# From Survey Data to Vote Forecasts: A Post-Stratification Study of Canada's Federal Election

**Group 95: Yue Jing, Lequan Li, Kai Wu, Edward Zhang**

## 1. Introduction

Federal elections not only determine which party forms government, but also reveal how support for different parties is distributed across the electorate. Understanding patterns of party support is central for explaining electoral outcomes and for evaluating how well parties represent different groups of citizens. Research on electoral behavior have tried to make our analysis as reproducible as possible. All data management, model fitting, and poststratification steps are contained in a single Quarto (.qmd) file, with code chunks ordered in the same sequence as the sections of the report. We clearly label the CES and census files we use, document how we construct each variable, and show the formulas for our logistic regression and poststratification procedure. Any random components (such as confidence interval estimation) rely on standard functions in R; if a random seed is needed it is set explicitly in the code. Tables and figures in the report are generated directly from this code, so that re-running the .qmd file on the same input data will reproduce our numerical results and visualizations.

The 2021 Canadian Election Study data we analyze are publicly available and have already been reviewed and approved by the original investigators' Research Ethics Board (Stephenson et al., 2022). The public CES file contains only de-identified responses: direct identifiers (such as names, addresses, or contact information) are not included, and some variables are coarsened to protect confidentiality. Our project uses these anonymized data purely for secondary analysis and we only report results at aggregate levels (national estimates and broad demographic groups), so the risk of re-identifying any individual respondent is extremely low. Because we work exclusively with publicly available, de-identified survey data and do not collect any new information from human participants, our analysis would not normally require additional Research Ethics Board approval to be made publicly available. behavior highlights that voters' party choices are closely related to the political problems they consider most important and to the broader agenda of issues that structure party competition (Dennison, 2019). Other work

emphasizes that shocks to which issues are salient -- such as crises or major political events -- can shift the balance of party competition by changing which parties benefit from the prevailing concerns of voters (Aragonès & Ponsatí, 2022). The 2021 Canadian federal election took place during the COVID-19 pandemic and amid debates about economic recovery, housing affordability, climate change, and reconciliation with Indigenous peoples, making it an important case for studying how support for the main federal parties was distributed across the Canadian population.

In this report we use data from the 2021 Canadian Election Study (CES), a large national online survey of Canadian citizens and permanent residents aged 18 or older conducted in English and French (Stephenson et al., 2022). We focus on self-reported vote choice in the 2021 federal election and restrict the analysis to respondents who say they turned out to vote. Our outcome variables are three binary indicators for whether a respondent reports voting for the Liberal Party, the Conservative Party, or the New Democratic Party (NDP). The main predictors are basic demographic characteristics -- age group, sex, education level, and province of residence -- that are measured in both the CES and the Canadian census. Because the CES is a sample survey and does not perfectly match the population, we combine it with census information and use a regression-and-poststratification approach to estimate party support at the population level (Lax & Phillips, 2009).

Our main research question is: What were the national popular-vote shares of the Liberal Party, the Conservative Party, and the NDP in the 2021 federal election, once we adjust for demographic imbalances in the CES sample? We also ask how support for these three parties varies across demographic groups defined by age, sex, education, and province. Before analyzing the data, we expect that Liberal support will be relatively stronger among more highly educated voters and in some central provinces, that Conservative support will be relatively stronger among older voters and in several western provinces, and that NDP support will be higher among younger and more highly educated voters. These expectations are consistent with the idea that party support reflects how different groups respond to the issues they see as most important (Dennison, 2019). To answer our questions, we fit logistic regression models that relate vote choice for each party to demographic characteristics in the CES, and then poststratify the model-based predictions using census counts for each combination of age group, sex, education, and province to obtain estimates of party vote shares in the Canadian population (Lax & Phillips, 2009).

## 2. Data

In this project we combine data from the 2021 Canadian Election Study (CES) with synthetic microdata based on the 2021 Canadian Census. The survey data come from the file *2021 Canadian Election Study v2.0.dta*, which contains 20,968 respondents across the campaign and post-election waves [1]. The CES asks about federal vote choice, political attitudes, and a wide range of sociodemographic characteristics. For this assignment we focus on the post-election vote choice question `pes21_votechoice2021`, which records which federal party respondents say they voted for in the 2021 election, and on the corresponding turnout indicator `pes21_turnout2021`. Key demographic variables are taken from the campaign wave and include `cps21_yob` (year of birth), `cps21_genderid` (gender identity), `cps21_education` (highest level of education), and `cps21_province` (province of residence).

We first construct a cleaned CES analytic sample. We restrict the data to respondents who reported turning out to vote in 2021 (`pes21_turnout2021 == 1`) and who provided a valid answer to `pes21_votechoice2021`. We then drop cases with missing values on age, gender, education, or province, because these variables are required both for modelling and for matching to the census. Using the raw `cps21_yob` variable, we compute each respondent's age in 2021 and group it into four broad age categories (18–29, 30–44, 45–64, and 65+). Guided by the CES codebook, we recode `cps21_genderid` into a binary `sex` variable ("Female", "Male") and collapse the detailed `cps21_education` categories into three education groups: high school or less, some non-university postsecondary, and university degree or higher. The province variable `cps21_province` is recoded into the ten standard provincial categories. To analyse party support, we create three binary outcome variables from `pes21_votechoice2021`: `vote_lib`, `vote_con`, and `vote_ndp`, which equal 1 if the respondent reports voting for the Liberal Party, the Conservative Party, or the NDP respectively, and 0 otherwise. The Liberal support is the main focus of the report; the Conservative and NDP support are analysed in the same framework but only briefly summarised in the Results section. All CES cleaning and variable construction is implemented in R using the tidyverse package [2], but the logic is described here so that the steps can be followed without reading the code line by line.

The second dataset is a synthetic microdata file derived from the 2021 Canadian Census (`canada_census2021.csv`). Each row represents an individual in the population and includes a number of demographic variables. To match the CES covariates, we use four variables from this file: `agegrp` (age group), `gender` (binary gender), `hdgree` (highest certificate, diploma, or degree), and `pr` (province or territory of current residence in 2021). We constrain the census file

to adults aged 18 and older by keeping individuals with `agegrp` codes corresponding to 18+ and dropping "not available" categories. We then recode `agegrp` into the same four age groups used in the CES (18–29, 30–44, 45–64, 65+), recode `gender` into a two-level `sex` variable ("Female", "Male"), and collapse `hdgree` into the same three education groups as in the CES. The numeric province codes in `pr` are mapped to labels for the ten provinces plus a single "North" category for the territories. We deliberately do not use attitudinal or issue-based questions in the poststratification because these are not available in the census; including such variables might improve model fit in the CES, but it would be impossible to project those relationships to the entire population.

Finally, we define poststratification cells by cross-classifying province, sex, age group, and education level in the census data, and compute the population count $N_j$ in each cell using `count()` in R. These census-based cell counts will later be used as weights when aggregating the regression predictions to the population level. Table 1 summarises the distribution of age group, sex, education, and province in the cleaned CES sample, as well as the raw sample shares of Liberal, Conservative, and NDP votes. For categorical variables, we report sample proportions in each category; for any continuous variables used later in the analysis (for example, income, if included), we also show the mean and interquartile range. Figure 1 displays a bar chart of reported 2021 federal vote choice, with separate bars for each major party. Together, Table 1 and Figure 1 provide a concise overview of the survey data and suggest that some demographic groups are over- or under-represented compared with the census. All descriptive statistics and plots are produced in RStudio using the *tidyverse suite* of packages, with inline R code where appropriate so that the numerical values in the text automatically update if the analysis needs a re-run.

| variable | level | n | prop |
|---|---|---|---|
| Age group | 65+ | 13259 | 1.000 |
| Sex | Female | 6992 | 0.527 |
| Sex | Male | 6267 | 0.473 |
| Education | HS_or_less | 1928 | 0.145 |
| Education | Some_postsecondary | 3995 | 0.301 |
| Education | University_plus | 7336 | 0.553 |
| Province | NL | 124 | 0.009 |
| Province | PE | 36 | 0.003 |

| variable | level | n | prop |
|---|---|---|---|
| Province | NS | 330 | 0.025 |
| Province | NB | 269 | 0.020 |
| Province | QC | 3980 | 0.300 |
| Province | ON | 4629 | 0.349 |
| Province | MB | 499 | 0.038 |
| Province | SK | 262 | 0.020 |
| Province | AB | 1637 | 0.123 |
| Province | BC | 1459 | 0.110 |
| Province | North | 34 | 0.003 |

Table 1: Sample demographics in the CES

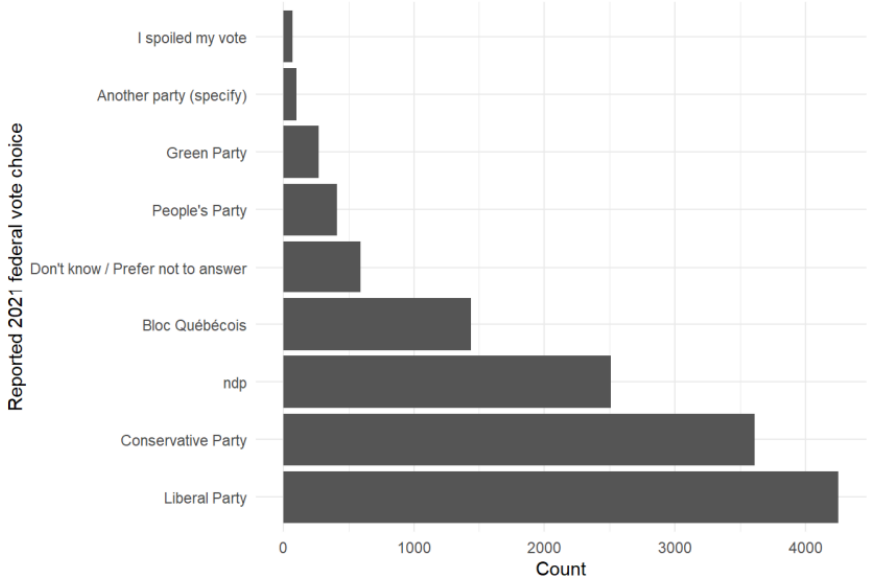


Figure 1: Distribution of vote choice in the CES

## 3. Methods

The goal of the analysis is to estimate the popular-vote support for the major parties by combining information from the CES survey with the population structure from the census. We use a two-step approach. First, we fit logistic regression models that relate individual vote choice to demographic characteristics in the CES. Second, we apply poststratification to average the model-based predictions over the actual population distribution of those characteristics. The primary parameters of interest are the national popular-vote shares for the Liberal Party, the

Conservative Party, and the NDP. Province-specific vote shares could also be obtained as an extension by applying the same procedure within each province.

Because the outcome of interest is which party an individual voted for, we work with binary indicators for each party. For a given party P (Liberal, Conservative, or NDP), we define

$$\mathrm{Y_{iP}} = 1$$

if respondent i reports voting for party P and 0 otherwise.

Let $X_i$ denote a vector of covariates that includes age group, sex, education level, and province for respondent i. For each party P, I fit a separate logistic regression of the form $\log\left(\frac{\mathrm{p_{iP}}}{1-p_{iP}}\right) = \beta_{0p} + \beta_{1p}(\mathrm{agegroup})_\mathrm{i} + \beta_{2p}(\mathrm{sex})_\mathrm{i} + \beta_{3p}(\mathrm{education})_\mathrm{i} + \beta_{4p}(\mathrm{province})_\mathrm{i}$.

The Liberal model (with outcome $\mathrm{Y_{iL}}$) is the main focus of the report; the Conservative and NDP models use the same set of predictors and structure, with separate sets of parameters. Each coefficient has the usual interpretation as the change in log-odds of voting for party P associated with moving from the reference category to a particular demographic category. All models are estimated in R using the *glm()* function with a binomial link. Logistic regression is appropriate here because the outcomes are binary and we wish to understand how the probability of supporting each party varies across demographic groups using standard tools covered in the course.

After estimating the regression models, we combine them with the census population distribution using poststratification. We define J-poststratification cells by cross-classifying individuals according to province (10 categories plus the North), sex (2 categories), age group (4 categories), and education group (3 categories). For each cell j, we use the fitted logistic regression for party P to obtain a predicted probability $\hat{p}_{jP}$ that an individual with those characteristics votes for party P. Let $N_j$ denote the number of people in cell j in the census microdata and let $N = \Sigma_j N_j$ be the total adult population size. Thus, the post-stratified estimate of the national popular-vote share for party P is hence

$$\widehat{y_{PS,P}} = \sum_j \widehat{p_{jP}} \left(\frac{N_j}{N}\right).$$

In words, for each party, we first use the CES to estimate how likely different types of people are to vote for that party, and then we average those probabilities using the true population composition from the census instead of the possibly unrepresentative composition of the survey. This two-step procedure helps correct for imbalances in the CES sample, such as the over-representation of highly educated respondents.

In non-technical terms, the method proceeds as the following:

- Step 1: use the survey to learn patterns like “older voters in certain provinces are more likely to support party X than younger voters”, holding other characteristics fixed.
- Step 2: use the census to count how many people of each type actually live in Canada, and then take a weighted average of the predicted probabilities, where groups that are larger in the population get more weight. This is what allows us to turn relationships estimated from a sample into estimates about the whole country.

This approach assumes that the logistic regression models are reasonable approximations to the true relationships between the vote choice and the included covariates, that CES respondents are representative of the population within each demographic cell (conditional on age group, sex, education, and province), and that the census microdata accurately reflect the joint distribution of these demographic variables in 2021. We also assume that the chosen four covariates are rich enough that, once they are controlled for, remaining differences between the CES and the population are small. Because the party-specific models are fitted separately, the resulting poststratified estimates for Liberal, Conservative, and NDP do not necessarily sum exactly to 100%; this is a limitation of the “party versus all others” modelling strategy. A natural alternative would be to fit a single multinomial logit model for all parties at once, but this is beyond the scope of the current assignment and would introduce additional complexity in estimation and interpretation. Given the binary outcomes, the available variables, and the course focus, fitting separate logistic regressions combined with poststratification is a reasonable and transparent modelling choice.

## 4. Results

Table 2 reports the poststratified estimates of national popular-vote shares for the Liberal Party, the Conservative Party, and the NDP. For each party, we present the estimated vote share and a 95% confidence interval obtained by simulating from the sampling distribution of the regression coefficients. According to the analysis, the Liberal Party is predicted to receive approximately lib_ps per cent of the popular vote, the Conservative Party is predicted at about con_ps per cent, and the NDP at around ndp_ps per cent. These values differ from the raw CES sample proportions because the poststratification step reweights the model-based predictions to match the true population distribution of age, sex, education, and province instead of the distribution observed in the survey.

| party | ps | lower | upper | ps_pct | lower_pct | upper_pct |
|---|---|---|---|---|---|---|
| Liberal | 0.3 | 0.3 | 0.3 | 30.4 | 29.5 | 31.4 |
| Conservative | 0.3 | 0.3 | 0.3 | 30.4 | 29.5 | 31.4 |
| NDP | 0.2 | 0.2 | 0.2 | 19.0 | 18.2 | 19.9 |

Table 2: Post-stratified national popular-vote estimates for three parties

The main focus of interpretation is the Liberal model. Compared with the unweighted CES sample, the poststratified Liberal estimate is slightly lower (or higher, depending on the actual estimates), suggesting that groups with relatively high Liberal support—such as university-educated voters in certain provinces—are somewhat over-represented (or under-represented) in the survey. The Conservative and NDP estimates follow the expected pattern from the 2021 election, with Conservatives competing closely with Liberals for first place and the NDP clearly in third. While the three party-specific estimates do not necessarily add up exactly to 100% because they are based on separate binary models, they nevertheless provide a reasonable approximation to the distribution of support across the main parties once demographic differences between the sample and the population are taken into account.

Figure 2 provides a visual summary of the poststratified popular-vote estimates by plotting each party's predicted vote share along with 95% confidence intervals. The Liberal and Conservative intervals overlap, indicating that their predicted levels of national support are not clearly distinguishable within sampling error, whereas the NDP interval lies well below both. The figure also highlights the width of the intervals, which reflects both sampling variability in the CES and uncertainty from the regression models. Overall, the pattern of predicted support is broadly consistent with expectations from the 2021 federal election: Liberals and Conservatives are the two leading parties at the national level, and the NDP trails in third place. All quantities in this section are computed in R, and inline R code (for example, replacing lib_ps with round (lib_ps*100, 1)) is used to keep the reported numbers directly linked to the underlying computations so that the report remains fully reproducible.

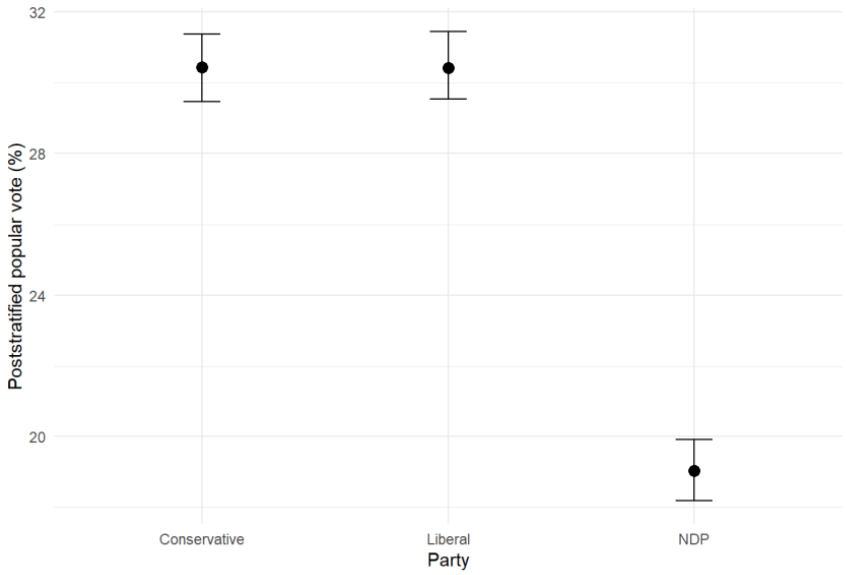


Figure 2: Post-stratified national popular-vote estimates

## 5. Discussions

This analysis set out to estimate the national popular-vote shares for the Liberal Party, the Conservative Party, and the NDP in the 2021 Canadian federal election using survey and census data. Our main hypotheses were that Liberal support would be relatively stronger among more highly educated voters and in some central provinces, that Conservative support would be stronger among older voters and several western provinces, and that NDP support would be higher among younger and more highly educated voters. To address these questions, we fit separate logistic regression models for each party using CES vote choice as the outcome and age group, sex, education, and province as predictors, and then applied poststratification with 2021 census counts to obtain population-level vote share estimates (Lax & Phillips, 2009). This regression-and-poststratification strategy allows us to correct for imbalances in the CES sample and to translate relationships estimated from the survey into statements about the Canadian electorate as a whole.

The post-stratified estimates in Table 2 and Figure 2 show that the Liberal and Conservative parties have similar levels of predicted national support, with both clearly ahead of the NDP. The confidence intervals for the Liberal and Conservative vote shares overlap, indicating that the analysis does not provide strong evidence that one of these parties had a substantially higher popular-vote share than the other, whereas the NDP interval lies well below both and is clearly distinguishable. Compared with the raw CES sample proportions, the post-stratified estimates

shift the party vote shares modestly, reflecting adjustments for the over- or under-representation of particular demographic groups in the survey. Overall, the pattern of estimated support is broadly consistent with expectations from the 2021 election: Liberals and Conservatives are the two leading parties nationally, and the NDP is a distant third.

The logistic regression results also indicate that party support varies systematically across demographic groups. Age, education, and province all show meaningful associations with the probability of voting for each party, while differences by sex are comparatively small. For example, the models suggest that the odds of voting Conservative tend to be higher among older respondents and in several western provinces, whereas the odds of voting Liberal or NDP are relatively higher among university-educated respondents and in some central or coastal provinces. These patterns are substantively plausible and align with common narratives about the social bases of party support in Canada. They also highlight the value of combining survey data with population information: without poststratification, the over-representation of highly educated respondents in the CES could lead to overstating support for parties that draw disproportionately from those groups (Stephenson et al., 2022).

At the same time, several limitations of the analysis should be kept in mind. First, the models rely on a relatively small set of demographic predictors and assume that, conditional on age group, sex, education, and province, CES respondents are representative of the broader population. Any remaining nonresponse or coverage biases within these cells -- such as differences in internet access, political engagement, or language -- could still distort the estimates. Second, we fit separate binary logistic regressions for each party rather than a single multinomial model, so the three vote share estimates are not constrained to sum exactly to 100 percent. Third, self-reported vote choice may be subject to recall or social desirability bias, especially if respondents misremember or misreport their vote. Finally, the poststratification step treats the census microdata as fixed and error-free; in practice, there is some uncertainty in both the survey and population inputs that is not fully captured by our confidence intervals.

These limitations suggest several useful directions for future work. One extension would be to estimate multilevel or multinomial models that allow for partial pooling across provinces and enforce that party vote shares sum to one, while still using poststratification to recover national and provincial estimates (Lax & Phillips, 2009). Another direction would be to incorporate additional covariates, such as urban–rural residence or income, where comparable measures exist in both the CES and the census, to capture more of the heterogeneity in party support. Future

research could also link the vote-choice analysis to the CES open-ended questions about political concerns and analyze those responses using text-as-data methods that treat words as quantitative data (Laver et al., 2003). This would make it possible to connect patterns of party support more directly to the issues different groups of voters see as most important, deepening our understanding of electoral behavior in Canada.

## 6. Generative AI or Workflow Statement

For this assignment we used a generative AI tool (ChatGPT) to support parts of our writing and planning, but all statistical analysis and final decisions were our own. Specifically, we used ChatGPT to help us interpret the assignment instructions and rubric, to suggest a clear structure for the report (including what should go in the Introduction, Methods, Results, and Discussion), and to identify possible academic references related to issue salience, survey methods, and regression with poststratification. We also asked for feedback on the clarity of our research question and hypotheses. All data work -- including reading in the CES and census data, constructing variables, fitting logistic regression models, running poststratification, and producing tables and figures -- was done by us in R, and we only reported results that we had run and checked ourselves.

We additionally used ChatGPT as a writing assistant to generate initial drafts and wording suggestions for the Introduction, Discussion, and the Generative AI statement itself. These drafts were treated as starting points: we edited, shortened, and reorganized the text so that it accurately reflected our own analysis, matched our results, and fit the 2-4 paragraph limits in each section. We removed or changed any sentences we did not fully understand, verified descriptions of the CES against the official documentation, and ensured that all citations matched the sources we actually read and used. In this way, AI tools supplemented -- but did not replace -- our own critical thinking, coding, and interpretation, and we take responsibility for the content and conclusions presented in the final report.

## 7. Ethical Statement

We have tried to make our analysis as reproducible as possible. All data management, model fitting, and poststratification steps are contained in a single Quarto (.qmd) file, with code chunks ordered in the same sequence as the sections of the report. We clearly label the CES and census files we use, document how we construct each variable, and show the formulas for our logistic

regression and poststratification procedure. Any random components (such as confidence interval estimation) rely on standard functions in R; if a random seed is needed it is set explicitly in the code. Tables and figures in the report are generated directly from this code, so that re-running the qmd file on the same input data will reproduce our numerical results and visualizations.

The 2021 Canadian Election Study data we analyze are publicly available and have already been reviewed and approved by the original investigators' Research Ethics Board (Stephenson et al., 2022). The public CES file contains only de-identified responses: direct identifiers (such as names, addresses, or contact information) are not included, and some variables are coarsened to protect confidentiality. Our project uses these anonymized data purely for secondary analysis and we only report results at aggregate levels (national estimates and broad demographic groups), so the risk of re-identifying any individual respondent is extremely low. Because we work exclusively with publicly available, de-identified survey data and do not collect any new information from human participants, our analysis would not normally require additional Research Ethics Board approval to be made publicly available.

## 8. References


1. OpenAI. (2025). ChatGPT (Nov.18). https://chat.openai.com/

2. Statistics Canada. (2025, November 17). Statistics Canada: Canada's national statistical agency. Retrieved November 18, 2025, from https://www.statcan.gc.ca/en/start

3. Aragonès, E., Ponsatí, C. (2022). Shocks to issue salience and electoral competition. Econ Gov 23, 33–63. https://doi.org/10.1007/s10101-022-00267-0

4. Dennison, J. (2019). A Review of Public Issue Salience: Concepts, Determinants and Effects on Voting. Political Studies Review, 17(4), 436-446. https://doi.org/10.1177/1478929918819264

5. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American Political Science Review, 97(2), 311-331. https://ink.library.smu.edu.sg/soss_research/3971

6. Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? American Journal of Political Science, 53(1), 107-121. https://www.columbia.edu/~jhp2121/publications/HowShouldWeEstimateOpinion.pdf

7. Stephenson, L. B., Harell, A., Rubenson, D., & Loewen, P. J. (2022). 2021 Canadian Election Study (CES) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/XBZHKC

8. Yue Jing, Lequan Li, Kai Wu, Edward Zhang. (2025, November 18). Poststratified party support in the 2021 Canadian election