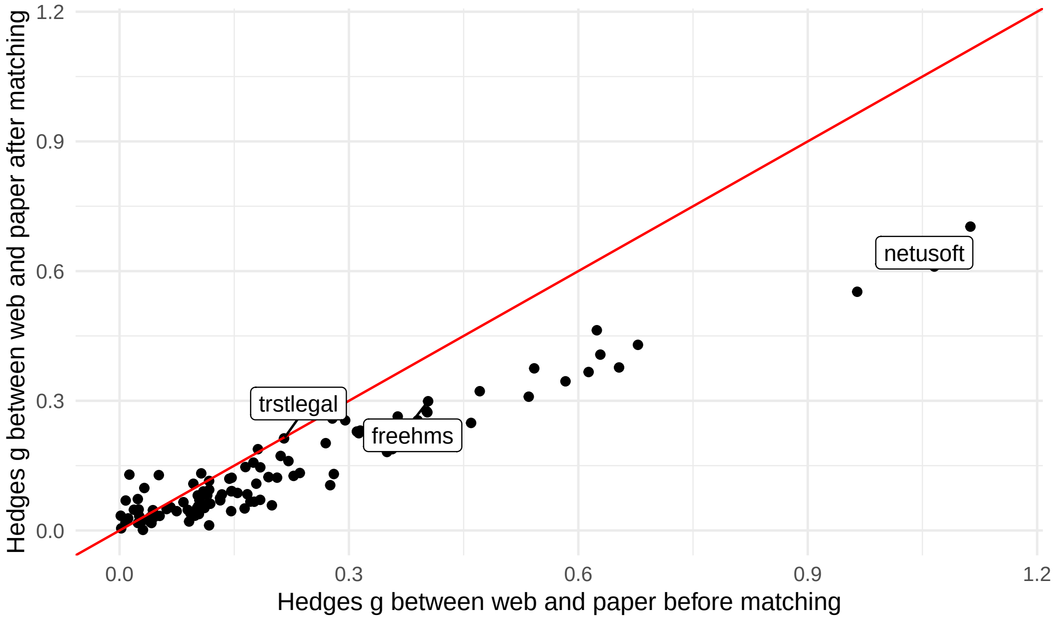
Variables were chosen based on three criteria: 1. variables should not be susceptible to large mode measurement effects, 2. variables should be related to mode selection effects and 3. variables should be related to the outcome variables in the ESS. Details on the evaluation of these criteria are found in the parallel report.

The matching procedure was carried out using coarsened exact matching and using the {Matchit} package in R. Matching was only done within a country; respondents from for example Poland could only be matched to other respondents from Poland. It was possible to match 1 respondent to multiple respondents in the other mode. Out of a total of 12998 respondents who completed the ESS on paper, 7914 respondents were matched. Out of all 7496 web respondents, 4923 were matched. Subsequent analyses are only carried out with the 7496 + 4923 that were matched, and where covariate balance is 100%.

Figure 12 below shows that the Hedges g-values decrease when we compare them before (x-axis) and after matching. The red line indicates that mode effects are the same, and it is clear that for the majority of variables, we find much smaller differences after matching. This is especially true for variables where we found large differences before matching. Overall, the reduction in mode effects after matching are of medium size. On average, the selection effects we identified account for about 50% of observed mode differences. This means that the other half of the mode effect between web and paper responses are unaccounted for, and are probably due to other, unobserved selection effects, or because of a mode measurement effect. It is quite likely that we have unobserved selection effects, and so conclude that if mode measurement effects exist between paper and web, that they are small.

Figure 12: Mode effects on means before and after matching

Mode effects between self−completion web and paper after matching on 7 covariates



To illustrate the problems we sometimes see when comparing web and paper responses, we again zoom in to some variables. We here choose to focus on three variables, that exhibit different patterns:

1. The variable "netusoft" which measures the frequency of Internet use. Here, we see the largest difference between paper and web of all variables, that are reduced by about 50% after matching,

2. The variable "freehms", which measures respondents attitude about the statement whether 'Gay men and Lesbians are free to live as they wish'. Here where we see smaller differences that are only slightly reduced by matching,

3. The variable "trstlgl", which asks respondents whether they are satisfied with the legal system in the country. Here we find a medium-sized mode effect that is not changed at all after matching.

Table 7: Frequency of Internet use

	paper	web
1 - Never	0.14	0.02
2 - Only occasionally	0.12	0.02
3 - A few times a week	0.08	0.03
4 - Most days	0.12	0.09
5 - Every day	0.53	0.84

For the variable 'Frequency of Internet use' we find that respondents on the web report to use the Internet a lot. 84% of respondents uses the Internet every day. For paper responses we see that only 53% uses the Internet every day, while 12% or 14% use the Internet 'only occasionally' or 'never'. We see that the distribution of the responses for the web responses is actually quite similar to the distribution of the responses in face-to-face as was reported in Table 4. Although the covariates in matching such as age and level of education are correlated with the frequency of Internet use, and selection into mode, they do not fully explain selection effects. Still, it seems quite likely that the difference that we find between web and paper responses is caused by selection effects into both modes since the matching procedure reduced the size of the effect by about half.

Table 8: Gay men and Lesbians are free to live as they wish

	paper	web
1 - Agree strongly	0.39	0.56
2 - Agree	0.36	0.28
3 - Neither agree not disagree	0.15	0.10
4 - Disagree	0.06	0.03
5 - Disagree strongly	0.04	0.03

For the question asking respondents whether they think ‘Gay men and Lesbians are free to live as they wish’, in Table 8 we find that respondents completing the ESS on web are more likely to strongly agree with the statement. For this variable, matching did not reduce the size of the mode effect by as much as for the frequency of internet use. This could be because we are missing relevant covariates in the matching procedure; we do find that age, educational level, and having no Internet access are strong predictors, however. The question itself is a 5-point scale, and there is no reason why we would expect a mode measurement effect for this variable. The most likely explanation of differences between modes is that there are selection effects that remain unexplained. It appears that paper respondents were more likely to espouse more socially conservative views on this subject.

Table 9: Trust in the legal system

	paper	web
0 - No trust at all	0.12	0.07
1	0.05	0.04
2	0.08	0.07
3	0.09	0.09
4	0.08	0.07
5	0.16	0.14
6	0.07	0.09
7	0.11	0.14
8	0.12	0.15
9	0.06	0.09
10 - Complete trust	0.07	0.04

Finally, Table 9 shows the distributions of paper and web responses for the variable ‘trstlgl’ which measures trust in the country’s legal system on an 11-point scale. This is a variable where perhaps the horizontal (on paper) or vertical (on web) placement of the answer scale could make a difference. Table 9 shows however that the distributions are quite similar, but that people on the web are on average slightly more trusting. The matching procedure in this instance did not reduce any of these differences, suggesting that the small differences were mostly explained by the demographic compositions of both groups of respondents not included in the matching, or that there is indeed a small measurement effect, that can be caused by the placement of the scale.

Conclusion

This report shows that for countries that used face-to-face interviewing in both round 9 and 10 of theESS, we find no strong differences on means, variances and covariances between the rounds. This suggests that time and the Covid pandemic had minimal effects on measurement in the ESS. When we compare countries that switched from using face-to-face interviewing in round 9 to self-interviewing in round 10, some changes in means are observed, but not for variances and covariances. There are approximately 25 variables in the ESS for which the change in standardized means that comes with the change of mode is larger than about 0.20, which is considered to be a small effect in intervention research. Overall, our findings are positive for the ESS's transition to self-completion in the sense that the number of variables for which mode-effects occur are limited, and they mainly affect means only.

There are two possible reasons why mode effects occur. Firstly, the switch to self-interviewing attracts different kinds of respondents than face-to-face interviewing. When selection effects — i.e., effects due solely to the composition of respondents in either mode – directly or indirectly are related to the variables of interest in the ESS, this can cause a change in means. A second reason why mode effects occur is due to measurement differences. In this study, we were not able to systematically isolate each of the two causes of mode effects. There is also the possibility that differences that we find are caused by unknown country-specific factors affecting the group of countries that opted to use self-completion at round 10 . It appears that these country-specific factors, if present, are unlikely to lead to large changes, since we see that countries that use face-to-face interviewing in both round 9 and 10, on aggregate, did not experience changes on means, variances and covariances. Nevertheless, there remains some uncertainty of the causes we find for the mode effects.

For variables where we found relatively large mode effects, we evaluated whether these are consistent across countries and whether they can perhaps partly be explained by the fact that within self-interviewing, both the paper and web interview mode are used. We find first that mode effects are pretty consistent across countries. This finding is helpful, because it means that any future approach to try and reduce the mode effect due to changes in fieldwork or questionnaire design do not have to be country-specific. The finding does not help to solve the issue to what extent mode effects can be explained by selection or measurement effects. It is quite possible that selection effects are consistent across countries, and indeed, earlier experiments in the ESS in the context of the CRONOS panel found some evidence for this (e.g. Maslovskaya & Lugtig, 2022). When the results for the countries using self-interviewing in round 10 are further split by paper and web responses we also find quite big differences in means between the two self-completion modes. To some extent, differences can here be explained by selection effects.

The transition towards self-completion that the ESS is taking on in the next 5 years will benefit from more detailed assessment of these mode effects for individual questions. From a theoretical perspective, we could expect some types of questions (e.g. with long answer scales, or those sensitive to socially desirable answer behavior) to exhibit stronger mode measurement effects. It would be worthwhile to systematically code question characteristics based on theories on mode measurement differences (presentation, communication, answering process) and then relate these question characteristics to the size of the observed mode effect. This would pinpoint the cause of mode effects in more detail, and also provide clues about what to potentially do to reduce mode measurement effects.

Related to this, this report did not answer the question of how mode effects can be resolved. In some cases, a review of question content and cognitive testing could provide in-depth information on whether measurement effects are present, and whether these are perhaps amenable by changing the ESS questionnaire. It is also worthwhile to spend more energy on understanding selection effects and nonresponse bias in countries where the ESS is making the switch from face-to-face to self-interviewing. Experiments where the assignment of mode at the level of the sample would be useful for further understanding selection and nonresponse bias, especially when information is available on all sample members. Experiments where the interview mode is assigned to a respondent after he/she has agreed to participate would be very useful to investigate measurement effects in more detail.

References:

- Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. Psychological bulletin, 92(2), 490.

- Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. International Journal of Public Opinion Research, 21(1), 111-121.

- Kelley, K., & Preacher, K. J. (2012). On effect size. Psychological methods, 17(2), 137.

- Klausch, T., Schouten, B., Buelens, B., & Van Den Brakel, J. (2017). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. Journal of Survey Statistics and Methodology, 5(4), 409-432.

- Maslovskaya, O., & Lugtig, P. (2022). Representativeness in six waves of CROss-National Online Survey (CRONOS) panel. Journal of the Royal Statistical Society Series A: Statistics in Society, 185(3), 851-871.

- Lugtig, P. (2024b) ESS round 10 mode experiments in the United Kingdom and Finland. Findings on mode effects. ESS internal report, July 2024

- Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2014). Evaluating mode effects in mixed-mode data through the back-door and front-door. Journal Of Official Statistics, 30(1), 1-21.