Supplemental Information for *The Misreporting Trade-off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries*

Patrick M. Kuhn[*] Nick Vivyan[†]

Contents

[*] Associate Professor in Comparative Politics, Durham University. E-mail: p.m.kuhn@durham.ac.uk.

[†] Professor of Politics, Durham University. E-mail: nick.vivyan@durham.ac.uk.

A Further details on surveys and true turnout measurement

A.1 NZES survey and true turnout measurement

New Zealand Election Study (NZES) fieldwork was conducted by the Centre for Methods and Policy Applications in the Social Sciences (COMPASS) at the University of Auckland, beginning 27 September, four days after the general election of 23 September 2017. Respondents were contacted by mail and could respond either by completing a mailback questionnaire or by completing an online version of the survey. A reminder postcard was sent after two weeks and a second copy of the questionnaire mailed to non-respondents after about five weeks. The survey was mailed to all individuals who responded to the 2014 NZES who could be found on the electoral roll in 2017 and to a set of new individuals sampled randomly from the electoral rolls.[1] Sampling was stratified by age (18-31 and 32 or older) and electorate (general electorate and Māori electorate), with oversamples of Māori and the young. Overall, 10997 questionnaires were mailed out and the total number of responses was 3455. 1339 were respondents from the 2014 NZES and the remaining 2116 were new respondents. 3043 responses came via mailback questionnaire and 412 via the online version of the survey. Among 2014 respondents who were re-contacted, the response rate was 61.6 per cent. On a conservative basis (not removing any of the original sample for non-availability), the response rate among the new individuals sampled for 2017 (weighted to account for oversampling of some groups) was 30.6 per cent.

The 2017 NZES team validated respondent turnout by manual inspection of the marked electoral rolls. As the rolls themselves are the source of the NZES sample, matching errors between sample and roll are extremely unlikely. The computerized rolls contain age data in five-year bands, and gender can be assigned for over 99 per cent of names. To guard against inclusion of responses from persons not sampled, respondents' reported age and gender are checked against the roll data and where these do not match the record is removed from the data set. The marked rolls record who voted and who did not, what type of vote (ordinary, special, or overseas), and are recorded at polling places, centrally collected by the New Zealand Electoral Commission, and made available for public inspection after the election at 20 Electoral Offices located in cities and towns around New Zealand.

The 2017 NZES study design was reviewed and approved by the Human Ethics Committee of the Victoria University of Wellington (NZ) (Case number 0000025131). As in previous iterations of the NZES, 2017 respondents were not directly informed that their turnout would be verified via the marked electoral rolls but the NZES does report that turnout data is corrected from the rolls on its website. The marked rolls are available for public inspection in New Zealand. As in normal practice for de-identification, names and addresses were stripped from the dataset for analysis by the NZES team before the data was released for others to use and shared with the authors.

[1] The rolls are regularly maintained by the New Zealand Electoral Commission and continuously updated to a relatively high standard. In 2014, address accuracy of the writ day roll was independently estimated at 96.9 per cent (18 Justice and Electoral Committee, 2016). Enrollment is possible until the day before the election, with another 4 per cent of the coverage target added by then.

A.2 London survey and true turnout measurement

The London survey was fielded by YouGov to panelists recorded as residing in Greater London and the target population was the adult population of Greater London. YouGov maintain an online panel of over 800,000 UK adults (recruited via their own website, advertising, and partnerships with other websites) and hold data on the socio-demographic characteristics and newspaper readership of each panel member. Drawing on this information, YouGov uses targeted quota sampling, not random probability sampling, to select a sub-sample of panelists for participation in each survey. Quotas are based on the distribution of age, gender, education, social grade, party support, ethnicity and political attention in the British adult population. YouGov has multiple surveys running at any time and uses a proprietary algorithm to determine, on a rolling basis, which panelists to email invites to and how to allocate invitees to surveys when they respond.[2]

Due to the way respondents are assigned to surveys YouGov do not calculate a per survey participation rate. However, the overall rate at which YouGov panelists invited to participate in a survey do respond is 21%.

YouGov provided us with a file containing only the names and addresses of respondents so that we could locate respondents on the marked electoral register[3] and then merged our verification data into the survey responses before sending us the anonymized data stripped of all personal information. Thus at no point were we able to connect individual survey responses to identified individuals.

Overall, we visited the offices of 31 of the 32 London Local Authorities ('London boroughs') in which respondents reside. The 'missing' borough is Kensington and Chelsea. Officers from this authority informed us that they were unable to arrange access to the marked register during the data collection period due to the pressures of administrating a local election.

The overall number of London respondents for whom we have a definitive true turnout measure (i.e., who were successfully matched to the marked electoral register and whose true turnout was observable on the register) was 2595, which is 81.4% of all 3189 of the survey respondents and 82.4% of all survey respondents whose Local Authority office was visited. The rate of definitive true turnout measurements is lower than that obtained for the NZES. This is because YouGov do not sample directly from the electoral register (as was the case for the NZES), such that the resulting sample may contain respondents who are not on the register and respondents whose self-reported name and address is out of date or contains other errors meaning they cannot be located on the official register.

We recorded eight possible outcomes, as follows:

1. *Voted*: the named individual is found at the given address on the register clearly voted. This is a definitive validation outcome.

2. *Did not vote*: the named individual is found at the given address on the register and clearly did not vote. This is a definitive validation outcome.

3. *Not eligible*: the named individual is found at the given address on the register and was marked as not eligible to vote (e.g., underage or non-UK EU citizen). This is a definitive validation outcome.

[2] Any given survey thus contains a reasonable number of panelists who are 'slow' to respond to invites. Along with the modest cash incentives YouGov offer to survey participants, this is designed to increase the rate at which less politically engaged panelists take part a survey.

[3] This is the copy of the electoral register used at polling stations on Polling Day and which is marked to indicate when a listed elector has voted. Paper copies of the marked registers covering electors in a given Local Authority area are stored in Local Authority offices and are, for one year after Polling Day, available for in-person inspection.

4. *Absentee/proxy missing information*: the named individual is found at the given address on the register and is marked as an absentee voter or proxy voter, but the turnout records for such voters (which are stored in a separate file) were not available in the local authority in question. This is a indefinite validation outcome.

5. *Not at address*: the named individual was not found at the address given. This is an indefinite validation outcome, as the individual may have been registered at another address at the time of the election or may have incorrectly reported their address to YouGov.

6. *Address not found on register*: the reported address was not on the register. This is an indefinite validation outcome, as the individual may have incorrectly reported their address to YouGov.

7. *LA not attempted*: we did not attempt to (or in some cases were unable to) validate the marked registers for the Local Authority in which the individual is recorded as residing. This is an indefinite validation outcome.

8. *Other non-definitive*: This includes all remaining cases where we were unable to obtain a definitive turnout measure for an individual. Individuals in this group are mainly those who had address information recorded with YouGov which was either inconsistent, or insufficient to locate the polling station for which the marked register should be searched. Some others resided in electoral wards which had recently been reassigned to a different Local Authority, such that the register for these wards were not stored at the Local Authority offices at which we searched for them.

Table A.1 reports the frequency of each validation outcome.

Table A.1: Frequencies of Turnout Validation Outcomes, London Survey

Validation outcome	Freq	Percent
Voted	2244	70
Did not vote	226	7
Not eligible	125	4
Absentee/proxy missing info	0	0
Not at address	135	4
Address not found on register	281	9
LA not attempted	40	1
Other non-definitive	138	4

The design of the London study was reviewed and approved by the Ethics and Risk Committee of the School of Government and International Affairs (SGIA) at Durham University (UK). As with the NZES, respondents were not informed that we would attempt to validate their electoral turnout using the marked electoral register. However, the marked register is available for public inspection in the UK and, due to the data management process described above (where only YouGov could connect the data containing respondent names, addresses, and validated vote measures to the data containing respondent survey responses), at no point were the authors able to connect survey responses to identified individuals. Furthermore, informing respondents that electoral records would be checked to verify their turnout would have compromised our research design. First, it is likely to reduce direct question strategic misreporting by making respondents more concerned about being caught out lying about their turnout than they would be in a typical survey with no vote validation – in line with the so-called “pipeline to the truth” effect reported by Hanmer, Banks and White (2014). Second, respondents prone to strategically misreport may withdraw from the survey on learning their record would be checked. Either of these effects would have given the direct question in our studies an accuracy advantage that is not present in a typical survey on turnout.

B Direct turnout question responses are subject to strategic misreporting

Here we perform additional analysis of the London survey to provide further evidence that turnout is a sensitive topic. To do so, we exploit our measure of respondents' true election turnout and a measure of how comfortable or uncomfortable a respondent thinks they would feel revealing their turnout behavior in a survey. The latter measure comes from a question asked of those London survey respondents assigned to the sensitive list treatment group. The question was placed three questions after the list experiment question and was phrased as follows:

> `How comfortable do you feel revealing whether you did or did not vote in the last general election?`
>
> - `Extremely comfortable`
> - `Quite comfortable`
> - `Neither comfortable or uncomfortable`
> - `Quite uncomfortable`
> - `Extremely uncomfortable`
> - `Don't know`

We exclude 'don't know' answers and recode responses along a 1-5 scale such that 'extremely uncomfortable' corresponds to a minimum score of 1 and 'extremely comfortable' represents a maximum score of 5. We then use OLS to regress this measure of comfort revealing turnout on a binary indicator of whether a respondent actually voted or not in the general election. The results of this regression are reported in the first column of Table B.1. In the second column we add controls for respondent age group, gender and highest educational qualification. In both models true turnout is significantly and strongly positively related to a respondent's reported level of comfort revealing their turnout behavior. Average comfort answering the turnout question is around 0.6 units higher among actual voters. This estimated difference is around 3/4 of the standard deviation of self-reported comfort in the sample (0.84). These results are consistent with the notion that turnout is seen as a desirable behavior and that people therefore dislike admitting to nonvoting.

Table B.1: True Turnout and Respondent Self-Reported Comfort Revealing Turnout, London Survey

	(1)	(2)
Intercept	4.012***	4.172***
	(0.063)	(0.125)
True voter	0.606***	0.603***
	(0.068)	(0.068)
Demographic controls?	No	Yes
Observations	1,273	1,261
R^2	0.060	0.068
Note:	*p<0.1; **p<0.05; ***p<0.01	

C List experiment Diagnostics

In this Appendix we report standard list experiment diagnostics recommended in the literature.

First, to check that treatment assignment was not associated with respondent characteristics, we performed χ^2 tests of independence, reported in Tables C.1 and C.2 for New Zealand and London, respectively. Across all characteristics examined for each survey (for New Zealand, age group, gender, educational qualifications, and indicator for whether a respondent is registered in the Maori electorate; for London, age group, gender, educational qualifications and social grade) we failed to reject the null hypothesis of no association with treatment assignment. Thus, we are confident that our randomization was successful.

Table C.1: Randomization Checks: New Zealand List Experiment

Respondent attribute	Chi.sq	df	P-value
Age group	1.45	4	0.84
Gender	0.01	1	0.92
Qualification	3.31	2	0.19
Maori	0.10	1	0.75

Note: For each of the respondent attributes listed we conduct a χ^2 test of the null hypothesis of no association between the attribute and treatment assignment in the list experiment. Age group has five levels: 18-24, 25-34, 35-49, 50-64, 65+. Gender has two values: male and female. Qualification has three levels: highest qualification is school-level or below; highest education is post-school; highest qualification is university degree or above. The variable Maori measures whether or not a respondent is registered for the Maori electorate.

Table C.2: Randomization Checks: London List Experiment

Respondent attribute	Chi.sq	df	P-value
Age group	3.15	4	0.53
Gender	0.30	1	0.58
Qualification	2.03	3	0.57
Social grade	0.37	1	0.54

Note: For each of the respondent attributes listed we conduct a χ^2 test of the null hypothesis of no association between the attribute and treatment assignment in the list experiment. Age group has five levels: 18-24, 25-34, 35-49, 50-64, 65+. Gender has two values: male and female. Qualification has four levels: 'None/Other/Don't know', 'Level 1 to 2', 'Level 3', 'Level 4 or above'. Social grade has two values: ABC1, C2DE.

Second, we examine diagnostics for the 'no design effects' assumption, which states that respondents' reported item count for the control items is not affected by treatment assignment (Blair and Imai, 2012). Following Blair and Imai (2012), for both the New Zealand and London experiments, Table C.3 reports the estimated frequency of each respondent 'type', defined in terms of respondent control item count and sensitive item status (where 1 corresponds to voting). For both experiments, none of the estimated frequencies are negative, which would provide evidence of a violation of the 'no design effects' assumption (Blair and Imai, 2012). Furthermore, when we implement the Blair and Imai (2012, 63-65) formal test for design effects, we fail to reject the null hypothesis of no design effect for both New Zealand ($P > 0.99$) and London ($P > 0.99$).

Third, as recommended by Blair and Imai (2012), we check for possible 'floor' and 'ceiling' effects in each experiment by examining the observed frequency of item counts

Table C.3: Estimated Proportion of Respondent Types

	New Zealand		London	
	est.	s.e.	est.	s.e.
$\mathcal{J}$(0,1)	0.003	0.003	0.055	0.009
$\mathcal{J}$(1,1)	0.034	0.008	0.307	0.015
$\mathcal{J}$(2,1)	0.226	0.014	0.302	0.016
$\mathcal{J}$(3,1)	0.594	0.013	0.121	0.010
$\mathcal{J}$(4,1)	0.043	0.005	0.023	0.004
$\mathcal{J}$(0,0)	0.008	0.002	0.038	0.005
$\mathcal{J}$(1,0)	0.025	0.005	0.053	0.011
$\mathcal{J}$(2,0)	0.042	0.010	0.066	0.018
$\mathcal{J}$(3,0)	0.006	0.016	0.031	0.013
$\mathcal{J}$(4,0)	0.018	0.008	0.005	0.006

NOTE: The table shows the estimated proportion of respondents of each 'type' $\mathcal{J}(z, y)$, where $z \in \{0, ..., 4\}$ denotes the true number of affirmative answers to the control items and $y \in \{0, 1\}$ denotes true turnout.

in the control and treatment list conditions (Table C.4). Floor effects may occur among respondents who negate all control items and who are non-voters. Any such respondents assigned to the treatment condition would reveal their non-voter status if they answer the turnout item truthfully (giving an item count of zero). If this set of respondents recognize this and strategically misreport their turnout as a result, this will violate the 'no liars' assumption underpinning the analysis of list experiments (Blair and Imai, 2012) and will lead the list experiment to underestimate the prevalence of non-voting. Ceiling effects may occur among respondents who affirm all control items and who are non-voters. The list experiments used here do mask these respondents' answer to the sensitive item in the treatment condition (observing a reported item count of 4 out of 5 in the treatment condition, the researcher cannot know which item has been negated), so ceiling effects should be less egregious than floor effects. Nevertheless, respondents may worry that their non-voter status may be revealed by answering truthfully, and may therefore misreport their turnout, again violating the no liars assumption and leading to an underestimate of non-voting.

Analysis of the New Zealand control group in Table C.4 (a) suggests that the potential for floor effect is small – of 1,716 respondents in the control group, only 1% negate all control items – while there is some mild potential for ceiling effects – 6% of control respondents affirm all control items. Analysis of the London control group in Table C.4 (b) suggests that there is some mild potential for floor effects – of 1,563 respondents in the control group, 9% negate all control items – while the potential for ceiling effects is small – 3% of control respondents affirm all control items.

How concerned should we be about the potential for ceiling effects in New Zealand and floor effects in London? First, we note that the proportion of all-affirmers or all-negaters in the New Zealand and London list control groups, respectively, is lower than proportions of all-affirmers or all-negaters in the control groups of existing published list experiments (e.g., Blair, Imai and Lyall, 2014; Corstange, 2018; Kuhn and Vivyan, 2018). Second, the degree to which we should be concerned about the potential for ceiling effects in New Zealand and floor effects in London also depends on the degree to which non-voters in each experiment are disproportionately likely to give control item counts of four and zero, respectively. The more this is the case, the more non-voters will have an incentive to lie about their turnout status and the less the reduction of false negatives (relative to the direct question) the list experiment can achieve. Using true scores measures, we

Table C.4: Observed Counts by List Experiment Treatment Group

(a) New Zealand

		Count						
Group		0	1	2	3	4	5	All
Control	Frequency	20	101	460	1030	105	0	1716
	Percentage	1	6	27	60	6	0	100
Treatment	Frequency	14	48	129	393	1036	73	1693
	Percentage	1	3	8	23	61	4	100

(b) London

		Count						
Group		0	1	2	3	4	5	All
Control	Frequency	145	562	575	237	44	0	1563
	Percentage	9	36	37	15	3	0	100
Treatment	Frequency	59	170	585	523	198	36	1571
	Percentage	4	11	37	33	13	2	100

NOTE: Each table shows the frequency of reported item counts in the control condition and sensitive item condition of the list experiment.

find that only a moderate proportion of all actual-nonvoters in either the New Zealand and London surveys appear to be vulnerable to ceiling and floor effects, respectively. Using the true turnout measure, we find that around 5% of those actual nonvoters in the New Zealand list control group report the maximum item count (and would therefore be subject to ceiling effects in the treatment condition), while around 20% of actual nonvoters in the London list control group report the minimum item count (and would therefore be subject to floor effects in the treatment condition). This finding suggests that floor or ceiling effects are unlikely to explain the list experiment's poor classification performance relative to the direct question in both surveys we study.[4]

Fourth, because in both the New Zealand and London surveys we have a direct question measure of turnout for all respondents (recorded after the list experiment in New Zealand and a baseline measure recorded by YouGov prior to our survey in London), we can perform the placebo test developed by Aronow et al. (2015). This simultaneously tests the no design effect and no liars assumptions, as well as two additional assumptions posited by Aronow et al. (2015): a 'monotonicity' assumption, which states that respondents never falsely confess to the norm-defiant behavior in response to a direct question; and a 'treatment independence' assumption that list experiment treatment assignment (sensitive vs control list) is uncorrelated with direct question response.[5] The test involves generating a difference-in-means estimate of the sensitive item among those respondents who admit to the sensitive behavior when asked the direct question (so-called 'confessors'). Given our list experiment setup (where a zero response to the turnout questions indicates norm-defiance), the resulting difference-in-means should be statistically indistinguishable from zero if all of the above assumptions hold. At the 0.05 significance level,

[4] Further support for this conclusion when one considers that, whereas ceiling and floor effects only generate false negative errors (because they encourage actual nonvoters to misreport their turnout), further analysis of the confusion matrices presented in Table 2 shows that, and even if one eliminated all false negative errors from the New Zealand and London list experiments their overall accuracy would still be lower than the corresponding direct turnout question due to false positives among actual voters.

[5] The 'monotonicity' and 'treatment independence' assumptions are not strictly necessary for standard difference-in-means analysis of list experiments but are necessary for techniques developed by Aronow et al. (2015) for jointly exploiting direct question and list experiment responses to generate more efficient prevalence estimates.

we find this to be the case for both list experiments: for New Zealand the difference-in-means estimate is 0.12 with a P-value of 0.517; for London, the difference-in-means estimate is 0.2 with a P-value of 0.054.

Finally, we perform the Blair, Chou and Imai (2019) diagnostic designed to detect list experiment measurement error. This uses a Hausman specification test to check for large differences in the parameter estimates that result when applying maximum likelihood and nonlinear least squares models to list experiment data. We perform the test for an empty list experiment regression model with no covariates. In both the New Zealand (test statistic = 0.07; $df = 2$; $P = 0.97$) and London (test statistic = 0.52; $df = 2$; $P = 0.77$) list experiments, the Hausman tests fail to reject the null hypothesis of identical results. Thus for each list experiment this diagnostic yields little sign of measurement error, whether due to non-strategic misreporting or other types of misreporting.

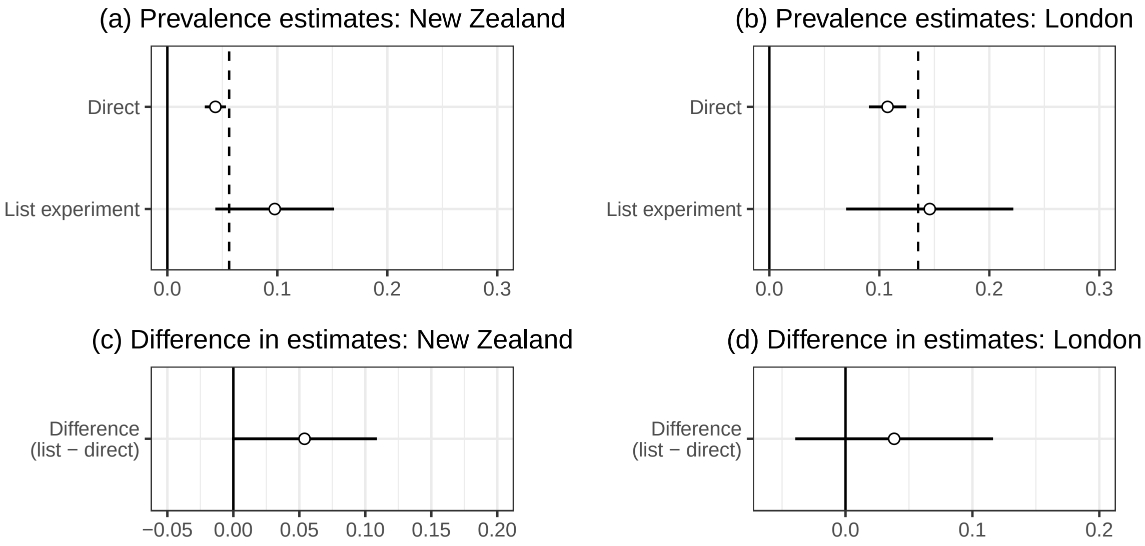
Overall, applying the key list experiment diagnostics recommended in the literature, there are no clear indications that the New Zealand and London list experiments violate key assumptions or are likely to yield problematic measures of the sensitive trait.

D Sample prevalence validation

Here we perform *sample prevalence validation* – i.e., comparing direct and list experiment nonvoting prevalence estimates against the true sample prevalence.

Subsetting to the subsample of New Zealand and London respondents with definitive true turnout measurements, Figure D.1 shows the nonvoting prevalence estimates resulting from the direct and list questions, with a dashed line representing the known true prevalence in the subsamples.

Figure D.1: Estimated Prevalence vs True Sample Prevalence



Note: Plots (a) and (b) show, for New Zealand and London respectively, direct and list estimates of non-voting prevalence in the subsample of respondents for whom we have definitive true turnout measures. Vertical dashed lines indicate the known true subsample non-voting rate. Plots (c) and (d) show differences between direct and list estimates.

The first notable feature of Figure D.1 is that, in both New Zealand and London, the true *sample* nonvoting rates are ten or more points lower than the corresponding true *population* nonvoting rates shown above in Figure 1. This illustrates the advantage of sample prevalence validation over population prevalence validation: both samples contain a disproportionate number of true voters compared to the population, meaning comparison of a prevalence estimate to the population prevalence benchmark can provide a misleading sense of which measure is better at eliminating reporting errors.

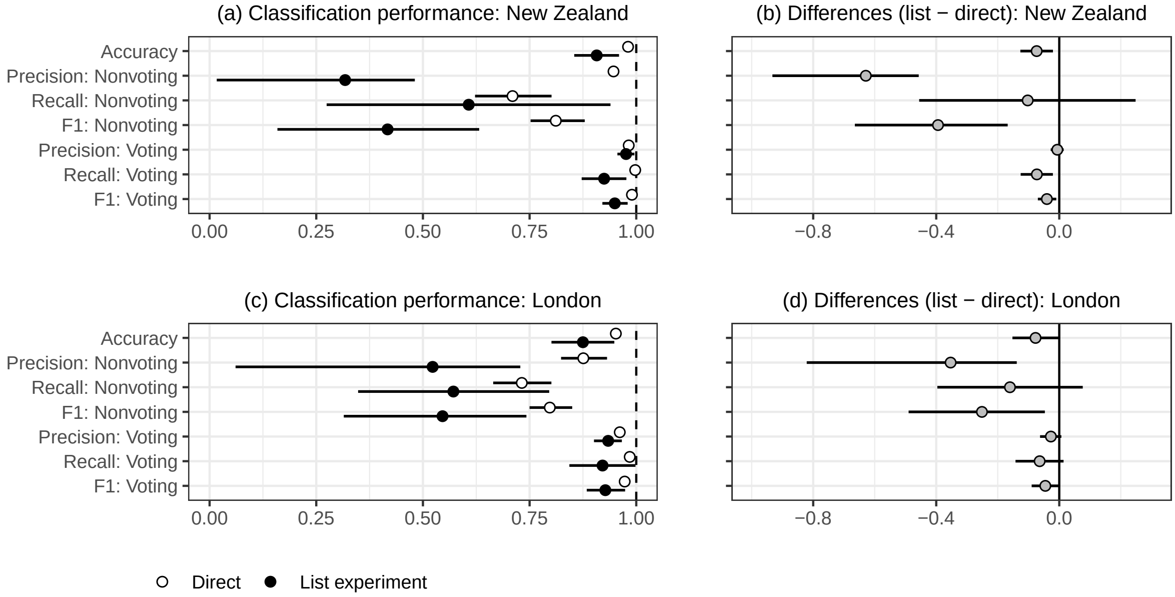
How well do the list experiments perform compared to the direct questions at recovering prevalence estimates close to the true sample benchmark? In New Zealand, the list experiment under-performs the direct question in terms of point estimates, as it overestimates the true sample nonvoting rate to a greater extent (4.1 points) than the direct question under-estimates the true sample nonvoting rate (1.3 points). In the subsample of London respondents for whom we observe true turnout, the list prevalence point estimate is closer to the true sample prevalence than is the direct question: it overestimates the true sample nonvoting rate to a lesser extent (1.1 points), than the direct question under-estimates it (2.8 points). However, plots (c) and (d) show that, in both the New Zealand and London surveys, the differences between the list and direct prevalence estimates are not statistically distinguishable from zero with 95% confidence. We therefore have somewhat contrasting findings for the two cases when it comes to sample prevalence validation. In New Zealand, there is some suggestion that the list experiment under-performs the direct question. In London, the list experiment slightly outperforms the direct question by point estimate.

But even these sample prevalence validation results could be misleading. It could be, for example, that in London the list experiment generates a greater estimate of nonvoting prevalence than the direct question – and therefore generates an estimate closer to the true sample prevalence – by increasing the rate of false positives rather than true positives. The partition validation results presented in the main text avoid this issue by separating true from false positives and true from false negatives.

E Alternative Measures of List Experiment and Direct Question Classification Performance

When norm-compliers heavily outweigh -defiers in a sample, a measure can achieve high accuracy simply by classifying all respondents as norm-compliers. In this appendix we therefore compare list experiment and direct question classification performance using three additional summary measures commonly used in the information retrieval and machine learning literature (e.g., Manning, Raghavan and Schütze, 2009, 154-157): *precision*, the fraction of measured positives that are true positives; *recall*, the fraction of actual positives that are classified as such; and F_1, which combines information concerning precision and recall, and is defined as the weighted harmonic mean of the two measures (we give equal weighting to precision and recall throughout our analysis). All three statistics require us to set an outcome on the sensitive variable that counts as a "positive". Below, we calculate each statistic twice: once with norm-defiers and once with norm-compliers as the "positive" outcome.

Figure E.1: Classification Performance of Direct and List Experiment Turnout Measures



NOTE: Plots (a) and (c) display, for New Zealand and London respectively, direct and list classification performance by various criteria. Dashed vertical lines indicate a perfect score. Plots (b) and (d) display differences in list and direct question performance for each criteria. Horizontal lines indicate bootstrapped 95% confidence intervals.

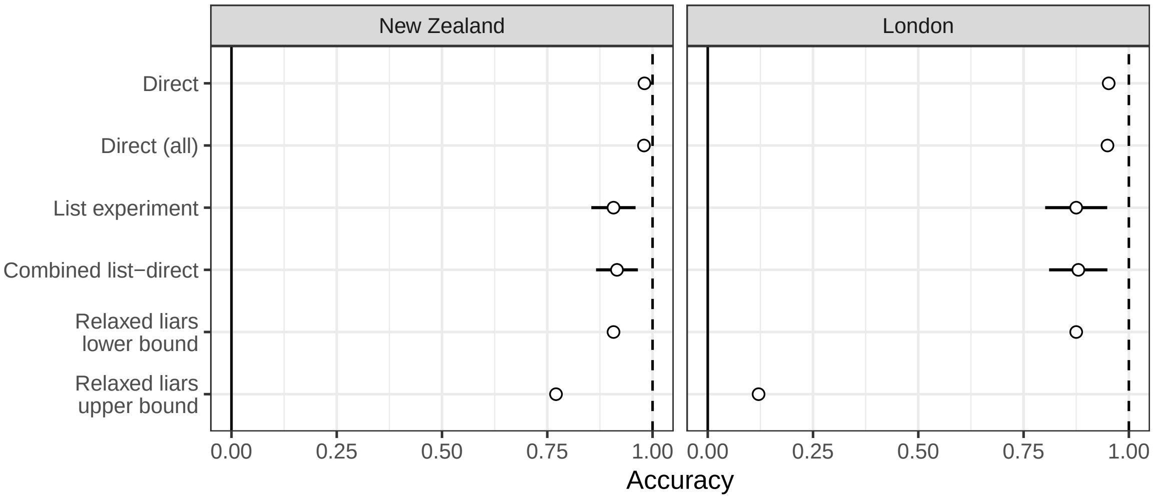
Figure E.1 summarizes overall direct question and list experiment classification performance, as measured according to accuracy, precision, recall and F_1 statistic, with the latter three defined with respect to actual nonvoters first, and actual voters second. For both New Zealand and London, the list experiment performs worse than the direct question by precision, recall, and F_1 statistic, whether these are calculated with respect to classification of nonvoters or voters. For London, the differences in list and direct question precision and F_1 for classification of nonvoters are statistically distinguishable from zero. For New Zealand, the differences in list and direct question precision and F_1 for classification of nonvoters are also statistically distinguishable from zero, as are the differences in list and direct recall and F_1 for classification of voters. Note in particular that recall of voting is worse for the list experiment than the direct question in New Zealand (the difference is distinguishable from zero with 95% for New Zealand but not London):

the finding that the list experiment classifies a lower fraction of actual voters as voters is consistent with the list experiment inducing additional non-strategic misreporting among actual voters (norm-compliers), compared to the direct question. Note also that the list experiments perform no better – and probably worse – than the direct question when it comes to recall of nonvoters (unsurprisingly given the relatively small number of actual nonvoters in each sample, 95% confidence intervals for the difference in list and direct question recall of nonvoters overlap wit zero in both New Zealand and London): i.e., they do not seem to reduce false negatives among norm-defiers, which is generally considered the main goal of list experiments.

F Alternative Direct and List Experiment Measures

In the main text we use partition validation to assess the classification performance of two turnout measures in each of our surveys: a direct turnout question measure based only on responses from the list experiment control group; and a list experiment turnout measure generated using standard difference-in-means estimation. Do our conclusions regarding relative performance of the direct and list experiment turnout measures change if we use alternative approaches to generate these measures based on the available data? Figure F.1 addresses this question. It shows, for the New Zealand (left panel) and London (right panel) surveys, the overall classification accuracy of a number of alternative direct and list experiment turnout measures.

Figure F.1: Accuracy of Alternative Direct and List Experiment Measures



Note: Displays accuracy for various alternative direct question and list experiment estimators of turnout for the New Zealand (left) and London (right) surveys, respectively. Horizontal lines indicate 95% confidence intervals, obtained via bootstrap.

The uppermost points in each plot again show the classification accuracy of our main direct turnout question measure: direct question responses asked only of list control group respondents. The second-top points in each plot show the classification accuracy of a direct turnout question asked of all survey respondents (recall that, for New Zealand, this ‘Direct (all)’ measure is based on the same question as the main direct measure, but includes responses from all respondents; for London, the ‘Direct (all)’ measure is based on a turnout question asked of all respondents shortly after the 2017 UK General Election and before our survey). Both versions of the direct question measure of turnout exhibit similarly high levels of accuracy: 98.1 points (‘Direct’ measure) and 98 points (‘Direct (all)’) in New Zealand; 95.2 points (‘Direct’ measure) and 94.9 points (‘Direct (all)’) in London. The third-top point in each plot corresponds to the difference-in-means list experiment turnout measure examined above. Consistent with results above, this measure has lower accuracy than either of the list experiment measures (90.7 points in New Zealand; 87.5 points in London).

Next we consider classification accuracy when, instead of standard difference-in-means to exploit the list experiment, one uses the estimator developed by Aronow et al. (2015), which ‘combines’ list experiment responses with information from the direct question asked of all respondents. Specifically, under the assumption that respondents who admit to the norm-defiant behavior when asked the direct question do not lie, this measure uses direct question responses for those respondents, and estimates the proportion of norm-defiers in the remainder of the sample by applying difference-in-means analysis to the list item counts of these remaining responses. The overall prevalence estimate is the weighted average of these two prevalence rates, weighted by relative sample size. The fourth-top point in each panel of Figure F.1 shows how, when one applies partition validation to this measure (by applying the estimator for actual voters and actual nonvoters separately, then combining results via a weighted average), classification accuracy is similar to that obtained for the standard list experiment measure, and clearly worse than that of the direct questions: (91.5 points in New Zealand; 88 points in London).

Finally, we consider classification accuracy one uses the 'relaxed liars' method for analyzing list experiments (Li, 2019). This provides lower and upper bounds for nonvoting prevalence. The former is equivalent to the standard difference-in-means estimate and assumes 'no liars'. The latter relaxes the no liars assumption and instead assumes that (a) only norm-defier respondents lie about the sensitive item and (b) norm-defiers are more likely to lie about the sensitive item when they are exposed to floor effects (or ceiling effects if an affirmative response to the sensitive item indicates norm-defiance). The fifth- and sixth-top points in each panel of Figure F.1 show classification accuracy when one applies partition validation for the lower and upper bounds, respectively (by applying the bound estimator for actual voters and actual nonvoters separately, then combining results via a weighted average).[6] As expected, the lower bound has the same overall accuracy as the standard list experiment measure. The upper bound – relaxing no liars – has substantially worse overall accuracy than other list experiment or direct question measures: (77.1 points in New Zealand; 12.1 points in London). Further analysis reveals that this is driven by the upper bound over-estimating the rate of nonvoting among actual voters, and thus increasing the false negative rate compared to other measures.

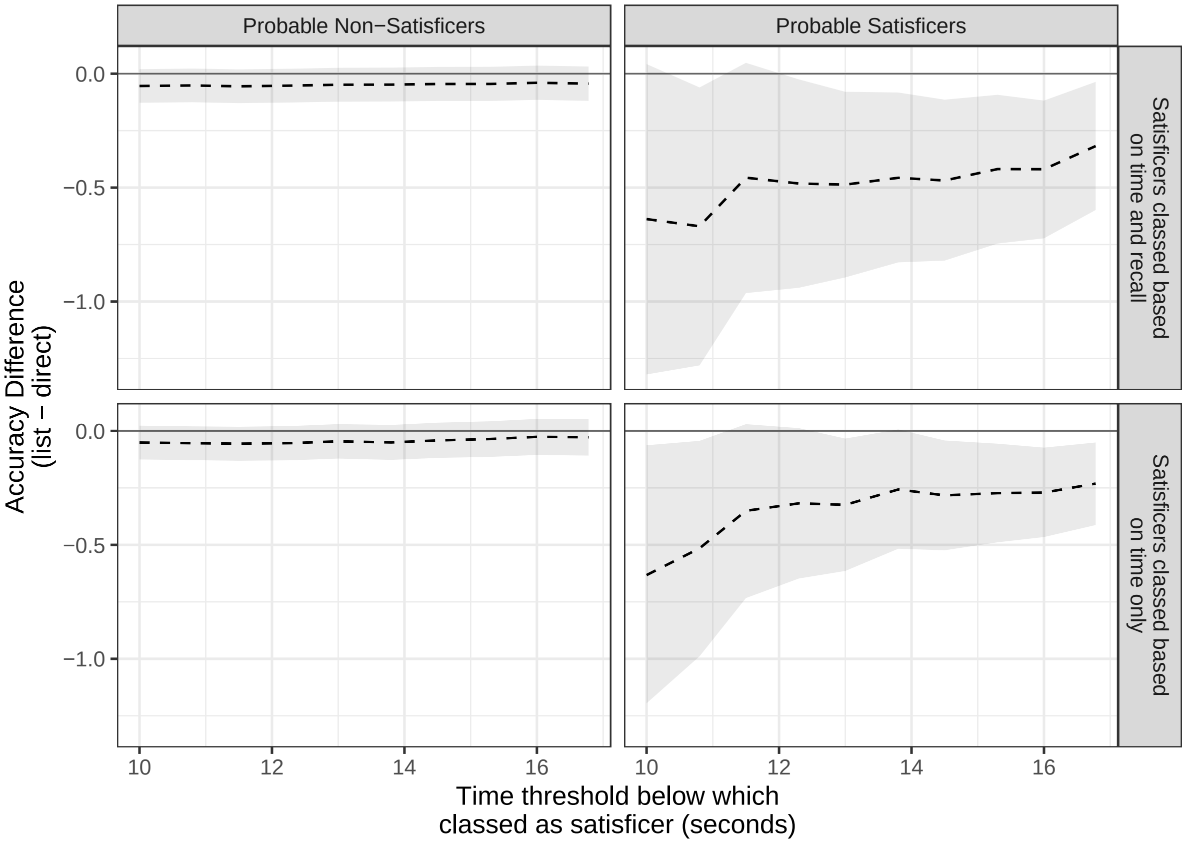
In sum, none of the alternative list experiment estimators examined here make the list experiment more accurate than either of the direct turnout measures examined.

[6] Following Li (2019), we suppress confidence intervals for these bounds in the figure.

G Relationship between non-strategic misreporting and alternative measures of satisficing

In the main text we show that the list experiment measure of turnout in our London survey is particularly inaccurate – and the drop in list accuracy versus direct question accuracy particularly pronounced – among those respondents who exhibit satisficing-consistent behavior when answering the list experiment. There, we identified list experiment satisficers as those respondents who (a) were unable to correctly recall either the first or last items on the list in a follow-up question and (b) were in the bottom quartile in terms of time taken to answer the list experiment question. In this appendix we show that finding is robust to different measurement strategies for identifying satisficers.

Figure G.1: Difference in List vs Direct Question Accuracy for Different Measures of List Satisficing



Note: Lines show estimated difference in list and direct question accuracy (y-axis) for different response time thresholds used to classify list experiment satisficers (x-axis). Top row gives results where satisficers are identified based on recall of list items as well as response time. Bottom row gives results where list satisficers are identified based on response time only. Left column gives results for identified non-satisficers. Right column gives results for identified satisficers. Gray bands are 95% confidence intervals.

Figure G.1 shows the difference in list and direct question measurement accuracy for probable list non-satisficers and probable list satisficers, respectively, as the criteria for identifying list satisficers varies. First, we vary whether list satisficers are identified based on list experiment response time and inability to recall list items (top row) or response time alone (bottom row). Second, we vary the list experiment response time threshold below which a respondent is classed as a satisficer (x-axis of each panel). We vary this from 10 seconds (the 5th percentile of response time for the list experiment) to just under 17 seconds (the 25th percentile).

Regardless of how we identify satisficers, the difference between list and direct question measurement accuracy among probable non-satisficers in Figure G.1 is consistently small and indistinguishable from zero with 95% confidence. In contrast, the difference between list and direct question measurement accuracy among probable satisficers is consistently negative, greater than 0.25 in absolute terms (on a -1 to +1 scale) and is mostly distinguishable from zero with 95% confidence. In sum, across the varying measures of of list experiment satisficing covered in Figure G.1, the drop in list experiment reporting accuracy relative to direct question reporting accuracy emerges mainly among measured list satisficers. This is again consistent with the notion that the list experiment question format induces satisficing and non-strategic misreporting that respondents do not engage in when responding to the direct question

H Trade-off between strategic and non-strategic bias in accuracy

Based on the formal model of direct question strategic misreporting and list experiment non-strategic misreporting in Section 2 we are able to derive formally the overall accuracy trade-off between the two types of misreporting. The expected accuracy of the direct question ($\mathbb{E}(\hat{\tau}^{\text{Direct}})$) given true norm-defier prevalence (π) and the proportion of norm-defiers who strategically misreport (θ) is

$$\mathbb{E}(\hat{\tau}^{\text{Direct}}) = (1-\theta)\pi + (1-\pi) = 1 - \theta\pi. \tag{1}$$

We derive expected accuracy of the list experiment assuming that a respondents' status as a list experiment non-strategic misreporter (S_i^*) is independent of their true status on the sensitive variable (X_i^*). Under this assumption, the expected accuracy of the list experiment ($\mathbb{E}(\hat{\tau}^{\text{List}})$) given π, the proportion of list experiment non-strategic misreporters (λ), and the expected list DiM among non-strategic misreporters ($\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})$), is

$$\begin{aligned}
\mathbb{E}(\hat{\tau}^{\text{List}}) \qquad &= \pi\mathbb{E}(\hat{\pi}^{\text{List}}_{X_i^*=1}) + (1-\pi)(1-\mathbb{E}(\hat{\pi}^{\text{List}}_{X_i^*=0})); && (2)\\
&= \pi\left((1-\lambda) + \lambda\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})\right) + (1-\pi)\left(1-\lambda\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})\right); && (3)\\
&= 1 - \lambda\left(\pi + \mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})(1-2\pi)\right). && (4)
\end{aligned}$$

Based on these expressions for expected direct question and list experiment accuracy, we can derive an indifference function which, for a given level of true norm-defier prevalence (π), proportion of list experiment non-strategic misreporters (λ), and expected list DiM among non-strategic misreporters ($\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})$), gives θ^*, the proportion of norm-defiers that must strategically misreport for the direct question in expectation such that expected list and direct question accuracy are equalized:

$$\begin{aligned}
\mathbb{E}(\hat{\tau}^{\text{Direct}}) &= \mathbb{E}(\hat{\tau}^{\text{List}}); && (5)\\
1-\theta\pi &= 1 - \lambda\left(\pi + \mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})(1-2\pi)\right); && (6)\\
\theta^* &= \lambda\frac{\pi + \mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})(1-2\pi)}{\pi}. && (7)
\end{aligned}$$

When $\theta \leq \theta^*$, expected list experiment accuracy is lower than expected direct question accuracy. When $\theta > \theta^*$, expected list experiment accuracy is greater than expected direct question accuracy.

I Does adding a placebo item to the control list eliminate non-strategic misreporting bias?

Ahlquist (2018) and Blair, Chou and Imai (2019) demonstrate that non-strategic misreporting can bias the DiM prevalence estimator and we note in Section 2 that the direction and size of this bias will generally be difficult to gauge given observable information in practical applications.

To address non-strategic misreporting in a list experiment Riambau and Ostwald (2020) suggest including a placebo statement (i.e., a statement no respondent can truthfully affirm) in the control list, such that the total number of available items is equalized in the control and treatment list. If non-strategic misreporters in list experiments report item counts that are a function of the total listed items, this should lead to a DiM of zero among non-strategic misreporters.

We show here that this solution does not eliminate non-strategic misreporting bias in list experiments, but does allow researchers to sign the direction of the bias.

If non-strategic misreporter status (S_i^*) is independent of a respondents status on the sensitive item (X_i^*), then the expected list prevalence estimate ($\mathbb{E}(\hat{\pi}^{\text{List}})$) is equal to

$$\mathbb{E}(\hat{\pi}^{\text{List}}) = (1 - \lambda)\pi + \lambda\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}), \tag{8}$$

where λ represents the proportion of list experiment non-strategic misreporters, π represents true norm-defier prevalence, and $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})$ is the expected DiM among list experiment non-strategic misreporters. It is easy to see that if and only if $\lambda = 0$, or $\lambda > 0$ and $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) = \pi$, will $\mathbb{E}(\hat{\pi}^{\text{List}}) = \pi$. In all other case the list prevalence estimator will be biased.

Now assume $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) = 0$ due to the inclusion of a placebo item in the control list as recommended by Riambau and Ostwald (2020). Then Equation 8 simplifies to

$$\mathbb{E}(\hat{\pi}^{\text{List}}) = (1 - \lambda)\pi. \tag{9}$$

This expression makes clear that, for all $\lambda > 0$, $\mathbb{E}(\hat{\pi}^{\text{List}}) < \pi$. Hence, the inclusion of a placebo item does not eliminate non-strategic misreporting bias in list prevalence estimates, as non-strategic misreporters now contribute a DiM of 0 to the list estimate in expectation – which implies false negatives among true norm-defiers. Nevertheless, the non-strategic misreporting bias in the list prevalence estimate is now clearly signed as negative.

References

Ahlquist, John S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimates." *Political Analysis* 26(1):34–53.

Aronow, Peter, Alexander Coppock, Forrest W. Crawford and Donald P. Green. 2015. "Combining List Experiments and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3(1):43–66.

Blair, Graeme and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1):47–77.

Blair, Graeme, Kosuke Imai and Jason Lyall. 2014. "Comparing and combining list and endorsement experiments: Evidence from Afghanistan." *American Journal of Political Science* 58(4):1043–1063.

Blair, Graeme, Winston Chou and Kosuke Imai. 2019. "List Experiments with Measurement Error." *Political Analysis* 27(4):455–480.

Corstange, Daniel. 2018. "Clientelism in competitive and uncompetitive elections." *Comparative Political Studies* 51(1):76–104.

Hanmer, Michael J., Antoine J. Banks and Ismail K. White. 2014. "Experiments to Reduce the Over-Reporting of Voting: A Pipeline to the Truth." *Political Analysis* 22(1):130–141.

Justice and Electoral Committee. 2016. "Inquiry into the 2014 general election." Report of the Justice and Electoral Committee, New Zealand House of Representatives `https://www.parliament.nz/resource/en-nz/51DBSCH_SCR68922_1/878b9b3603f17a6986fa56f6b0414924993c24e7` [Last accessed: 26 January 2020].

Kuhn, Patrick and Nick Vivyan. 2018. "Reducing Turnout Misreporting in Online Surveys." *Public Opinion Quarterly* 82(2):300–321.

Li, Yimeng. 2019. "Relaxing the No Liars Assumption in List Experiment Analyses." *Political Analysis* 27(4):540–555.

Manning, Christopher D., Prabhakar Raghavan and Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Riambau, Guillem and Kai Ostwald. 2020. "Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore." *Political Science Research and Methods* . `https://doi.org/10.1017/psrm.2020.18`.