Telephone versus Online Survey Modes for Election Studies: Comparing Canadian Public Opinion and Vote Choice in the 2015 Federal Election

Charles Breton *University of British Columbia*
Fred Cutler *University of British Columbia*
Sarah Lachance *University of British Columbia*
Alex Mierke-Zatwarnicki *Harvard University*

Surveys conducted over the internet have become ubiquitous and reliable. However, internet surveys are generally not the probability samples that researchers prize, samples that, for many years, they got by using telephone and face-to-face designs. Nevertheless, the barriers to representativeness for internet surveys are coming down and telephone surveys are increasingly plagued by low response rates and sampling challenges (Sala and Lillini, 2015). We may have reached a point where internet surveys using commercial panels will satisfy the needs of academic researchers at least as well as telephone surveys (Ansolabehere and Schaffner, 2014; Pasek, 2016). There is also the matter of cost: internet surveys are 5 to 15 times less expensive than comparable phone surveys and over 50 times less expensive than face-to-face, giving the internet mode a big advantage in terms of statistical power. Internet surveys also offer a much greater set of potential stimuli, such as images and video, easier experimentation and varied question types. So the mixed-mode 2015 Canadian Election Study (CES), which

Charles Breton, Department of Political Science, University of British Columbia, 1866 Main Mall, Vancouver BC, V6T1Z1, email: cbreton@mail.ubc.ca (corresponding author)
Fred Cutler, Department of Political Science, 1866 Main Mall, Vancouver BC, V6T1Z1, email: fred.cutler@ubc.ca
Sarah Lachance, Department of Political Science, 1866 Main Mall, Vancouver BC, V6T1Z1, email: s.lachance@alumni.ubc.ca
Alex Mierke-Zatwarnicki, Department of Government, Harvard University, 1737 Cambridge St., Cambridge MA 02138, email: amierkezatwarnicki@g.harvard.edu

Canadian Journal of Political Science / Revue canadienne de science politique
Page 1 of 32 doi:10.1017/S0008423917000610

combined telephone, internet and mail modes, is now typical of election studies.[1]

Election studies like the CES must optimize their design on sample size, cost and data quality. Investigators and funding agencies therefore need up-to-date evidence on these criteria. This paper's chief purpose is to compare telephone interview data (conducted by ISR-York) with web data (using online panellists purchased from Survey Sampling International (SSI)). A tangential benefit of the present study is the presentation of a significant amount of public opinion data from the CES.

Much of the literature on internet surveys goes out of date within a few years of publication. Most of the available internet-to-phone comparisons examine voluntary opt-in internet panels (see Pasek, 2016), which were common until recently. But the relevant comparison is now with commercial panels where the sample providers aggressively recruit respondent-panellists *and* provide meaningful incentives *differentially* so as to iteratively construct a large, diverse panel from which the most representative sample possible can be drawn. A second reason for another such study is that the comparability of the modes may be, to a large extent, particular to the phenomenon being studied. Though previous literature encompasses surveys on topics such as health and tourism, both *selection effects* and *measurement effects* may be specific to surveys on public affairs (Vannieuwenhuyze et al., 2010), especially ones that aim to measure sensitive or private political attitudes and behaviour.

Pasek (2016) reviews the considerations relevant to comparing survey modes. We refer the reader there for a full literature review on survey mode comparison (or see Bytzek and Bieber, 2016). There is a handful of published work that compares survey modes for surveys of political behaviour. Chang and Krosnick, with data from 2000, showed that the internet sample was "biased toward being highly engaged in and knowledgeable about the survey's topic (politics)" (2009: 641). They concluded that "the nonprobability Internet method yielded the most accurate self-reports from the most biased sample, while the probability Internet sample manifested the optimal combination of sample composition accuracy and self-report accuracy" (2009: 641). This seemed to be the state of affairs until quite recently, but the lay of the land in 2000 was quite different from what it is today given the diffusion of cellphones, the advent of smartphones, and the ubiquity and ease of web technologies.

The 2005 British Election Study used both face-to-face interviewing and an internet panel, finding little difference for understanding the determinants of vote choice (Sanders et al., 2007) and only slight differences in the marginal distributions of attitudes. Stephenson and Crête (2011) came to the same conclusion when comparing random-digit dialling (RDD) telephone and opt-in internet panel modes for the 2007 election in Quebec. Both studies found no mode differences with respect to hitting

Abstract. Election studies must optimize on sample size, cost and data quality. The 2015 Canadian Election Study was the first CES to employ a full mixed-mode design, aiming to take advantage of the opportunities of each mode while preserving enough commonality to compare them. This paper examines the phone interviews conducted by ISR-York and the online questionnaires from panellists purchased from a sample provider. We compare data quality and representativeness. We conduct a comprehensive comparison of the distributions of responses across modes and a comparative analysis of inferences about voting. We find that the cost/power advantages of the online mode will likely make it the mode of choice for subsequent election studies.

Résumé. Les études électorales doivent optimiser la taille des échantillons, leur coût et la qualité des données. L'Étude électorale canadienne de 2015 a été la première ÉÉC qui a adopté un plan à mode de collecte mixte, visant à tirer parti des possibilités de chacun des modes tout en préservant suffisamment d'éléments communs pour permettre la comparaison. Cet article examine les interviews téléphoniques menées par l'Institut de recherche sociale (IRS) de l'Université York et les questionnaires des répondants en ligne achetés auprès d'un fournisseur d'échantillons. Nous comparons la qualité des données et la représentativité. Nous effectuons une comparaison complète des distributions des réponses selon les modes et une analyse comparative des inférences au sujet du vote. Nous constatons que du point de vue des avantages coûts-efficacité, il est probable que le mode en ligne représentera le mode de choix des études électorales ultérieures.

the actual election result. The 2009 German Longitudinal Election Study data found that the opt-in online panel did the poorest job matching the election results but produced identical conclusions about the determinants of vote choice (Bytzek and Bieber, 2016). The most informative study for our purposes used data from 2010 to compare phone, internet, and mail surveys about politics in the US. Ansolabehere and Schaffner asked: "Does survey mode still matter?" The answer was a resounding "no": "researchers will not consistently get more accurate results, nor reach substantially different conclusions, when using one mode relative to another" (2014: 301). In fact, that study's *telephone* survey produced the highest total error on the marginal distributions of key variables that were benchmarked to population figures.

Plan of the Paper

The weight of the recent evidence implies minimal differences between modes, so we do not posit any theories or hypotheses about differences. Like other studies, we compare the survey modes on "each of three types of inference—point estimates, relations between variables, and trends over time" (Pasek, 2016: 2). We add elements that have not yet appeared in the literature: a detailed description of the quality of the web data, comparison of their performance in capturing campaign dynamics and a wider range of benchmarks to which both the web and phone data can be

compared. First, we describe the surveys. Next we examine measures of data quality applied to the two modes and then compare the modes to known benchmarks. We then compare the survey modes with respect to campaign dynamics, political engagement, attitudes and correlational inferences. Finally, we give recommendations for future election studies.

Preview of the Findings

We find that:

1. The internet mode is so much less expensive per interview that it provides significantly greater statistical power.
2. There is no glaring difference on the quality of individual responses to questions.
3. The internet sample is slightly *better* than the telephone in its ability to match verified population quantities from the census and other Statistics Canada surveys.
4. The telephone mode suffers from a combination of:
   a. self-selection into the sample of citizens who are more engaged with politics and generally more optimistic and open to difference, and
   b. social desirability bias on responses to questions about disadvantaged groups.
5. There is no clear difference between modes on political attitudes more generally nor for models of vote choice.

The Surveys

The 2015 CES was a dual-mode telephone (RDD) and internet study, with respondents sampled and recruited independently.[2] We call these the phone and web surveys. Table 1 gives descriptive information. The surveys were conducted from shortly before the start of the campaign until election day and then reinterviews took place after election day.

The telephone survey has been the cornerstone of the CES since 1988, when it was first conducted from the Institute for Social Research at York University with a rolling cross-section design. The survey has been remarkably constant since. The typical design consists of a campaign period study (CPS), a post-election reinterview (PES), and then a mailback questionnaire (MBS). The number of completed interviews has typically been around 4000 (CPS), 3000 (PES) and 1300 (MBS). This relatively large sample size has been thought necessary to allow for analysis of campaign dynamics and analysis by province. The total cost in 2015 dollars has been about $90 per CPS-PES completed interview, with a total study cost between $250,000 and $400,000.

TABLE 1
Survey Details

					Duration (mins.)		Cost	
Mode	Survey	Completed	Quality	Period	Mean	Median	Per complete	Total
Web	Campaign	7181	6596	Aug. 11th Oct. 19th	48	16	$13.81	$55,600
	Post	4402	4024	Oct. 28th Nov. 13th	137	20		
Phone	Campaign	4202	4165	Oct. 1st Oct. 18th	17	16	$107.54	$319,380
	Post	2988	2970	Oct. 20th Dec. 23rd	24	22		

The cost per complete are calculated based on respondents who have completed both questionnaires (CPS and PES) and who passed the quality threshold.

ISR uses a random-digit dialling list purchased from a leading provider. It includes cellphones, but only accidentally, and is not constructed to be representative of the proportions of land and cell lines in the "population" of telephone numbers. Interviewers use CATI terminals and place calls without a predictive dialer. This typically results in a response-rate of around 40 per cent, well above the market-research industry norm of 10 per cent and close to Statistics Canada's nearly 50 per cent RDD response rate. In 2015, the CPS response rate was 37 per cent and the PES re-interview rate was 71 per cent.

The web survey used a sample purchased from Survey Sampling International (SSI), which sent invitations out to its panel and directed them to a survey instrument implemented by the CES on the Qualtrics platform. Web respondents were invited to do the post-election survey but not a third survey equivalent to the mailback survey. SSI uses extremely complex, proprietary methods to recruit internet panellists and one-off survey respondents and to weight the sample frame for maximal representativeness. It is not possible to compute anything like a response rate, since SSI offers multiple surveys at random to respondents. Incentives vary across SSI's subpanels, but generally have a value for surveys of this length between one and four dollars. Among respondents with good data quality, the re-interview rate for the 2015 web PES was 61 per cent.

Cellphones and Landlines, Merging and Weighting

Survey researchers now confront the fact that many people cannot be reached by landline. In 2013, Statistics Canada found one in five

households had no landline, rising to 60 per cent of households where all persons were under 35. Given this trend, most telephone surveys use dual-frame cellphone-and-landline lists. Unfortunately, the 2015 CES could not afford the cost of a cellphone-only sample list, so cellphones only occur randomly in the overall list. To study the consequences, we collected data on the phone respondents' type of phone line (cellphone, landline, or VoIP), as well as the web respondents' answers to the question: "Do you have a landline that you use to make calls?" Only 5 per cent of the phone sample did the interview via cellphone. But, far from being young and less politically engaged, they are three years older (mean = 59 years) and slightly *more* interested in politics on the 0–10 scale (mean = 7.5) than the others.

The SSI cellphone-only question is more useful. The proportion in the web sample without a landline is 25.6 per cent, which is probably very close to the population proportion, given that Statistics Canada estimated it at 21 per cent in 2013. In the web sample, the average age of those who do not use a landline is 9 years lower (mean = 39 years) and interest is a half-point lower (mean = 6.4) than those with a landline. We use this indicator below to try to separate selection from measurement effects and evaluate the seriousness of the problem of the phone CES not using a cellphone-only sample.

The web and phone data cannot be weighted identically due to their very different sampling methods. We can, however, apply post-stratification weights to age and gender, as well as, separately, province and household-size in the phone data. However, the phone data is significantly skewed older and the mean age remains three years too old even with weighting (See Figure A1 in the appendix). The web survey is weighted simply to provincial population proportions because its age and gender distribution is so close to the population.

Data Quality

Are the web data good enough to meet academic standards for an election study? Where possible we compare modes on equivalent criteria, but since the phone is the standard, the burden of proof is on the web. Table 2 presents indicators based on a growing literature on quality in self-administered surveys. We look for respondents with high levels of non-response, who take too little time, who responded to question batteries in a "straight line," or whose responses are contradictory or nonsensical.

We begin with time. Our judgment is that the CPS should never take less than 7 minutes and the PES never less than 10 minutes, even for a technologically adept, politically sophisticated respondent who reads quickly and has an excellent internet connection.[3] On the web, only 0.5 per cent

TABLE 2
Comparison of Data Quality

Criterion	Web (SSI)	Phone (ISR)
CPS Completed Questionnaire	94%	N/A
CPS < 5 min.	0.2%	N/A
CPS < 7 min.	0.5%	N/A
CPS < 10 min.	2.4%	N/A
PES < 10 min.	6.4%	N/A
Leader traits < 15 sec.	6.1%	N/A
Failed screener	16.4%	N/A
Vote for & never same party	0.9%	0.2%
DK/Refused > 2 on 22 CPS questions	21%	16%
DK/Refused > 10 on 22 CPS questions	2%	1%
Predicted DK/Refused on 22 CPS questions	1.7%	1.2%
No leader feelings	11%	4.6%
Party feelings std. dev.	27.5	24.6
Leader feelings std. dev.	26.5	24.9
Straight line on at least one leader's traits	20.4%	N/A
Straight line on spending in policy areas	2.9%	4.3%
All missing on thermometer battery	3.2%	2.8%
Missing on two of dem_sat, econ_ret, education	2.7%	0.7%
Recommended Discard Data	13.1%	0.88%

Percentages are based on completed questionnaires unless indicated.

completed the CPS in less than 7 minutes, while 6.4 per cent took less than 10 minutes for the PES. For time-on-questions, we look at leader traits presented on consecutive screens. Processing these and thinking about a response should take at least five seconds per screen, such that being under 15 seconds in total (20 in Quebec) should indicate a respondent not making a genuine attempt to answer thoughtfully. The data show that 6.1 per cent of web respondents used less than 15 seconds.[4]

On non-response ("don't know," "refuse" or skip responses) the web data compare favourably with the phone survey: a nonlinear prediction model for 22 CPS questions common across the modes gives predicted values of 1.17 for the phone and 1.65 for the web. The percentage with more than two out of 22 non-responses is similar: 21 per cent on the web and 16 per cent on the phone. Web respondents *are* responding to the questions.

There are significant differences between modes in a few rows of the table. The percentage who had no response on the feeling thermometers for three major party leaders (Harper, Trudeau, Mulcair (ROC) / Duceppe (QC)) was 11.4 in the web data and only 4.6 on the phone. Obviously, a live interviewer promotes responses. Yet the web data seem to fare better on differentiation in the spending items.

Looking at the data quality measures altogether, we are optimistic that web data will be as good as phone data if the sample provider uses best practices for sample selection, monitoring and incentives. We also recommend that users sweep the data and discard respondents judged not to have taken the questionnaire seriously (Berinsky et al., 2014; DeSimone et al., 2014).[5]

We take a comprehensive, balanced approach to screening the data, and calculate for CES users a "discard" variable (for campaign and post-election waves separately) that allows removal of respondents who satisfied *one or more* of the following "failure conditions":

- Incomplete: Did not get to last 5 per cent of questionnaire.
- Complete CPS interview less than 7 minutes.
- No party feeling thermometer responses.
- No leader feeling thermometer responses.
- Less than 15 seconds on the three leader trait screens.
- Refused or don't know or skipped on two of *education*, *democratic satisfaction*, and economic retrospection.[6]
- Complete PES interview less than 10 minutes.

This results in a discard rate of 13.1 per cent in the web data and just under 1 per cent in the phone data.[7] In what follows, we use only the higher quality data, though it made little difference. The web data's large cost advantage makes discarding this much data unproblematic for statistical power.

Panel Attrition

Given the CES design as a pre- and post-election panel study, we also examine the reinterview rate and the extent to which selection bias is exacerbated in PES reinterviews. The phone has an 8 point advantage on reinterviews in the good quality data: 71.9 per cent (2841/3950) to 63.6 per cent (3917/6758). However, selection effects are strong and different in the two modes, pushing the phone post-election data even more in the direction of more educated, more attentive respondents. The web PES leans in the direction of older, higher income respondents. We show selection effects for education, knowledge, income, and age in the appendix (Figure A2). To illustrate, a 25-year-old respondent with only high school, no facts correct and the lowest income level has a reinterview rate predicted at 54 per cent in the phone and 39 per cent in the web, while a person with a post-graduate degree, all four facts correct, the highest income level, is predicted to be reinterviewed at 88 per cent in the phone and 78 per cent on the web. While panel selection bias seems worse on the web, the effect is driven by age, so simple weighting can

correct the problem. That is not possible for the variables influencing attrition in the phone data.

Representativeness: Comparing with Benchmarks

Comparing survey modes to objective benchmarks is critical. The idea is to assess surveys' accuracy on validated measures of population parameters (Ansolabehere and Shaffner, 2014). The benchmarks are constructed from sources such Statistics Canada or Elections Canada, documented in the appendix (Table A1).

We use the weighted data to calculate point estimates and 95 per cent confidence intervals after each variable has been dichotomized (Ansolabehere and Shaffner, 2014; Pasek, 2016). Our figures show point estimates as dots and the validated benchmarks as vertical lines.

In Figure 1 we see important differences between modes and benchmarks. Both samples over-represent people over 50 and those with a bachelor's degree, though the phone does this to a greater extent. The phone and web samples also both have a slightly higher proportion of Francophones than the population. And both modes oversample atheists and undersample those who consider religion very important. The phone has too many married respondents.

Previous CES have always had fewer unemployed respondents than the StatsCan figures. The two survey modes in our data are a fair distance apart on employment-unemployment. The phone is below the benchmark by 1.5 points and the web is above it by 2.7. This is perhaps not surprising given that the SSI web panel is composed of individuals who get rewards to participate in surveys and thus the unemployed might have strong reasons to do web surveys.

Excluding turnout, the biggest difference between our samples and the benchmark is on the variable measuring volunteering, but only on the phone. A measure of volunteering appeared on the CES in a deliberate attempt to gauge representativeness on a validated measure of social and community engagement. The benchmark comes from Statistics Canada's General Social Survey of 2013, a large phone survey with a very good response rate (46%).[8] Looking at volunteering in Figure 1, we see the web hits the population target while the phone overrepresents volunteers at 14 percentage points above the benchmark value (phone: 56%, web:43%, StatCan benchmark: 42%). That is, CES phone respondents are significantly more socially engaged than the general population, even after weighting for age. Social desirability bias cannot be driving this effect, as the benchmark is also a phone survey. This must be self-selection; volunteering is driven by factors very similar to political engagement and the phone survey is much more attractive to the socially and politically engaged.

Fig. 1 - B/W online, B/W in print

FIGURE 1
Comparing Modes with Benchmarks: Demographic and Politically Relevant Variables

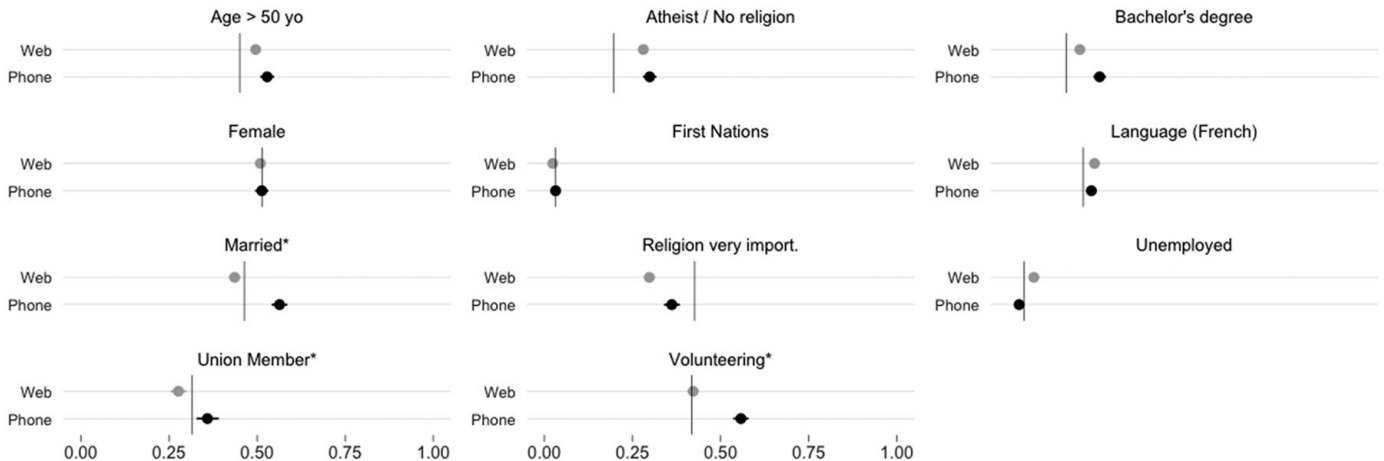


Figure 1 shows point estimates and 95 per cent confidence intervals for each mode after each variable has been dichotomized. These point estimates are shown as dots and the corresponding validated benchmarks as vertical lines.

Though both samples overestimate turnout, the web does a bit better, as seen in Figure 2. The phone is 25 points too high for both 2011 and 2015 turnout, leaving very little room to study factors influencing turnout. The web is 19 points high for the immediate reporting of 2015 turnout, but only 10 points high on 2011 turnout. The social desirability bias to misreport turnout must be stronger on the phone (contra Sanders et al., 2007: 264).

Finally, we see in Figure 2 that the modes are virtually indistinguishable on 2015 vote intention (labelled forecast), reported vote in 2015 (PES), and vote recall from 2011. Given the wide usage of the CES in studies of voting behaviour this is reassuring.

So far, we have examined each variable separately. To get a "final score," we now consider the total error for each survey mode. Like Ansolabehere and Shaffner (2014), we calculate the average difference for each mode as well as the mean squared error (MSE). The MSE, the average squared difference between each variable and its validated benchmark, is the most common measure of total survey error (see Biemer, 2010; Groves and Lyberg, 2010).[9] Table 3 gives the results, variable by variable and in total.

Table 3 only shows the difference between point estimates and the benchmark (see Appendix Table A.2 for the full results). Shaded cells indicate that the 95 per cent confidence interval included the benchmark value. The table also shows what happens when the web and phone data is pooled.[10] In sum, both modes offer relatively low total survey errors, though the web sample outperforms the phone. In this respect, we corroborate recent studies in political science that give the advantage to internet election studies.[11]

Campaign dynamics

Do phone and web surveys capture the same campaign dynamics in vote intention and attitudes? There is no scholarly literature on this question. One major virtue of the phone is controlling sample release for a rolling cross-section design (RXS, see Johnston and Brady, 2002). Web samples purchased from a provider with a proprietary recruitment system require compromises from a true rolling cross-section. Often, researchers cannot get detailed information on the procedures followed by the sample provider for sample release. The CES team asked SSI to target a certain number of completed interviews per day, which required adjusting the number of invitations sent out each day, sometimes as a reaction to previous days' number of completed interviews. Does this threaten our ability to make inferences about dynamics using the web data?

The daily sample sizes were not much different between the two modes, both averaging 100 to 150 per day. The web had a slight advantage

Fig. 2 - B/W online, B/W in print

FIGURE 2
Comparing Modes with Benchmarks: Vote, Turnout, and Vote Forecast

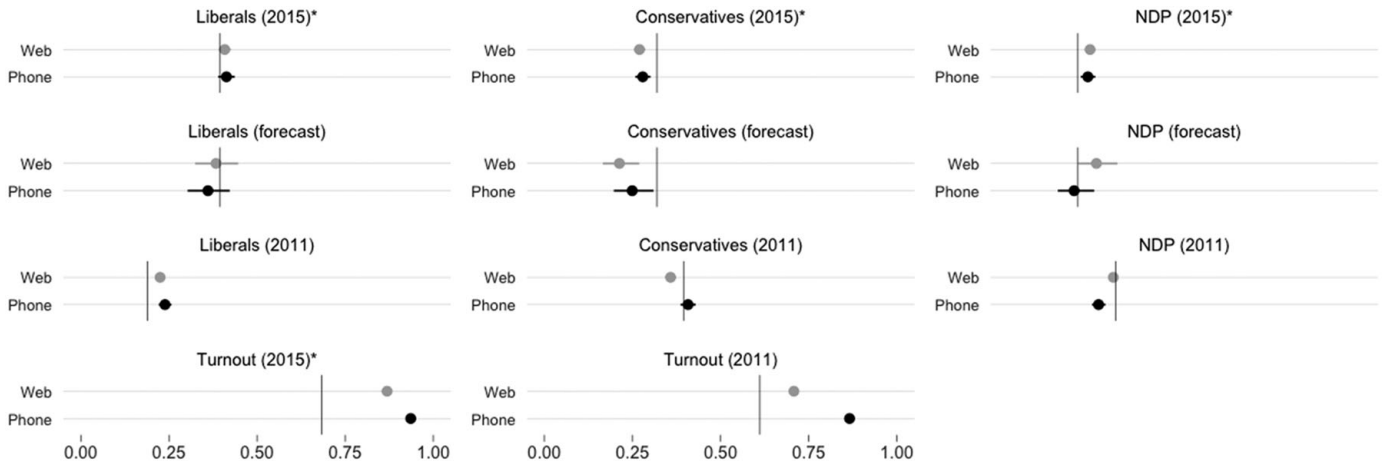


Figure 2 shows point estimates and 95 per cent confidence intervals for voting variables. These point estimates are shown as dots and the corresponding validated benchmarks as vertical lines.

TABLE 3
Differences between Modes and Benchmarks

Variable	Response	Web	Phone	Combined
Socio-Demographic				
Gender	Female	−.006	−.002	−.004
Age	50yo +	.045	.078	.057
Education	Bachelor's	.039	.095	.061
Married	Yes	−.028	.100	.025
First language	French	.032	.023	.029
Ethnic Group	First Nations	−.009	.000	−.005
Average difference		.027	.050	.030
MSE		.001	.004	.002
Other				
Volunteering	Yes	.004	.139	.060
Unemployed	Yes	.027	−.015	.011
Union member	Yes	−.038	.044	.000
Religion	No (Atheist)	.084	.103	.091
Religion Imp.	Very Important	−.128	−.065	−.103
Average difference		.056	.073	.053
MSE		.005	.007	.005
Vote				
Vote forecast	Liberal	−.011	−.034	−.024
2015	Conservatives	−.106	−.070	−.086
	NDP	.054	−.010	.019
Vote choice in	Liberal	.013	.018	.015
2015	Conservatives	−.049	−.040	−.046
	NDP	.036	.029	.033
Vote choice in	Liberal	.036	.050	.042
2011	Conservatives	−.038	.011	−.018
	NDP	−.007	−.049	−.025
Turnout 2015	Yes	.186	.253	0.225
Turnout 2011	Yes	.097	.255	0.152
Average difference		.061	.074	.062
MSE		.007	.013	.008
Total Average difference		.050	.067	.051
Total MSE		.005	.009	.005

The table shows differences between point estimates from the two samples and the actual value used as a benchmark. Shaded cells indicate values for which the 95% confidence interval includes the benchmark value. See appendix for benchmarks sources.

but its daily sizes were more volatile. Even the strictly controlled telephone rolling cross-section is far from perfect, however; it takes a while for the survey to ramp up, resulting in low power in the first week of the campaign, and the end of the campaign shows as much volatility as in the web survey.

Can the web mode at least match what we can learn from the phone RXS. As Pasek argues, "probability and nonprobability samples may not reflect the same distributions of attitudes and behaviours at any given

time, and, yet, they may reveal similar patterns of change over time… Yet… if aggregate changes are driven by a subset of the public rather than mass movement, trends from nonrepresentative samples could be misleading" (2016: 1).

While statistical testing is theoretically possible, graphical analysis tells us what we need to know. We show two figures depicting movement in key political variables: vote intention and feelings about the leaders, leaving feelings about the parties for the appendix. The figures cover only the period when the phone survey was in the field. The lines are produced by local regression of the (weighted) daily means. We do not present confidence bands, since both modes have very uniform and similar margins of error.

Overall, similarity across modes is the theme in these graphs. Apart from one notable divergence, they track the same forces over the campaign. Looking at vote intention first, we have something of a benchmark against which to compare our two modes. The solid lines in Figure 3 depict an aggregation of the polls, which amounts to daily samples well over one-thousand, using a variety of polling methodologies. It is widely accepted that an aggregation of polls is very likely to be very close to the true levels of support on a given day (Pickup and Johnston, 2008). Indeed the polls converged very close to the popular vote results.[12]

FIGURE 3
Vote Dynamics

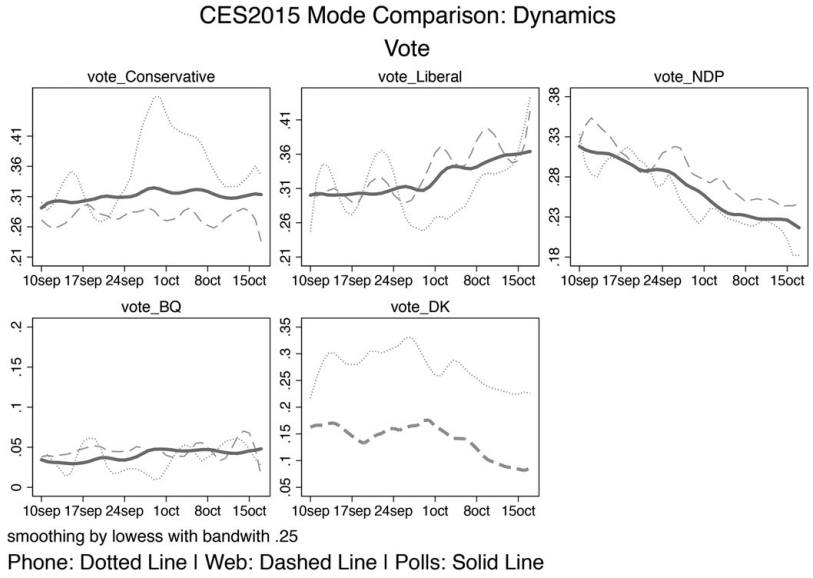


Fig. 3 - Colour online, B/W in print

While both modes track the movement in the polls, one key lack of correspondence is a cause for concern. The phone data show a dramatic 10-point rise in Conservative support at mid-campaign, only to fall back to the same level as the polls by campaign's end. The polls do exhibit a very slight Conservative rise that is not picked up in the web data. But the divergence from the polls and web data suggests that the phone is unrepresentative in some way for this period. In fact, it seems to sample Conservative partisans at a significantly higher rate (6–8%) over this ten-day period. Unfortunately, we cannot explain this quirk of the phone survey and note that, in view of the last eight CES studies, this is extremely atypical. We hope this quirk does not detract from the general point that both modes generally show the same dynamics.

For the other parties, the web and phone data seem to be quite similar and track the polls very well. The Liberal movement in the phone data lags the polls and the web, which is likely just the mirror image of the divergence in the Conservative share. Both modes track the severe drop in NDP support. Reassuringly, the decline in indecision shows voters making up their minds in both modes, though indecision is significantly higher overall in the phone data.

In Figure 4, showing Leader Feeling Thermometers, we see both modes picking up gains for Trudeau, a last-minute decline for Harper and

FIGURE 4
Leader Dynamics

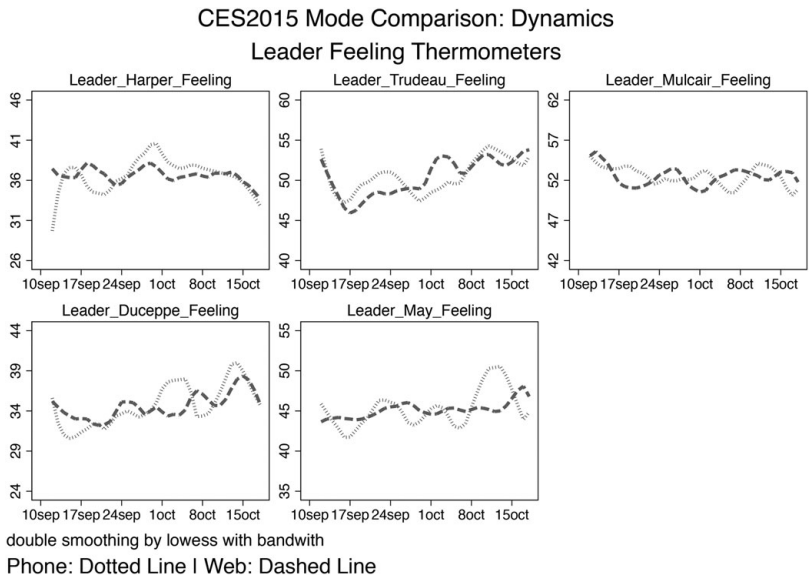


Fig. 4 - B/W online, B/W in print

Duceppe, and a modest two-stage decline for Mulcair. Again, the picture is of similar dynamics with the phone somehow producing more volatility. Our Party Feeling graph (appendix) has the same character, though movement is weaker, as it should be. One explanation for greater phone volatility is slightly smaller sample sizes.

We have little opportunity to formally test for similar dynamics in the web and phone data series because there was little movement on most measures included in both. There *was* significant movement on the question asking which party would be best for the economy. We estimated various time-series models that confirm the significant movement (particularly views on the Liberals vis-à-vis the NDP) but the models did not differ significantly by survey mode. Figure A3 in the online appendix shows how close the two modes are in tracking this attitude through the campaign.

The foregoing analysis constitutes the first evidence on the comparability of phone and web survey modes for capturing campaign dynamics. We are certain that neither mode is wholly superior on this criterion. However, the advantage in *daily* statistical power possible from a much-lower-cost internet sample makes it the obvious choice for campaign dynamics, especially if those dynamics are thought to be found only among particular subsets of the electorate (see Fournier et al., 2012).

Political Engagement

We now examine measures of political engagement: interest in politics, attention to the campaign and political issues, knowledge, participation and media consumption. As is clear from Figure 5, the differences are not enormous but phone respondents are significantly more interested in politics and reported more attention to the campaign. The same is true on attention to issues. Respondents in both samples say they are attentive to most issues, with Health Care at the top. For all seven issues, phone respondents report greater attention, with Defence and Immigration exhibiting the widest differences and Welfare the least.

On political knowledge, the CES contains four factual questions.[13] The two modes split the spoils here, each one on top for two questions out of four. The mean total of facts correct is significantly higher on the phone, but once we control for age the difference is non-significant. It is possible that some web respondents looked up the correct answer on the more difficult questions, despite the fact that they are paid panellists and have no real incentive to produce correct answers. This is one area where the phone may be slightly more accurate.

Media consumption and political participation are more tangible indicators of political engagement. Not surprisingly, phone respondents say

FIGURE 5
Political Interest and Attention

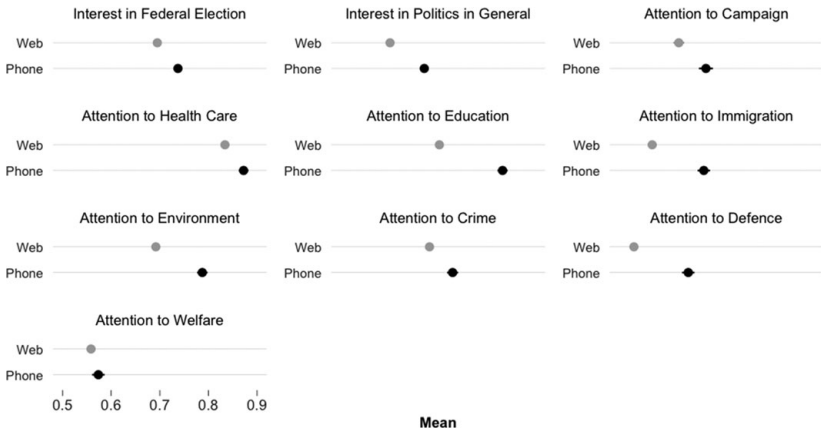


Fig. 5 - B/W online, B/W in print

This figure shows the mean value for each mode on questions measuring respondents' interest and attention to politics. Values for "Interest in the Federal Election" and "Politics in General" are scales that go from 0 to 10 where 0 is "No interest at all" and 10 is "A great deal of interest" and have been rescaled from 0 to 1. For level of attention, all measures have three response options that have been rescaled to range from 0 to 1 (No attention, 0; A little, .5; A lot, 1).

they consume more news in general. The key question asks "On average, how many minutes or hours per day do you spend watching, reading, and listening to news?" When we code the open-ended phone responses to the web categories, we find the distributions quite close, with the exception of more web respondents in the lowest 1–10 minutes per day category (Table 4). Notably, web respondents report higher levels of reading and exchanging news over the internet.

TABLE 4
Minutes per day reported news consumption

	Phone	Web
None	2%	3%
1–10 min	3%	12%
10–30 min	23%	24%
30–60 min	38%	29%
60–120 min	22%	20%
2 hr +	10%	11%
N (weighted)	2862	4531

We can also compare modes on political participation (aside from turnout discussed above) with the caveat that these are self-reports. In this case, differences in participation across survey modes vary according to the type of participation, but phone respondents, in general, say they participate more. While web respondents are more likely to sign petitions, to be politically active on the internet, and to take part in a marches, rallies or protests, phone respondents are more likely to volunteer and to boycott products for ethical or political reasons. These differences comport with our general picture of the compositional differences between the phone and web samples.

Political Attitudes and Social Desirability Bias

The CES survey measured a wide variety of political attitudes, from values to issue opinions to orientations toward groups. In the online appendix we provide a table with the distributions of these variables by mode. To quickly summarize, we find little difference across modes in domain-specific spending preferences, except for immigration. On a range of foundational and current policy issues, differences were mostly non-significant. We do see web respondents slightly more likely to want the Senate abolished, to think government doesn't care what they think, and to think politicians are ready to lie to get elected, but these differences all but disappear when controlling for age. These opinion results are very encouraging for future studies hoping to get more respondents per dollar by using the web mode.

On economic attitudes, however, there are some notable differences. Two-thirds of web respondents say that the Canadian economy has gotten worse, 12 points higher than the phone. The groups also differ in their attribution of blame for the economy: 48 per cent of web respondents claim that the current government has made the economy worse, compared to only 27 per cent of phone respondents.

Similarly, we see a big difference on respondents' levels of social trust. When asked the standard question ("Generally speaking would you say that most people can be trusted or that you need to be very careful when dealing with people?"), 54 per cent of phone respondents gave the trusting answer while only 31 per cent of web respondents did so. Neither the trust nor the economy differences disappear in the face of multiple socio-demographic controls. We suspect that social-psychological factors related to the difference in trust are responsible for much of the selection bias into the phone CES in the first place, and see this large attitudinal divergence as closely related to the difference in the behavioural measure of volunteering. These are important differences in sample composition that cannot be ignored.

When we turn to respondents' feelings about groups on a 0–100 "thermometer" scale we begin to see the most important systematic pattern of

differences on attitudes. Figure 6 displays the results for political parties while Figure 7 displays thermometers for different groups. There is little difference for the three major parties, little difference on Canada and the US, but large differences in feelings about marginalized or minority groups ("outgroups").

The obvious suspect is a social desirability bias where respondents who feel negatively about these groups are inclined to hide their feelings to a live interviewer (Chang and Krosnick, 2009; Kreuter et al., 2009). We therefore dig deeper into this set of questions, where the web and phone CES data would paint a different picture of Canadian public opinion. Recall that the samples differ on age, education, knowledge, and political engagement, all factors that have been shown to influence attitudes about minority and marginalized groups. Yet the fact that some feeling thermometer means are similar across modes rules out the possibility that any differences are inherent to the survey mode. Sample composition and social desirability bias on certain questions are likely jointly responsible for the differences across the two modes, so we try to disentangle the two where possible.

First, we look at questions related to immigration in Figure 8. Phone respondents are much more positive toward immigration and immigrants. The proportion of web respondents saying that Canada should admit fewer immigrants, spend less on immigrants and that immigrants take jobs from other Canadians is around 20 percentage points higher than on the phone. The fourth panel shows an even more striking difference. Asked about banning "Muslim women from covering their faces in

FIGURE 6
Feeling Thermometers: Political Parties

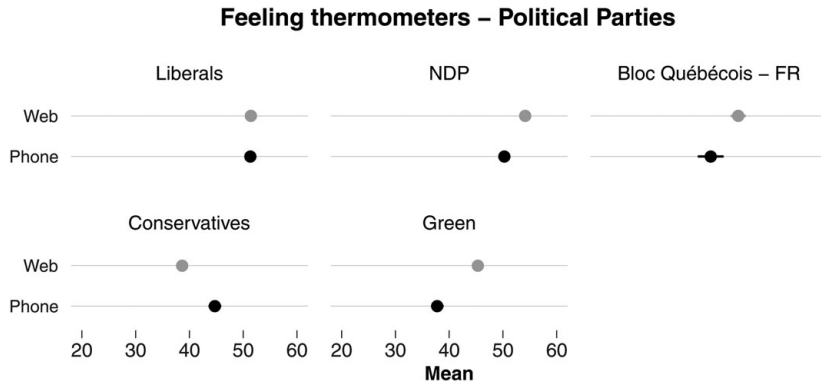
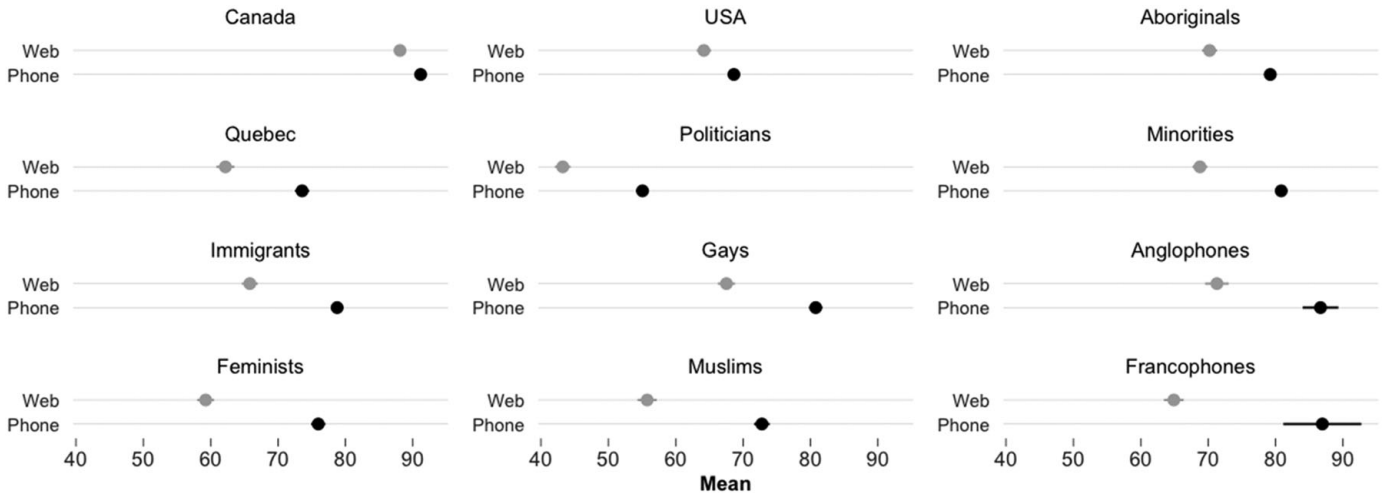


Fig. 6 - B/W online, B/W in print

This figure shows the mean responses to feeling thermometer where answers can range from 0 (Really dislike) to 100 (Really like). Facets are ordered based on the difference between the two modes. Bloc Québécois only includes Francophone respondents in Quebec.

Fig. 7 - B/W online, B/W in print

FIGURE 7
Feeling Thermometers: Groups



This figure shows the mean responses to each feeling thermometer where answers can range from 0 (Really dislike) to 100 (Really like). Facets are ordered based on the difference between the two modes. Feeling thermometers for Anglophones and Francophones only include respondents who are not "members" of that group.

Fig. 8 - B/W online, B/W in print

FIGURE 8
Immigration and Minorities

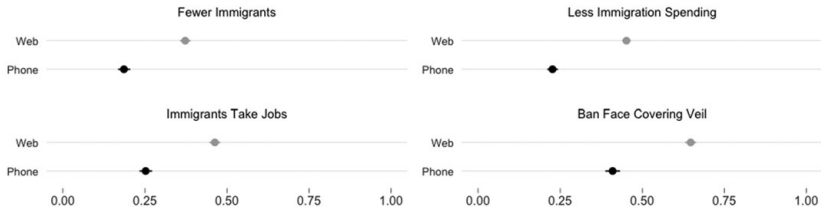


Figure 8 shows point estimates for each mode on questions about immigration and minorities. All answers have been dichotomized where 0 can be interpreted as meaning "No" and 1 as meaning "Yes". The only exception is "Ban Face Covering Veil" which was already dichotomous (Yes/No).

public" a majority of the phone sample opposes this while an even bigger majority of the web sample supports it.

Based on the responses to feeling thermometer questions, social desirability bias seems a likely culprit, but is probably not the whole story. Our attempt to sort out the causes is a regression of each of the thermometers on the variables that differentiate the samples—age, trust, volunteering—plus an indicator for mode. We also include a measure of whether the phone survey was done on a cellphone and whether the web respondent has a landline or not.

In the top line of Table 5 we see that the mode difference persists very strongly even controlling for the sample composition even though some of the compositional effects are quite large. The pattern of mode differences is the same as in the means shown in Figure 7: not much difference at all for Canada and the US, increasing from politicians and Aboriginals through minorities, gays and immigrants, to the biggest mode effect for Muslims and feminists. We are confident that social desirability bias is a big part of the mode difference and that the web mode measures these attitudes more accurately, as well as having more useful variance.[14] We have even more confidence because a separate model (not shown) confirms that the mode effect on the feminist thermometer is significantly weaker for women than for men.

In sum, while the phone and web groups give similar responses on many attitudes, there are very large differences in the marginal distributions on some critically important variables for our understanding of Canadian public opinion. These differences must be considered as scholars design multi-mode election studies. They might even call into question some of the findings from previous telephone studies.

TABLE 5
Sample Selection or Social Desirability

Feelings about groups by mode	Canada b/se	Quebec b/se	USA b/se	Politicians b/se	Aboriginal b/se	Gays b/se	Feminists b/se	Minorities b/se	Immigrants b/se	Muslims b/se
Web mode	**−2.30**	**−12.94**	**−3.65**	**−9.64**	**−8.80**	**−12.10**	**−15.92**	**−10.79**	**−10.73**	**−13.84**
	(0.53)	(1.07)	(0.87)	(0.95)	(0.91)	(1.04)	(1.04)	(0.87)	(0.92)	(1.11)
Pos/Neg wording	**−6.38**	**−4.26**	**−2.85**	0.41	**−4.69**	**−2.43**	−1.30	**−2.41**	**−3.00**	**−3.82**
	(0.58)	(1.08)	(0.88)	(0.96)	(0.97)	(1.10)	(1.06)	(0.92)	(0.99)	(1.15)
Age in years	**0.13**	**−0.14**	**0.14**	−0.03	**−0.04**	**−0.27**	**−0.13**	**−0.19**	**−0.17**	**−0.32**
	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
Most people can be trusted	**2.75**	**9.22**	**3.22**	**7.44**	**5.56**	**9.90**	**8.46**	**8.82**	**10.38**	**13.55**
	(0.44)	(0.87)	(0.70)	(0.76)	(0.74)	(0.84)	(0.86)	(0.70)	(0.74)	(0.88)
Volunteered	−0.02	**2.62**	0.38	**3.77**	**1.84**	0.17	**2.68**	**2.92**	**3.64**	**3.03**
	(0.45)	(0.83)	(0.68)	(0.75)	(0.72)	(0.82)	(0.82)	(0.68)	(0.74)	(0.87)
Cell only (web)	−0.80	3.39	**−3.70**	−3.35	**6.40**	3.30	**4.94**	3.84	2.76	**6.10**
	(1.30)	(2.89)	(2.01)	(2.21)	(2.18)	(2.38)	(2.83)	(2.37)	(2.38)	(2.70)
Have landline (web & phone)	−0.66	3.35	−2.29	−1.88	**5.54**	1.13	**4.78**	**4.11**	2.56	**5.33**
	(1.11)	(2.55)	(1.70)	(1.88)	(1.85)	(1.98)	(2.49)	(2.05)	(2.03)	(2.31)
Constant	**85.97**	**68.43**	**61.96**	**51.82**	**73.79**	**87.73**	**73.23**	**81.51**	**78.46**	**77.14**
	(1.34)	(3.02)	(2.02)	(2.30)	(2.20)	(2.35)	(2.94)	(2.50)	(2.50)	(2.79)
R^2	0.11	0.12	0.05	0.09	0.08	0.12	0.14	0.14	0.14	0.17
N	5151	4936	5064	5053	4896	4867	4880	4878	4912	4829

Coefficients in bold are more than 1.64 times their standard error.

Inference about Vote Choice

In sister studies, researchers have looked for differences in inferences in regression models across modes (Ansolabehere and Schaffner, 2014; Bytzek and Bieber, 2016; Sanders et al., 2007). The general conclusion has been that differences are rare. Obviously, the range of possible behaviours and the unlimited possible models explaining those behaviours make analyses illustrative rather than determinative. Bearing this in mind, we estimate a vote choice model for 2015 in both modes.

The challenge here is that vote choice models in multi-party Canada typically involve a huge number of coefficients. The most recent CES vote choice model (Fournier et al., 2013) included thirty independent variables per party. To be comprehensive we would have to compare literally hundreds of coefficients across modes, parties, and variables. Here, then, we simplify by estimating a model like the one Fournier and colleagues (2013) used for the 2011 election, do so only for the rest of Canada (ROC) without Quebec, and omit the Green party choice.[15] We simplify further by estimating logit models that follow the decision structure most voters would have used: one for the choice between the Conservatives and the other parties (an approval model), and one for choice between the NDP and Liberals.[16] We do not mean to ignore important choices; we merely present part of the choice to examine the differences across survey modes.

Table 6 presents the two estimations in two panels, with 35 coefficients for the Conservative choice and 37 for the NDP-Liberal choice. The panels' left columns show results from the modes pooled together. Of the 72 estimated coefficients, 54 show a significant effect in one or both modes. Of these, 22 are significantly different from zero in *only one of the modes*. This is a wider divergence across modes than has been found by other researchers. And the pattern of differences defies characterization. Table 7 shows what each mode would have missed: effects indistinguishable from zero in one mode but significant in the other mode or the pooled data. Notably, the most sensitive salient issues in the campaign—the face-veil ban and immigration—appear unimportant if we only use a phone sample. This is a very serious shortcoming—damning for the phone mode—and it is likely due to the bias discussed above.

No judgment can be made as to the superiority of one mode or the other for modelling vote choice. The R^2 statistics are very similar.[17] We cannot even say that one has certain tendencies. Clearly, both will produce interesting conclusions about vote choice, but even with these large samples researchers using each mode would tell different stories about what mattered and which were winning and losing strategies. This is troubling indeed. If a story must be told, we advise suggest pooling the data and letting the chips fall where they may.

TABLE 6
Vote Choice

Vote Choice 2015 by Mode: ROC					Vote Choice 2015 by Mode: ROC				
Conservative vs. Other					**Liberal vs. NDP**				
Probit with Marginal Effects	All marg eff	Phone marg eff	Web marg eff	Difference	Probit with Marginal Effects	All marg eff	Phone marg eff	Web marg eff	Difference
Atlantic Provs (d)	**−0.082**	**−0.071**	**−0.091**	0.02	Atlantic Provs (d)				
Western Provs (d)					Western Provs (d)	**−0.11**	**−0.147**	**−0.078**	−0.069
Female (d)					Female (d)				
Age <35 (d)	**0.041**	0.01	**0.056**	−0.046	Age <35 (d)	**−0.048**	0.051	**−0.082**	0.133
Union Memb in HH (d)	**−0.029**	**−0.039**	−0.02	−0.019	Union Memb in HH (d)	**−0.043**	−0.03	−0.058	0.028
Bible Word of God (d)					Bible Word of God (d)				
Francophone (d)	**−0.065**	**−0.073**	−0.038	−0.035	Francophone (d)	0.032	**0.087**	−0.019	0.106
Non-Eng/Fr First Lang. (d)					Non-Eng/Fr First Lang. (d)				
PID_Conservative (d)	**0.126**	**0.064**	**0.192**	−0.128	PID_Liberal (d)	**0.094**	0.055	**0.124**	−0.069
PID_Other_Parties (d)	**−0.119**	**−0.144**	**−0.094**	−0.05	PID_NDP (d)	**−0.194**	**−0.175**	**−0.222**	0.047
Feelings: Harper	**0.406**	**0.362**	**0.422**	−0.06	Feelings: Trudeau	**0.63**	**0.599**	**0.66**	−0.061
Cons. Best for Econ (d)	**0.244**	**0.26**	**0.225**	0.035	Feelings: Mulcair	**−0.633**	**−0.524**	**−0.774**	0.25
Income 5 categories	**0.105**	**0.124**	**0.091**	0.033	Income 5 categories	**0.109**	0.051	**0.206**	−0.155
Fragile Income (d)	**0.088**	**0.213**	0.064	0.149	Fragile Income (d)				
Nat. Econ. Worse (d)					Nat. Econ. Worse (d)				
Pers. Econ. Worse (d)	**0.037**	0.058	0.032	0.026	Pers. Econ. Worse (d)				
Spending on Left Issues	**−0.075**	**−0.078**	**−0.084**	0.006	Spending on Left Issues				
Spending on Right Issues	**0.036**	0.033	**0.042**	−0.009	Spending on Right Issues				
Feelings: USA	**0.085**	0.079	0.067	0.012	Feelings: USA	**0.097**	0.058	**0.132**	−0.074
Market Liberalism	**0.054**	**0.063**	0.039	0.024	Market Liberalism	**0.104**	**0.118**	**0.119**	−0.001
Feelings: Gays					Feelings: Gays				
Feelings: Politicians	**−0.179**	**−0.12**	**−0.205**	0.085	Feelings: Politicians	0.047	−0.018	**0.144**	−0.162

Increase Taxes (d)	**−0.1**	**−0.075**	**−0.108**	0.033	Increase Taxes (d)	**0.097**	**0.098**	**0.087**	0.011
Decrease Taxes (d)	−0.019	**−0.039**	−0.005	−0.034	Decrease Taxes (d)	**0.051**	**0.071**	0.024	0.047
No Guns	**−0.047**	−0.013	**−0.084**	0.071	No Guns				
2 Tier Health					2 Tier Health				
Terror Crackdown	**0.075**	**0.084**	**0.065**	0.019	Terror Crackdown	**0.078**	0.078	0.063	0.015
Fewer Immigrants (d)					Fewer Immigrants (d)	**−0.09**	−0.022	**−0.111**	0.089
Supp. Same-Sex Marr. (d)	−0.028	**−0.055**	−0.012	−0.043	Supp. Same-Sex Marr. (d)				
Supp. Climate Tax (d)					Supp. Climate Tax (d)				
Supp. Niqab Ban (d)	**0.037**	0.015	**0.052**	−0.037	Supp. Niqab Ban (d)				
Supp. Military v ISIS (d)	**0.076**	**0.073**	0.064	0.009	Supp. Military v ISIS (d)				
Supp. Deficit Stimulus (d)	**−0.146**	**−0.156**	**−0.128**	−0.028	Supp. Deficit Stimulus (d)				
Supp. Public Daycare (d)	**−0.031**	−0.004	**−0.049**	0.045	Supp. Public Daycare (d)	**−0.05**	**−0.079**	−0.019	−0.06
Shrink Income Gap (d)	**−0.063**	0.034	**−0.132**	0.166	Shrink Income Gap (d)	−0.028	**−0.104**	0.03	−0.134
					Lib. Best for Econ. (d)	**0.136**	**0.179**	**0.093**	0.086
					NDP Best for Econ. (d)	**−0.088**	**−0.141**	−0.042	−0.099
Pseudo R-Square	0.662	0.687	0.655		Pseudo R-Square	0.38	0.377	0.41	
Number of Cases	3487	1580	1907		Number of Cases	2258	1025	1233	

Estimated by Probit. Marginal effects (probabilities) shown when all other variables at means.

Bold indicates $p < 0.1$ (d) is Dummy variable

Difference column shaded darker the larger are the differences between modes.

Blank cells indicate insignificant coefficients in all three columns (modes). Blank variable labels indicate variable not included in estimation.

TABLE 7
Effects "Missed" by Mode

Conservative vs. Other		Liberal vs. NDP	
Phone	Web	Phone	Web
Age > 35	Union Member	Age < 35	Francophone
Spending on Right Issues	Francophone	Income	Decrease Taxes
Gun Ownership	Fragile Income	Feelings: USA	Public Daycare
Niqab Ban	Market Liberalism	Feelings: Politicians	Shrink Income Gap
Public Daycare	Decrease Taxes	Fewer Immigrants	NDP Best for Economy
Shrink Income Gap	Same-Sex Marriage Military vs. ISIS	PID Liberal	

The model presented here reflects common practice. But perhaps we should estimate simpler models (Achen, 2005). Just to be sure, we estimated a significantly smaller model with 19 and 17 variables respectively. Twenty-one of the 36 were significant in one or the other or the pooled data but again, 9 of these 21 were only significant in one mode or the other. The divergence seems to be built into the differences in sample composition and mode-specific measurement.

Assessment: Statistical Power per Dollar

Given the significant difference in cost between the two modes, the advantages of the web over the phone in terms of statistical power are clear. The fact that the web enables larger sample sizes at a lower cost is especially important for finding the sorts of effects that election scholars are often interested in. Questions about campaign dynamics usually involve effects that are relatively small and found in specific subgroups, making statistical power paramount. For instance, if one were to ask "Did feelings toward Gilles Duceppe change significantly over the course of the campaign among Francophones in Quebec?" the answer would vary by mode because the web has twice the sample size of the phone.[18]

To better illustrate the difference in power and cost, we go back to the 2011 election, where two campaign events were said to have had an impact on feelings toward leaders in Quebec and ultimately on the outcome of the election (Fournier et al., 2013). The first event is Jack Layton's appearance on the Quebec talk show "Tout le monde en parle," while the second is BQ's leader Gilles Duceppe appearing with Parti Québécois leader Pauline Marois at a PQ convention. According to Fournier and colleagues,

the effect of the former was quite large while Duceppe appearing with Marois had a relatively small negative effect. This is simply an example of cases where we would need different minimum sample sizes to detect these effects reliably.

Figure 9 shows a post-hoc power analysis of these two campaign dynamics. The two panels of this figure can be read as an answer to the question: Assuming that the effects found in the 2011 sample are the "true" effects, how much does it cost to find a statistically significant effect? To do this, we use the results from 2011 to produce 500 simulations at different sample sizes (from N = 100 to N = 8000) and calculate the proportion of these 500 simulations that find a statistically significant difference in feeling thermometers before and after each event for each of these sample sizes. A power of .8—finding a statistically significant difference 80 per cent of the time—is often considered the threshold for a sufficient statistical power. The sample sizes are then multiplied by the cost per complete from 2015 (web: $13.81, phone: $107.54) to form the x-axis.

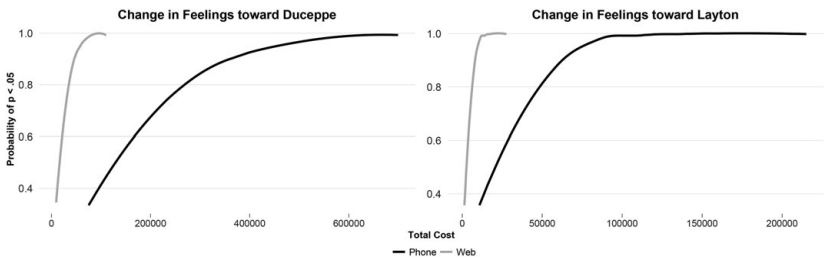
Even for an effect as big as the Layton one, the cost of achieving a power of .8 is about $48,000 for the phone but only $6,000 for the web. For the much smaller Duceppe-Marois effect the cost for the same power would be $247,000 using the phone and $33,000 with a web sample. In other words, even for large effects, the difference in cost is so important that it becomes difficult to justify having a phone sample from a cost perspective.

Discussion

We have highlighted differences between the phone and web modes of the CES 2015 and this leaves us perhaps slightly at odds with the relatively few and minor differences found in comparable studies in other countries.

FIGURE 9
Power and Cost

Fig. 9 - B/W online, B/W in print



The figure displays post-hoc power analysis for two campaign effects among Franchophones in Quebec in the 2011 electoral campaign. The total cost is calculated by multiplying the sample size by the cost per complete for each mode in 2015 CES.

We do not want to leave the impression that the differences are insurmountable, that either mode is to be abandoned entirely, that something has changed to produce larger differences than a few years ago, or that Canada is a special case. In fact, there are a great many commonalities across the two modes. It is impossible to say which is more representative, which better tracks campaign dynamics, which provides better estimates of engagement, knowledge and sophistication among Canadian voters, which paints a clearer picture of political attitudes, and which allows a richer or more accurate understanding of voter choice. This is probably a good situation, as it means researchers can use a mixed-mode design to address shortcomings in either mode (Dillman et al., 2014). We imagine, however, that the cost consideration will be dominant and election study teams will not be able to justify paying for the additional confidence provided by a large-enough telephone sample. Our findings support the conclusion that internet mode should now be the default for election studies. And our findings should embolden researchers who are using online-only studies of recent Canadian elections, such as the Local Parliament Project (Loewen and Rubenson, 2015).

Reviewing the findings, some important themes stand out. First, the internet mode is ten times less expensive per completed interview. As we have shown, the implications for statistical power, particularly among subgroups, are obvious. Moreover, the internet mode allows researchers to conduct more experimentation, to present richer stimuli and to examine a greater diversity of topics by running a core questionnaire with different add-on modules at random. The much lower cost and access to large panels would also allow targeted, incentivized interviewing of populations that are impossible to study with a telephone survey of 4000: Aboriginal people, youth, the unemployed or new Canadians, for example. We do not wish to be read as saying that the research community should accept the same CES for much less money. Instead, scholars should be thinking of the tremendous opportunities for small-group analysis, subtle campaign effects, powerful experimentation, and a longer interview period pre- and post-campaign that would all be possible for a cost similar to that of recent years (for instance, $300,000).

Second, it is clear that the telephone's day as the gold standard for election studies has passed. The phone data quality is still high, but even the very good response rate is falling, cellphones were omitted from the sampling frame, and the interview length is problematic. The relatively good data quality comes at the cost of representativeness: the phone mode suffers from massive self-selection effects. It over-represents politically engaged, socially integrated, optimistic, altruistic people. Answering the phone and spending 20 minutes, twice, answering questions about one's attitudes is now an unusually pro-social act.

On the other side of the ledger, questions about the quality and representativeness of commercial opt-in panels are diminishing. Firms like SSI

are concerned about, and take action toward, respondents who fail to take the surveys seriously and genuinely. In these panel providers' technical documentation they refer to behaviour of this sort as "fraud" and remove panellists immediately if it is detected. Indeed, the incentivization of participation may well, in combination with aggressive recruiting and maintenance, result in a *more* representative sample with more-than-acceptable data quality. That our vote model was not obviously superior in the phone data, and was in one important way worse, shows that most internet panellists respond in a serious, honest way.

The concern only a few years ago was that opt-in panels produced a sample that was younger, better educated, more male, and more affluent (Ansolabehere and Schaffner, 2014: 228; Shin et al., 2012: 217). In the 2015 CES it is the telephone mode that is too educated, too affluent, too socially integrated, and too politically sophisticated to be representative of the population. Notably, the CES had never adequately captured the unemployed and other disadvantaged populations; the internet mode now offers better representation of these citizens.

One possibility with the 2015 CES is to pool the web and phone data. If researchers accept both modes as valid measurement instruments there is no reason not to pool them. Some might argue that this throws the representativeness of the probability-sample phone study out the window, but we believe it is already long gone. Adding the web data actually pulls the phone data back close to the population benchmarks in most cases. For dynamics, pooling boosts daily statistical power. For inference, the pooled vote model looks more appealing to us, though who can really say which is closer to the true population model of vote choice?

The present study is the first to compare phone and web surveys for the purposes of understanding opinion dynamics, particularly over a short period like an election campaign. The RXS design continues to be central to the ethos of the community around the CES. We found that the internet sample was just as good on dynamics and if anything tracked the polls better than the phone survey. We require further analysis to learn whether the greater daily volatility in the phone mode comes from greater daily variance in representativeness or a sample that is much more attentive to the campaign. In the future, significantly larger daily samples from the internet mode could well improve understanding of campaign events and effects, as it has already done in other countries.

There is a great deal more work to be done to analyze these data from the 2015 CES. With the present paper and subsequent analyses it will be possible to provide a recommendation that maximizes the understanding scholars and the public can derive from election study surveys at a reasonable, sustainable cost. Given what we know right now, though, we are certain that future Canadian Election Studies will use the internet as their core survey mode.

Endnotes

1 The 2015 CES data are available at http://ces-eec.arts.ubc.ca/.
2 The CES also fielded an experimental study in another mode. Using an RDD list from the same source as the phone survey, respondents were recruited by telephone to do the web questionnaire. The usable N was 242. It was hypothesized that this technique would give better quality data than the internet mode at a much lower cost than the telephone interviewing from ISR-York. Even with a small sample we found that this was false. While the cost was about half the phone cost, the data quality was no better than the web data. The CES team thanks Parmida Esmaeilpour for her diligent management of this RDD-to-Web study.
3 We learned that one of the principal investigators of the CES timed himself doing the CPS survey quickly but honestly, with almost no need to read the questions, and took eight minutes to complete it.
4 Berinsky and colleagues (2014) recommended the use of screener questions. CES respondents were asked their favourite colour but instructed to ignore this and choose brown. The failure rate was 16.4 per cent. We examined a number of indicators of data quality conditional on failing this screener. In fact, a large majority of those who failed had completely sensible responses in general. We therefore do not use the screener to discard respondents.
5 Bytzek and Bieber's analysis of German election study data (2016) discarded 8.8 per cent of internet cases because they finished in less than 60 per cent of the average interview time.
6 These three variables are selected because they include a socio-demographic variable (education), an attitude (satisfaction), and a judgment of the state-of-the-world (economic retrospection) and appear at beginning, middle and end of the survey. All three are used extensively by researchers. None of them should be sensitive enough to motivate principled refusal. Two non-responses on these three variables probably indicates a respondent with little commitment to providing honest answers to all questions.
7 Our confidence in this measure is bolstered by the fact that it is not significantly different across any pair of provinces, nor by sex, and although it has a significant relationship with education, the maximum effect on the probability of being discarded is less than 1 per cent.
8 The data quality is excellent: the GSS has a reported 2011 federal election turnout of 71.1 per cent which is far closer to the true turnout than the CES phone survey.
9 For a more detailed examination of the logic behind the mean squared error as a measure of total survey error, see Biemer (2010).
10 Although the phone captures the benchmark on one more variable than the web (six and five respectively), the phone sample performs worse on almost all of the other variables. "Capturing" the benchmark can also be a product of a larger confidence interval.
11 Stephenson and Crête (2011) did not provide an MSE but we calculated it based on the results they provide. In their case, the phone sample does slightly better (.008 to .009). In the current study, if we only look at MSE for vote recall and include all five major parties, we find that the web fares much better than the phone (.005 to .011).
12 See https://en.wikipedia.org/wiki/Opinion_polling_in_the_Canadian_federal_election,_2015.
13 "Almost correct" answers are coded as correct: for instance, when there are numerous typos in a name or if it was pronounced incorrectly over the phone.
14 We confirmed that the distributions have a similar overall shape, with big pile-ups of responses (local modes) at 50 and 100.
15 Because the typical multinomial logit model suffers from the IIA property, our estimates of the choices between the other parties are, by definition, unaffected by the omission of the Green party.

16 In a formal sense this is not meaningfully different than a multinomial logit model (Alvarez and Nagler, 1998).
17 R2 is the only simple way to compare model fit for the same model in different datasets.
18 There is no statistically significant effect in the phone data ($p = .22$, $n = 649$), but there is one in the web data ($p<.01$, $n = 1310$), as well as when both dataset are merged ($p = .01$, $N = 1959$).

Supplementary materials

To view supplementary material for this article, please visit https://doi.org/10.1017/S0008423917000610.

References

Achen, Christopher. H. 2005. "Let's put garbage-can regressions and garbage-can probits where they belong." *Conflict Management and Peace Science* 22: 327–39.
Ansolabehere, Stephen and Brian Schaffner. 2014. "Does survey mode still matter? Findings from a 2010 multi-mode comparison." *Political Analysis* 22: 285–303.
Berinsky, A.J., Michele F. Margolis and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58: 739–53.
Biemer, Paul P. 2010. "Total survey error: Design, implementation, and Evaluation." *Public Opinion Quarterly* 74: 817–48.
Bytzek, Evelyn and Ina E. Bieber. 2016. "Does survey mode matter for studying electoral behaviour? Evidence from the 2009 German Longitudinal Election Study." *Electoral Studies* 43: 41–51.
Chang, Linchiat and Jon. A. Krosnick. 2009. "National surveys via RDD telephone interviewing versus the internet comparing sample representativeness and response quality." *Public Opinion Quarterly* 73: 641–78.
DeSimone Justin., A., P.D. Harms and Alice J. DeSimone. 2014. "Best practice recommendations for data screening." *Journal of Organizational Behavior* 36: 171–81.
Dillman, Donald A., Jolene D. Smyth and Leah M. Christian. 2014. *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. New York: John Wiley & Sons.
Fournier, Patrick, Fred Cutler and Stuart S. Soroka. 2012. "Who Responds to Election Campaigns? The Two-Moderator Model Revisited." Paper presented at the conference Duty and Choice: Participation and Preferences in Democratic Elections, Université de Montréal, January 20.
Fournier, Patrick, Fred Cutler, Stuart S. Soroka, Dietlind Stolle and Éric Bélanger. 2013. "Riding the orange wave: leadership, values, issues, and the 2011 Canadian election." *Canadian Journal of Political Science* 46: 863–97.
Groves, Robert M. and Lars Lyberg. 2010. "Total survey error: Past, present, and future." *Public Opinion Quarterly* 74: 849–79.
Johnston, Richard G.C. and Henry E. Brady. 2002. "The rolling cross-section design." *Electoral Studies* 21: 283–95.
Kreuter, Frauke, Stanley Presser and Roger Tourangeau. 2009. "Social desirability bias in CATI, IVR, and Web surveys the effects of mode and question sensitivity." *Public Opinion Quarterly* 72: 847–65.
Loewen, P. and D. Rubenson. 2015. "The Local Parliament Survey." http://www.localparliament.ca/ June 1, 2017).

Pasek, Josh. 2016. "When will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence." *International Journal of Public Opinion Research* 28: 269–91.

Pickup, Mark and Richard G.C. Johnston. 2008. "Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls." *International Journal of Forecasting* 24 (2): 272–84.

Sala, Emanuela and Roberto Lillini. 2015. "Undercoverage Bias in Telephone Surveys in Europe: The Italian Case." *International Journal of Public Opinion Research* 29 (1): 133–56.

Sanders, David, Harold D. Clarke, Marianne C. Stewart and Paul Whiteley. 2007. "Does mode matter for modeling political choice? Evidence from the 2005 British Election Study." *Political Analysis* 15: 257–85.

Shin, Eunjung, Timothy P. Johnson and Kumar Rao. 2012. "Survey mode effects on data quality: Comparison of web and mail modes in a US national panel survey." *Social Science Computer Review* 30: 212–28.

Stephenson, Laura B. and Jean Crête. 2011. "Studying Political Behavior: A comparison of internet and telephone survey." *International Journal of Public Opinion Research* 23 (1): 24–55.

Vannieuwenhuyze, Jorre, Geert Loosveldt and Geert Molenberghs. 2010. "A method for evaluating mode effects in mixed-mode surveys." *Public Opinion Quarterly* 74: 1027–45.