Polit Behav
DOI 10.1007/s11109-013-9226-4

ORIGINAL PAPER

When Context Matters: Assessing Geographical Heterogeneity of Get-Out-The-Vote Treatment Effects Using a Population Based Field Experiment

Edward Fieldhouse · David Cutts · Peter John · Paul Widdop

Abstract The presence of heterogeneity in treatment effects can create problems for researchers employing a narrow experimental pool in their research. In particular it is often questioned whether the results of a particular experiment can be extrapolated outside the specific location of the study. In this article, we use a population-based field experiment in order to test the extent to which treatment effects for impersonal mobilisation techniques (direct mail and telephone) are sensitive to where they are carried out (geography) and the context of the election in which they were conducted. We find that on the whole it does not much matter where an experiment is conducted: the treatment effects are to all intents and purposes geographically uniform. This has important implications for the external validity of get-out-the-vote field studies more generally, especially where single locations are used. However, there is one important exception to this: experiments carried out in high turnout locations at high salience elections may show larger effects than those carried out in low turnout areas.

Keywords Get-out-the-vote · Campaigns · Field experiment · Geography · Turnout

E. Fieldhouse (✉) · P. Widdop
Institute for Social Change, University of Manchester, Manchester, UK
e-mail: ed.fieldhouse@manchester.ac.uk

D. Cutts
Department of Politics, Languages and International Relations, University of Bath, Bath, UK

P. John
School of Public Policy, University College London, London, UK

Published online: 16 April 2013

Introduction

Get-out-the-vote (GOTV) field experiments have an important and long history in political science, going back to Eldersveld 1956 study and before that to Gosnell's (1926). More recently, Gerber, Green and colleagues (Gerber and Green 2000a, b, 2001; Gerber et al. 2003; Green 2004; Green and Gerber 2008) have used randomised control trials that show that face-to-face mobilisation has a strong effect on voter turnout and is far more effective than less personal methods, such as telephoning and direct mail (see also McNulty 2005). In a short space of time the number of these experiments have increased dramatically, covering different populations (adults, young people, different ethnic groups); mobilisation methods (door-to-door, phone-banks, direct mail, leafleting, election-day mobilisation, robo-calls, email, radio broadcasts, TV adverts, print media, and street signs); variations in delivery (timing, tone, quality); partisan and non-partisan interventions; bilingual or multilingual modes of delivery (see Green and Gerber 2008).[1] Green et al. (2010: 3–4) note that in respect to direct mail alone "from 1999 through 2009, a total of 93 independent experiments were conducted, encompassing 127 treatments reported in 40 distinct studies". An increasingly important line of enquiry is the heterogeneity of treatment effects. A number of studies have explored the conditions under which treatment effects vary from population to population, from study to study and by treatment design (Imai and Strauss 2011; Arceneaux and Nickerson 2009; Green et al. 2012).

The presence of heterogeneity in treatment effects creates potential problems for researchers employing a narrow experimental pool in their research. This is a persistent critique of experimental studies: a lack of generalizability or external validity (Mutz 2011). Whilst field experiments enjoy the advantage over laboratory experiments that the treatments are tested in realistic settings, it is often questioned whether the results of a particular experiment can be extrapolated outside the specific location and to a generalised situation (Druckman and Kam 2011). Mutz (2011) has argued that the traditional goal of internal validity need not be sacrificed in the search for external validity if researchers adopt population based experimental designs. However, because of the difficulty in carrying out large-scale field experiments across large areas and over time, most GOTV studies have been focused on a single area at a single election (or a small group of geographically proximate locations) and for a single group of the electorate (notable exceptions include Green et al. 2003; Nickerson 2006; Bennion and Nickerson 2010).

Meta-studies potentially allow researchers to compare treatment effects across studies and draw inferences about generalised effects (Green et al. 2012). However, the sheer variety of these kinds of experiments, encompassing variations in design and mobilisation methods as well as target population, militates against generalisation. A meta-analysis may suffer from a high degree of heterogeneity in various elements of design, (Crombie and Davies 2009; DerSimonian and Laird 1986) a problem for which there is no easy fix. When comparing studies, it is difficult (if not impossible) to separate variation caused by the use of different mobilisation

[1] See http://gotv.research.yale.edu for summaries of these approaches.

methods from variation caused by unit or geographical heterogeneity. The challenge is to design a study that can make population inferences in the presence of heterogeneity.

Whilst a narrow experimental pool does not necessarily threaten causal inference, if heterogeneity in treatment effects does exist, then to achieve valid causal inference it is necessary to (a) sample some variation on the key moderating variables; and (b) allow the treatment effect to vary, for example by including the interaction of these moderating variables with the treatment effect (Druckman and Kam 2011). But what are these key moderating variables? They could be in the individual or unit characteristics such as political sophistication or demographic characteristics. Here we focus on geographical or contextual factors. Elections are highly heterogeneous across space and it is likely that treatment effects may vary across different types of areas, for example those with high prevailing levels of turnout compared to those with lower levels, or marginal as supposed to safe seats. Whilst single location studies have the potential for examining variability in treatment effects across different categories of elector, such as high versus low propensity voters (e.g. Niven 2001), only studies with variance on all the relevant dimensions of electoral context are capable of identifying the potentially crucial role of local electoral context.

Given that we may theoretically expect heterogeneity across different political contexts within a single country, then ideally we need a nationally representative sample of voters across a sample of electoral districts and across different elections. In this article, unlike any other previous GOTV studies of which we are aware, we use such a design to test the extent to which treatment effects for impersonal mobilisation techniques (direct mail and telephone) are sensitive to where they are carried out (geography) and the context of the election in which they were conducted. One of the considerable advantages of a nationally representative multi-factorial design is that it renders possible the examination of the heterogeneity of treatment effects across space. Of course there are other dimensions of heterogeneity which we cannot capture with a nationally representative population-based experiment including variation by country, over time (beyond the two elections sampled) and for different types of intervention. However, because the study is based on a nationally representative sample of electors drawn from a random sample of electoral districts (wards) we are able to explicitly test whether the effectiveness of an impersonal nonpartisan intervention varies across different political contexts measured on a number of different dimensions. These are selected because of their potential theoretical relationship with treatment effects and are described in the following section.

Underlying Level of Turnout

At the individual level, it has been noted that electors with a high underlying propensity to vote are less likely to be swayed by a leaflet or phone call (Hillygus 2005). Conversely, those with a low underlying propensity to vote may be difficult to persuade to change their mind (Niven 2001). Integrating these ideas, Arceneaux

and Nickerson (2009) predict a curvilinear relationship between the individual level underlying propensity to vote (or level of interest) and the efficacy of intervention with the point of optimum efficacy depending on the salience of the election (Arceneaux and Nickerson 2009). Thus, in low saliency elections it is relatively high propensity voters who are more likely to be on the cusp of their personal voting (or indifference) threshold. Extending this to the aggregate (constituency) level we might expect that areas with middling levels of turnout are more likely to be productive for campaigners than those with very high or very low levels, in medium or high salience elections. In areas with very high levels of turnout, the *average* propensity to vote is likely to be exceptionally high and many voters would vote regardless of the intervention, except in low salience elections when more voters may be close to their voting threshold. By contrast, in very low turnout areas it is likely that electors, on average, are less susceptible to mobilization. In these areas, the average latent propensity to vote is lower and, given that the treatment is likely to raise this propensity by only a small amount, then the proportion that are raised above a critical threshold is likely to be low, except when the election salience is very high. In accordance with those who advocate the curvilinear argument, "GOTV efforts are likely to mobilize voters who fall in the middle of the voting propensity spectrum" (Arceneaux and Nickerson 2009: 3). By extension GOTV campaigns may be likely to mobilise those living in areas of mid-level turnout, though this may vary according to the saliency of the election (for example, contrasting a European and General Election as we are able to do here). In other words, we extend the logic of curvilinear contingent theory of turnout of Arceneaux and Nickerson (2009) to apply to geographical electoral districts, and more specifically the relationship between mobilization efficacy, the underlying or prevailing level of turnout and election salience.

Electoral Competitiveness

The competitiveness of the electoral contest has a bearing on where a party or candidate campaigns. Parties target campaign resources where the contest is close as it is in these marginal seats where party activism it is likely to have highest potential impact. A large body of literature shows that local party campaigns are effective at mobilising party supporters (Denver and Hands 1997; Johnston and Pattie 2006; Fieldhouse and Cutts 2008; Cutts 2006). Any non-partisan GOTV campaign must, therefore, vie with other campaigns for the attention of voters. Where party campaigns are intense, voters who are most likely to be persuadable by mobilisation techniques may be mobilised by parties regardless of the intervention being studied. In other words, the more marginal the seat, the more intense the party activism, and the greater the likelihood that the experimental GOTV treatment is to be "drowned out" by other interventions, since the control group will be likely to receive a large amount of election information that has nothing to do with the experiment.

There are also alternative reasons why the electoral competitiveness of the seat could drown out non-partisan GOTV effects. Those electors living in seats where the contest is highly competitive are likely to be aware of the seat status, and as a

consequence, more likely to have heightened levels of political awareness and have greater local political knowledge. Of course, this in itself may be a function of intensive party campaigning, but also other factors such as the media (old and new) and more politicised social networks. The decision about whether to participate or not is also more likely to be made in the knowledge that, unlike many electoral contests in other places, it could have a bearing on the final outcome.

In this study there is a range of geographical areas which make it possible to explore this relationship. Here we use a marginality variable—identifying those seats where the margin is less than 10 %—which not only captures the intensity of campaigns carried out by political parties but also reflects the higher levels of political knowledge and interest among those electors living in seats where the electoral contest is more competitive. Margin also has an additional advantage over the use of a campaign measure such as party campaign spending, insofar as it is easier to replicate in other contexts.

Party Control

As well as differing in respect to the prevailing level of turnout and the level of competiveness, parliamentary constituencies vary in a number of other politically relevant ways that may affect the efficacy of GOTV treatments. In general, such factors reflect the character of the constituency in relation to the prevailing political cleavages of the nation (Agnew 1987). The most important of these include the socio-economic and demographic profile of the seat, its' local political culture and history, and the personal profile and support of local candidates. Given that, by their very nature, these are all correlated with the popularity of each of the major political parties; party incumbency provides a useful proxy for these sources of variation. Thus, for example, the social profile of constituencies (whether it's predominantly working class or middle class) is highly correlated with the identity of the incumbent party. Moreover, in any given election the nature of the campaign may be shaped by whether the defending incumbent is from the governing party or the opposition. For example, for any given level of competitiveness, because of the relative unpopularity of the government at the time of the 2010 general election, sitting Labour MPs were more likely to be under threat of losing their seat than those of opposition parties. In order to capture these differences and to test for potential biases among experiments carried out exclusively in government controlled or opposition controlled seats, we split the sample according to whether the incumbent MP was from the Labour Party (the governing party going into both elections) or an opposition party.

The Electoral Context

Electoral turnout varies according to the electoral context (Marsh 2002; Franklin 2004; Fieldhouse et al. 2007). As noted above Arceneaux and Nickerson (2009) argue that the point of optimum efficacy of a treatment will depend on the salience of the election. Although plausible, there is limited hard-evidence that the salience

of the election is systematically related to the size of treatment effects across experiments. Green et al. (2010) for example, find no significant variation in treatment effects by salience of election across 41 experiments carried out in the US. In this study we are able to compare treatment effects for a second order (European) election with a first order (general) election.

Following from above we test the following null hypotheses:

$H_{0(1)}$: Treatment effects do not vary significantly between electoral wards (sampling units);
$H_{0(2)}$: Treatment effects do not vary with the prevailing level of turnout in the ward;
$H_{0(3)}$: Treatment effects do not vary with the marginality/competitiveness of the electoral district (constituency);
$H_{0(4)}$: Treatment effects do not vary with the party of the defending candidate;
$H_{0(5)}$: Treatment effects do not vary with the with the type of the election (general versus European).

The Study

The study was designed to examine the effect of non-partisan mobilisation, through telephone canvassing and direct mail, on voter turnout in the European elections in England on June 4th 2009 and the UK General Election on May 6th 2010 (see Fieldhouse et al. 2013). In a multistage design twenty-seven local authority districts were randomly sampled and three electoral wards were randomly selected from each sampled district. The sample of wards provided a close match to England as a whole on a range of social and political characteristics.[2] Using a database based on electoral registers and telephone records, 40,000 individuals were sampled from these eighty-one wards. By design all sampled wards contain individuals from treatment and control groups in randomly distributed proportions. The sample was restricted to one random person per household to avoid clustering, and to ensure households did not receive double treatments. This sample was further stratified according to telephone accessibility and therefore included two separate sub-samples made up of 26,500 telephone accessible electors (any record with a valid landline or mobile) and 13,500 individuals telephone inaccessible electors (anyone with no telephone contact information). Each sampled telephone accessible individual was randomly assigned to one of three treatment groups (telephone, mail, or mail and telephone) and telephone inaccessible to the mail or control group. Because of the different treatment combinations available and their different effectiveness, in the following analyses we split by (or control for) telephone accessibility.

After the randomisation was complete, any electors in the sample (treatment or control groups) that were not registered or not eligible to vote were removed,

[2] Turnout rate in sample wards was 37.2 % compared to 35.1 % in England (2009 European Elections). Turnout rate in 2010 in sample wards was 67.1 % compared to turnout rate in England (2010 General Election) of 65.5 %. The sample of wards also represents England as a whole across a range of social and demographic characteristics (see Appendix Table 6).

leaving a sample of 25,293 in 2009. This reduction reflects redundancy in the sampling frame particularly arising from non-registration (since we include only registered electors in the analysis). At the General Election of 2010, we canvassed the sample again, but with the difference that we randomly allocated a portion of the 2009 control group to a new mail and telephone treatment group. Members of the three 2009 treatment groups were assigned to receive a repeat dose of the same treatment in 2010. A proportion of the sample that was included in 2009 had left the electoral register in 2010 or had changed name/address details and was therefore excluded, leaving a sample of 21,984 in 2010. Further details of the study design are reported in Fieldhouse et al. 2013).

The intervention consisted of a GOTV campaign called 'Your Vote' which encouraged recipients of the treatment to vote for reasons of civic duty and expressive motivation. Telephone recipients received a brief phone call from a team of social science graduate students. Non-respondents were called back on at least five occasions at different times of the day to maximise the overall contact rate. The mail group received a personalised printed letter in a colour with almost identical message (tailored for the written word).

The total number of registered electors in the sample was 25,293 in 2009 and 21,984 in 2010. Of those in the telephone treatment group, 58 % were successfully contacted in 2009 and 78 % in 2010 (Fieldhouse et al. 2013). Official records of voter turnout were collected after both elections to verify the turnout of treatment and control groups. In 2009, 17 % of electors in our sample voted by post, and 20 % did so in 2010. As a result of electoral law, there is no public record that indicates whether, individually, these people cast their vote and therefore postal voters are treated as missing data and excluded from all analyses. Moreover, applications for postal vote could not be influenced by the treatment as the closing date for applications (11 days before polling day) had passed when the treatments were applied.

Results

Before examining whether there was any significant variation in treatment effects between areas and across elections, we start by summarising the estimated treatment effects for the GOTV experiment overall. In this paper we focus on the overall intent-to-treat effect (ITT) as defined by the comparison of the sample assigned to any treatment group and the control group, since this provides the largest available sample, and therefore the best test of heterogeneity between geographic areas. The ITT simply compares the treatment and control group on the basis of assignment. It gives a conservative estimate of the average treatment effects, as it does not adjust for non-contact. This approach is preferred here as contact rates were not available for all treatment types.[3] Table 1 shows the estimated ITT for the overall treatment for 2009 and 2010 split by telephone accessibility.

In both elections, the overall treatment effect was positive but statistically insignificant for the telephone inaccessible treatment group. In 2010 this largely

[3] We cannot know whether or not mail treatments were read or not.

Table 1 Intent to treat effects for overall treatment

	2009 ITT (standard error)	2010 ITT (standard error)
Telephone inaccessible	1.03 (1.24)	2.00 (1.44)
	$N = 5{,}589$	$N = 4{,}222$
Telephone accessible	1.37* (0.82)	2.87* (1.47)
	$N = 15{,}299$	$N = 13{,}256$

ITT is equal to the percentage point difference in the turnout between those assigned to any treatment and the control group. The standard errors $= \sqrt{(pq/n)}$. *P*-values derived from standard comparison of proportions *z*-test. Tests based on one-tailed test of significance as effects are hypothesised to be positive

* Significant at 0.05 (one-tailed test)

Table 2 Treatment effects for combined (mail and telephone) experiment, compared for 2009 and 2010

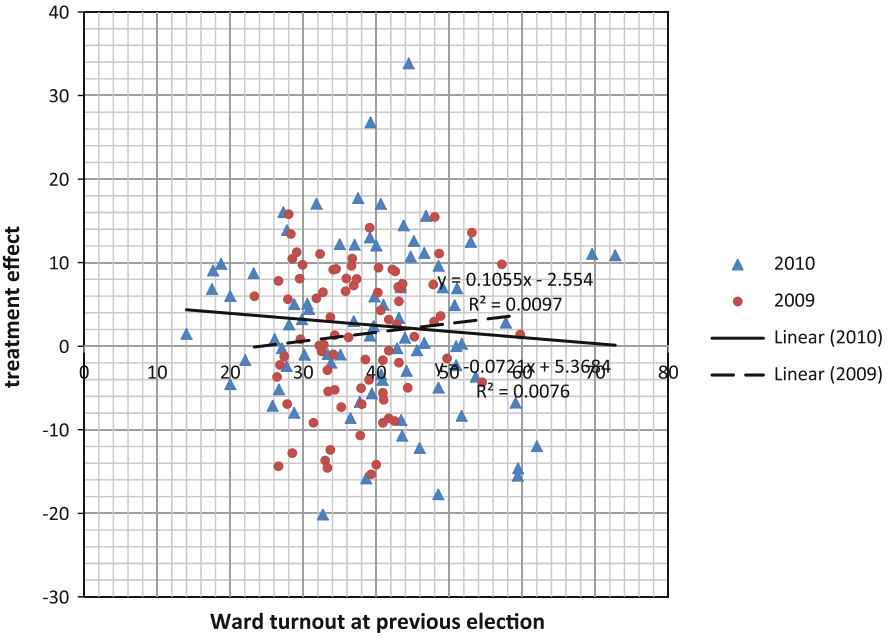
	2009	2010
N treatment	2,287	2,120
N control	5,179	2,352
Voted (treatment)	957	1,545
Voted (control)	2,058	1,620
Estimated intent-to-treat effect % (standard error)	2.11 (1.23)*	4.00 (1.36)*
Difference in TE = 1.89 *t*-statistic = 1.03		

* Significant at 0.05 (one-tailed test)

reflects the lesser effectiveness of the mail treatment effect compared to the telephone or combination effect, but in 2009 it also reflects a weaker mail effect in the telephone inaccessible group (see Appendix Table 7). Amongst the telephone accessible treatment group, the overall treatment effect was significant at both elections. The largest effects were for those receiving the combined treatment and, in 2010, for the telephone treatment.

Although Table 1 shows a larger effect in 2010 than in 2009, we cannot simply compare the overall treatment effect at the two elections. To make this comparison we must focus on the combination treatment (rather than the overall ITT), because the mail and telephone separate treatments are not strictly comparable between elections, as the 2009 mail and telephone groups were re-contacted in 2010. Table 2 therefore compares the effectiveness of the combination treatment across two different elections. The comparison of 2009 and 2010 gives an excellent test of the relevance of electoral context when comparing experiments, because the combination treatment was identical at both elections and carried out in exactly the same geographic locations.

The 2010 election was a first order election with a high-level of salience and the resultant level of turnout was much higher than in 2009 by a factor of two (nationally turnout was 65 % in 2010 compared to 34 % in 2009). Whilst there is reason to suppose the relationship between salience and the efficacy of GOTV treatments will depend on individual propensities to vote (Arceneaux and Nickerson 2009), overall the low level of interest in 2009 and the disillusionment with party politics prevalent at the time, appears to have limited the effectiveness of the 2009



Note. Previous turnout in ward in 2009 is derived from 2008 local election results, provided by Professor Michael Thrasher (The Elections Centre, University of Plymouth); and in 2010 from the control group turnout in 2009.

Fig. 1 Overall treatment effect for telephone accessible sample, 2009 and 2010 by percent turnout of ward at the previous election

treatment relative to 2010. However, the t-statistic for the difference in treatment effects is not significant and therefore we cannot discount $H_{0(5)}$. In other words there is no firm evidence that the treatment varies significantly between elections although the direction and magnitude of the effects do indicate that the treatment may have been more effective at the 2010 high salience election.

Comparing Between Areas and Within Elections

Figure 1 shows the relationship between ward turnout and the size of the overall treatment effect for telephone accessible electors for each ward in 2010, depending on the prevailing level of turnout in the ward, as measured by the turnout of the control group in the ward at the previous election.[4] Although each ward estimate is based on small numbers, there appears to be a very weak relationship between the

[4] We cannot use contemporaneous turnout in the control group as the measure of underlying turnout since this is used in the calculation of the treatment effect. Regression to the mean ensures that, by chance alone (notwithstanding the treatment effect), where the control group turnout is higher we would statistically be more likely to find a (comparatively) lower score for the treatment group. Therefore subtracting the control group turnout from the treatment group turnout to give the treatment effect, other things being equal, will give a negative slope coefficient. We therefore use turnout at the previous election which is akin to using prior voting record at the individual level.

underlying level of turnout and the treatment effect. In 2009 this relationship is slightly negative and in 2010 slightly positive but the *R*-squared at both elections is less than .01. This provides prima facie evidence that there is no strong or consistent relationship between the prevailing level of turnout and the efficacy of the treatment within a single election. In other words there is no systematic relationship between the underlying turnout level and the effectiveness of the treatment.

Modelling Variation in Treatment Effects

Above we showed that there is a weak relationship between the local treatment effect and the underlying level of turnout. However, although at the aggregate level this was a large-*N* experiment, when disaggregated to ward level, the sampling error around each individual ward estimate is quite large. In order to test the overall significance of variation in the treatment effect between wards we use multilevel (hierarchical) models, where vote is the dependent variable, and the independent variable is the treatment assignment (hence we are estimating the ITT). The hierarchical approach allows us test for variation in the level of turnout (the intercept); the treatment effect (the slope) and more particularly the covariance of the two. The covariance tells us whether the size of the treatment effect (the slope) is correlated with the local level of turnout (the intercept). It also allows us to test whether across the overall sample these random effects are statistically significant.

The hierarchical logistic models are fitted using MLwiN 3.2, with the estimates for the model derived using a Markov Chain Monte Carlo (MCMC) estimation procedure (Browne et al. 2005). Snijders and Bosker (2011) state that it is common to estimate hierarchical models using estimation methods based on marginal quasi-likelihood (MQL) or penalized (predictive) quasi-likelihood (PQL) procedures. However, when fitting binary response models, both of these quasi likelihood estimators can lead to an underestimation of the random effects, particularly when they are large and there are small numbers of observations within higher-level units, as is the case with our sample (Browne et al. 2005; Goldstein and Rasbash 1996; Rodriguez and Goldman 1995). Recent evidence also suggests that the Bayesian estimation procedure (MCMC method with diffuse priors) is less biased than either of the quasi-likelihood methods for binary response models (Browne et al. 2005). Moreover, if there is any higher level variation we want to be sure we find it, so it is imperative to use the MCMC approach.

Here, we used MLwiN software to estimate the starting values using first-order PQL, then 5,000 runs to derive the desired proposal distribution (discarded after convergence of the "burn in" period), followed by 50,000 simulated random draws to obtain the final estimates. We use the Metropolis–Hastings algorithm and the default diffuse gamma priors for variance parameters. The estimates in Table 3 are based on the mean of the simulated values, and the significance is derived from the standard error which is the standard deviation of the converged distribution. These estimates correspond to the traditional maximum likelihood estimate and its standard error.

Table 3 Multilevel MCMC Logistic Model Turnout with overall treatment

	2009 Coef (SE)	2010 Coef (SE)
Effect size (treatment)	0.062* (0.031)	0.116* (0.046)
Intercept variance	0.165* (0.032)	0.117* (0.031)
Slope variance	0.005 (0.004)	0.019 (0.012)
Covariance	0.002 (0.010)	0.001 (0.016)
N	20,888	17,117

Telephone accessibility included as a control

* Significant at $P \leq 0.05$

Table 3 shows the summaries of model results for the overall treatment effect, comparing any person allocated to any of the three treatment groups with the overall control group, regardless of whether they have telephone information or not. Telephone accessibility is controlled for with a covariate in the model. The overall treatment effects were statistically significant at the 5 % level in both elections, as represented by the overall effect size. Looking at the random effects, turnout varies significantly by ward at both elections, as represented by the intercept variance. This is unsurprising, and simply reflects geographical variation in the underlying level of turnout. What is more important is that there is no significant variance in the slope (the treatment effect) in either 2009 or 2010. There is also no significant covariance between the intercept and the slope, suggesting no systematic relationship between the local treatment effect and the level of turnout. The analyses were repeated for each of the separate experiments at both elections. In no instances across the two elections and across any of the methods of mobilisation, either alone or in combination, was there significant variance in the slope (the treatment effect), or the co-variance of slope and intercept (the tendency to vary according to the turnout rate).[5] We therefore cannot reject $H_{0(1)}$ or $H_{0(2)}$.

It is possible to compare the relative effectiveness of different models—in our case the baseline random intercepts model against the random slopes model—and evaluate their goodness fit by using the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002; van der Linde 2005). The DIC can be calculated from an MCMC run by calculating the value of the deviance at each iteration, and the deviance at the expected value of the unknown parameters. The DIC statistic also accounts for the number of parameters in the model, with a difference of less than 2 between models suggesting no difference, while a difference of 10 or above indicating an improvement in the goodness of fit (Burnham and Anderson 2002). A comparison of the DIC with random slopes and without (random intercept only) suggests there was no difference between the models for any of the treatments at either election (see Appendix Table 10 for further details). In other words, there was no improvement in model fit by relaxing the assumption that treatment effects are equal across geographical areas.

[5] See Appendix Tables 8 and 9.

Sources of Variation

The multilevel models allowed us to test for overall variation in the treatment effects and whether it varies with the overall level of turnout. We found no evidence that it does either. However it may be possible that there is some variation along the specific dimensions discussed above (electoral competitiveness and party control). As noted by Druckman and Kam (2011) where there is a theoretical expectation of heterogeneity in treatment effects we need sufficient variation in the key moderators (in our case political context), which is achieved through the sampling of 81 geographical locations. However, for valid causal inference these moderators must be interacted with the treatment. We test this by fitting fixed effect logit models with interactions between treatment effects and indicators for each of the relevant moderators. More specifically, we examine whether the impact of the intervention on turnout varies with electoral competitiveness of the seat (marginality), party control of the seat (Labour incumbency), and prior turnout (high, medium and low). Whilst there are some potential problems in using models containing covariates to adjust for imbalance (Bowers 2011), the model-based approach provides an excellent approximation of randomisation-based differences of means (Green 2009).[6] Moreover, there is no evidence of such imbalance in our sample and model estimated average treatment effects are almost identical to unadjusted effects (see Fieldhouse et al. 2013). The purpose of the models presented here is not to adjust for covariate imbalance or improve the estimation of the ITT per se, but to estimate the co-variation of the treatment effect and the contextual moderators defined above.[7] As a check on the model based results, we also stratified the sample according to the

[6] There has been much scholarly debate about the use of multiple regression to analyse experimental data. The main argument is that the introduction of assumptions associated with multiple regression are not justified by randomization and that the difference in means is the most appropriate estimator (Freedman 2005). Green (2009) provides a robust defence for the use of multiple regression in experimental analysis. Green (2009) uses a number of hypothetical examples and a voter mobilisation mail experiment to show that the discrepancy between the average multiple regression estimate and the true average treatment effect is negligible both in substantive terms and in relation to the standard error. In summary, multiple regression provides accurate estimates and standard errors, and this is the case even when the sample size is relatively small (Green 2009).

[7] Green and Kern (2012) do, however, claim that some obstacles exist including the possibility of specification error, multicollinearity when a large number of interaction terms are used and data-dredging where the researchers search for treatment-covariate interactions until they discover 'interesting' heterogeneity for some subsets of experimental units. Here we use the multiple regression method (inclusion of covariate and treatment-covariate interaction) as a method for estimating treatment effects and argue, like Green (2009), that it is identical to the traditional way of calculating the difference in means (splitting the sample). Our models carefully adhere to the set assumption. We explicitly test for specification error (using the linktest command in STATA 12) and find no evidence of this in our models (the _hatsq is insignificant, for instance in the incumbency model it has a P value of 0.28). We also find no evidence of serious multicollinearity. Our models only contain one interaction so the concerns raised (multiple interactions in the model) by Green and Kern (2012) is not valid in this case. Finally, the saliency of electoral competitiveness, underlying turnout and party control on 2009/10 turnout is well documented, not just here, but in the wider discipline and are selected for theoretical reasons.

Table 4 Logit model of treatment effects and electoral competitiveness (margin) and labour incumbency on turnout in the 2010 general election

	Treatment (T)			Main effect (X)			T*X			LL	Cases
	β	SE	Odds	β	SE	Odds	β	SE	Odds		
Marginality (X_1)	0.14*	0.05	1.15	0.27*	0.10	1.31	−0.11	0.07	0.90	−10,406	17,117
Labour incumbency (X_2)	0.19*	0.08	1.21	−0.21	0.11	0.81	−0.12	0.09	0.89	−10,381	17,117

Models include telephone accessibility as covariate

LL log likelihood

* Significant $P \leq 0.05$. Robust standard errors clustered by constituency ($N = 47$)

key contextual variables and calculated simple unadjusted treatment effects for the relevant groups.[8] These results are discussed further below.

Table 4 shows the overall treatment effect, the coefficients for two of the key contextual variables (marginality and incumbency) and the interaction between treatment effects and the contextual variable on turnout in the 2010 General Election.[9] Looking first at marginality, the overall treatment effect was statistically significant at the 5 % level. As expected, the 'margin' main effect was significant. Those individuals living in the most competitive seats were more likely to vote than electors living in much safer seats. However, there was no evidence that the treatment effects varied by the marginality of the seat. This was confirmed by splitting the sample into marginal and non-marginal wards and estimating treatment effects for the separate sub-groups (see Appendix, Tables 11, 12, 13). For both telephone accessible and inaccessible, although treatment effects were larger (and only significant) for non-marginal seats, the confidence intervals overlap, suggesting the treatment effects do not differ significantly. Similarly, we find no evidence that treatment effects vary by party control. People living in seats where there is a Labour incumbent were less likely to turn out, hardly surprising given the socio-economic characteristics of many of these constituencies and the electoral context (with Labour as the governing party losing support). As a consequence, party supporters in these areas where Labour were strong may have been less inclined to participate. However, this did not have any bearing on the efficacy of the treatment, and there is no significant interaction with the treatment effect. Again, this is confirmed by the split sample analysis. As for marginality, for the telephone accessible sample, the treatment was statistically significant in one group (non-Labour incumbents seats) but not the other (Labour seats), but the two samples did not differ statistically from each other. Given these findings, we therefore cannot reject $H_{0(3)}$ and $H_{0(4)}$.

[8] There are alternative ways of estimating heterogeneity of treatment effects based on Bayesian statistical decision theory (e.g. Imai and Strauss 2011).

[9] We also tested the effects of party spending using both a dichotomous variable (high spending versus low spending) and an overall spending measure obtained from the electoral returns of the three main parties during the 2010 official election campaign period. We found that both measures of spending had no significant effects reflecting the lack of variation in the spending variable.

Table 5 Logit model of treatment effects and prior turnout on turnout in the 2009 European elections and the 2010 general election

Treatment	2009			2010		
	B	SE	Odds	B	SE	Odds
Overall treatment	0.08	0.06	1.08	−0.02	0.06	0.98
High turnout	0.62*	0.07	1.86	0.42*	0.14	1.52
Mid turnout	0.33*	0.06	1.39	0.24*	0.11	1.27
T*High turnout	0.00	0.09	1.00	0.24*	0.10	1.27
T*Mid turnout	−0.05	0.08	0.95	0.15	0.09	1.15
Log likelihood	−13775.22			−10339.62		
N	20888			17,117		

Models include telephone accessibility as a covariate

LL log likelihood

* Significant $P \leq 0.05$. In all models, robust standard errors clustered by ward ($N = 81$). In 2009, Turnout is categorised on the basis of prior turnout in local elections (2006, 2007 and 2008). In 2010, Turnout is a categorical variable—high, mid and low—and is based on the 2009 European election GOTV sample (for each ward in the sample)

Table 5 shows the results of whether the treatment is related to prevailing turnout—through the splitting of the sample according to whether the overall level of turnout in the area is high, medium or low (allowing for a curvilinear relationship). We used previous local election turnout for the 2009 model (as defined in Fig. 1) and prior turnout in the 2009 European elections (from our control group sample) in the 2010 model. Because the 2009 election was a second-order low-salience election and the 2010 election was a first-order/high-salience election, the underlying turnout rates were defined in relative terms with three equal; sized categories at each election.[10] Unsurprisingly, in both 2009 and 2010, those individuals living in higher and medium turnout areas were significantly more likely to vote than those living in low turnout areas. Of more significance were the findings of the interaction between the treatment intervention and the local prevailing level of turnout. In 2010 (but not 2009) the overall treatment, had a significantly greater impact in high turnout areas. The split sample analysis (for the telephone accessible sample) also shows a larger effect in high turnout areas, though the confidence intervals do overlap (see Appendix Tables 11, 12).[11] The greater efficacy of the intervention in high turnout areas, at the high salience general election (where overall turnout was 65 %), is consistent with an individual level phenomenon of maximum treatment effects for high propensity voters (e.g. Green 2004). By

[10] In 2009 low turnout is defined as <32 %, mid turnout 32–45 % and high turnout >45 %. In 2010 low turnout is defined as <32 %, mid turnout 32–42 % and high turnout >42 %.

[11] Examination of the separate experiments (telephone, mail combi etc.) also supports this. None of these experiments showed a significant treatment effect in low turnout areas. However, in a number of cases the interactions with high turnout were significant, indicating significant treatment effects in high turnout areas. In 2010, both the double combination treatment and the new combination treatment showed a significant impact on turnout in higher turnout areas. A similar finding was found for the telephone treatment in 2009.

contrast, there is little support for the aggregate level equivalent of the (contingent) curvilinear theory (cf. Arceneaux and Nickerson 2009) which would predict the largest treatment effects in high-turnout areas in 2009 (a mid-salience election where the average turnout is around 50 % in high turnout wards) or in mid-turnout areas at the high salience 2010 election (again, where average turnout is around 50 %). However, it should be remembered that we are testing an aggregate level theory concerning the underlying level of turnout in the area, so we are not making any claim about the veracity of the individual level curvilinear theory, only that it does not appear to apply at the aggregate level in the way hypothesized.

Overall, there was some limited evidence that the treatment effects varied with the prevailing level of turnout in the area, with the treatments being very slightly more effective where turnout was already high in a high salience election.

Conclusions

The nationally representative sample allowed us to explore geographical variations in the effect of the treatment across two very different elections. This multi-factorial design not only allowed us to examine the heterogeneity of treatment effects but also to make comparisons between the treatments as applied to different sections of the population. We examined one theoretically important source of potential variability, namely heterogeneity across space. More specifically whether the treatments effects were equal across different types of area, those where a party was in control, where the seat was competitive and those areas with high prevailing levels of turnout compared to those with lower levels. We proposed a number of null hypotheses which explicitly tested this.

The findings were largely consistent. First, there was no conclusive evidence that the treatment varied significantly between elections, though there was some indicative evidence that the treatment was more effective in the high salience first order election of 2010. Second, there was no significant variation in the treatment effect across geographical areas. In 2009 and 2010, whilst turnout varied by ward (the intercept variance) there was no significant variance in the slope (the treatment effect) in any of the multilevel models. We then tested whether there was any variation in the treatment effects along specific dimensions including party control, the electoral competitiveness of the seat, and the prevailing level of turnout in the area. There was no evidence that the treatment effects varied significantly by the marginality of the seat or by party control. However, there are two significant caveats to this conclusion. First, whilst overall variation was largely insignificant, and the estimation of split sample treatment effects showed that subgroups did not generate statistically significant differences to each other, there were a number of instances where the ITT for some subgroups were statistically significant to zero and others were not (non-marginal sets, non-Labour incumbent seats and high turnout seats). This suggests that selection of geographic location can make a difference as to whether significant effects are uncovered or not, especially where effects are close to the threshold of statistical significance. Second, in 2010 the overall treatment had a significantly greater impact on turnout in high turnout areas. Just as some previous research has shown, treatments may be more effective amongst regular previous voters (Green 2004; Niven 2001). At

the aggregate level our GOTV treatments did appear to be more effective in higher turnout areas, in the higher salience general election. This is consistent with an individual level inference that it may be easier to nudge those already likely to vote, than it is to change the mind of ardent non-voters. However, our results relate to the characteristics of areas, not voters, so it is more accurate to say that campaigning may be most effective in high turnout locations at higher salience elections.

Notwithstanding this, overall it seems, taking the geography of treatment effects as a whole, it does not matter too much where an experiment is conducted: the treatment effects are to all intents and purposes uniform. This has important implications for the external validity of GOTV field studies more generally, especially where single locations are used (which lack of variation on key contextual moderators). It is possible to use these findings to conclude that the effects of single-location GOTV experiment can be extended to a wide range of locations (within a single election) without serious threat to causal validity. However, researchers should be warned that experiments carried out in high turnout locations are likely to show larger effects than those carried out in low turnout areas. Similarly campaigners might be interested to know that an additional leaflet or telephone call in a high turnout area may be more effect than the same leaflet in a low turnout area – though of course the additional voters may be less likely to be pivotal in those areas. Whilst these findings are important for researchers and campaigners alike, we should stress there are unanswered questions, not least whether larger samples or different electoral contexts might throw up more statistically significant patterns of variation. Future work based on meta-data could test whether the heterogeneity in existing studies conforms to the patterns found here. Beyond that, a nationally representative sample from other countries including the US is the natural next step, to compare findings with this British study.

Acknowledgments We are grateful to the Economic and Social Research Council who supported the research project on which this article is based (Award Number RES-000-22-2827).

Appendix

See Tables 6, 7, 8, 9, 10, 11, 12 and 13

Table 6 Comparison of sample ward characteristics and England population (2001 census data)

%	England	Sample ward mean
White british	90.9248	87.8559
Black	02.3047	02.9522
Asian	04.5754	07.1857
Muslim	03.1032	05.0652
Sikh	0.6662	0.6845
Hindu	1.1131	1.4300
Economically active	66.8599	65.3208
Economically inactive	33.1401	34.6792

Table 6 continued

%	England	Sample ward mean
Employed in agriculture	1.4772	1.3194
Employed in manufacturing	14.8316	15.5060
National socio-economic classification class 1 or 2	8.6118	7.9903
National socio-economic classification class 7 or 8	20.6671	20.8256
Never worked	1.0124	3.3632
Long-term unemployed	.7321	1.0814
Full time students	5.2467	7.0906
Households with 2 cars	23.5611	22.3803
With limiting long-term illness	17.9272	26.6337
Single parents	6.4151	9.5903
With no educational qualifications	28.8519	30.5522
With level 4 or 5 qualifications (at least college)	19.9033	19.3940
Owner occupiers	68.7195	67.3057
Aged 18–29	15.0572	15.6438
Aged 30–59	41.5297	41.1189
Aged 60+	20.7572	21.1440

Table 7 Intent to treat effects for original experiments

	Mail (tel. inaccessible	Mail (tel. accessible)	Telephone	Combined	Repeat combined
2009 ITT (standard error)	1.03 (1.27)	1.60 (1.01)	0.60 (1.08)	2.11* (1.23)	–
2010 ITT (standard error)	1.99 (1.50)	1.72 (1.20)	3.36* (1.25)	4.00* (1.36)	3.01* (1.38)

* Significant at 0.05 (one-tailed test)

Table 8 Multilevel MCMC logistic model of 2009 turnout with treatments

Treatment	Effect size		Intercept variance		Slope variance		Covariance		Cases
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	
Combi (mail + tel)	0.097	0.054	0.202*	0.044	0.017	0.014	−0.036	0.023	7,466
Telephone	0.029	0.046	0.178*	0.039	0.005	0.007	−0.009	0.015	8,645
Mail (tel accessible)	0.068	0.045	0.199*	0.042	0.011	0.011	−0.010	0.018	9,546
Mail (inaccessible)	0.066	0.058	0.143*	0.038	0.004	0.006	0.001	0.012	5,589
All Mail[a]	0.068	0.036	0.166*	0.033	0.007	0.006	0.005	0.012	15,135

[a] Includes telephone accessibility as covariate

* Significant at $P \leq 0.05$

Table 9 Multilevel MCMC logistic model of 2010 turnout with treatments

Treatment	Effect size		Intercept variance		Slope variance		Covariance		Cases
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	
Combi (mail + tel)	0.189*	0.073	0.103*	0.040	0.058	0.040	−0.035	0.032	4,374
Double combi	0.149*	0.070	0.166*	0.045	0.006	0.011	0.009	0.015	4,261
All combi 09 +10	0.176*	0.058	0.136*	0.037	0.003	0.004	0.002	0.010	6,330
Telephone	0.161*	0.062	0.127*	0.036	0.005	0.007	0.003	0.012	5,234
Mail (tel accessible)	0.069	0.067	0.103*	0.039	0.078	0.044	−0.025	0.033	6,010
Mail (inaccessible)	0.065	0.070	0.132*	0.047	0.025	0.021	−0.042	0.026	4,153
All mail[a]	0.075	0.050	0.113*	0.033	0.035	0.022	−0.019	0.022	10,163

* Significant at $P \leq 0.05$

[a] Includes telephone accessibility as covariate

Table 10 Comparison of deviance information criterion for each treatment in 2009 and 2010: random intercepts models and random slope models

	Random intercepts model only	Random slope
2009 Treatment groups		
Overall treatment	27147.09	27148.26
All Mail	19541.21	19542.34
Mail Accessible	12576.09	12577.24
Mail Inaccessible	7011.67	7012.01
Combi 2009	9858.66	9858.27
Telephone	11390.30	11390.30
2010 Treatment groups		
Overall treatment	20525.39	20525.44
All mail	12526.86	12526.43
Mail accessible	7269.51	7266.72
Mail inaccessible	5298.22	5297.28
All combi	7485.00	7485.15
Combi 2010 only	5235.10	5235.34
Double combi	5090.54	5090.14
Telephone	6226.67	6226.75

Table 11 Split sample treatment effects 2009

	Telephone accessible			Telephone inaccessible		
	High turnout	Med turnout	Low turnout	High turnout	Med turnout	Low turnout
N in the treatment group	1,190	6,814	2,116	394	1,921	808
N in the control	622	3,491	1,066	296	1,637	534
N voted in the treatment group	588	2,845	727	157	679	216

Table 11 continued

	Telephone accessible			Telephone inaccessible		
	High turnout	Med turnout	Low turnout	High turnout	Med turnout	Low turnout
N who voted in the control group	282	1,426	350	122	547	137
Treatment effect	4.07	0.90	1.52	1.37	1.93	1.08
Lower confidence	−0.77	−1.11	−1.98	−5.97	−1.2	−0.78
Upper confidence	+8.87	+2.91	+4.96	+8.75	+5.05	+5.81

Table 12 Split sample treatment effects (telephone accessible) 2010

	High turnout	Med turnout	Low turnout	Lab incumb	Other incumb	Marginal	Non marginal
N in the treatment group	2,404	5,917	2,338	6,527	4,132	3,405	7,254
N in the control	552	1,236	517	1,389	916	731	1,574
N voted in the treatment group	1,870	4,295	1,496	4,513	3,148	2,536	5,125
N who voted in the control group	402	867	326	933	662	533	1,062
Treatment effect	5.0	2.4	0.9	2.0	3.9	1.6	3.2
Lower confidence	+1.0	−0.3	−0.4	−0.7	+0.8	−0.2	+0.0
Upper confidence	+9.1	+5.3	+5.6	+4.7	+7.2	+5.2	+5.8

Table 13 Split sample treatment effects (telephone inaccessible) 2010

	High turnout	Med turnout	Low turnout	Lab incumb	Other incumb	Marginal	Non marginal
N in the treatment group	531	1,294	714	1,635	904	800	1739
N in the control	335	896	383	973	636	535	1079
N voted in the treatment group	386	883	420	1058	631	534	1155
N who voted in the control group	236	578	235	621	428	370	679
Treatment effect	2.2	3.7	2.5	0.9	2.5	2.4	3.5
Lower confidence	−3.8	−0.3	−3.6	−2.9	−2.2	−2.7	−0.1
Upper confidence	+8.5	+7.8	+8.5	+4.7	+7.2	+7.4	+7.1

References

Agnew, J. (1987). *Place and politics*. London: Allen and Unwin.

Arceneaux, K., & Nickerson, D. W. (2009). Who is mobilized to vote? A re-analysis of eleven randomized field experiments. *American Journal of Political Science, 53*(1), 1–16.

Bennion, E. A., & Nickerson, D. W. (2010). The cost of convenience: An experiment showing e-mail outreach decreases voter registration. *Political Research Quarterly, 64*, 858–869.

Bowers, J. (2011). Making effects manifest in randomized experiments. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 459–480). Cambridge, UK: Cambridge University Press.

Browne, W. J., Subramanian, S. V., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A, 168*(3), 599–613.

Burnham, K., & Anderson, D. (2002). *Model selection and multi-model inference: A practical-theoretical approach* (2nd ed.). New York: Springer.

Crombie, I. K., & Davies, H. T. O. (2009). *What is meta-analysis?*. London: Hayward Medical Communications.

Cutts, D. (2006). Continuous activism and electoral outcomes: The liberal democrats in bath. *Political Geography, 25*, 72–88.

Denver, D., & Hands, G. (1997). *Constituency electioneering in Great Britain*. London: Frank Cass.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188.

Druckman, J. N., & Kam, C. D. (2011). Students as experimental participants: A defense of the \narrow data base. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Eldersveld, S. J. (1956). Experimental propaganda techniques and voting behavior. *American Political Science Review, 50*, 154–165.

Fieldhouse, E., & Cutts, D. (2008). The effectiveness of local party campaigns in 2005: Combining evidence from campaign spending and agent survey data. *British Journal of Political Science, 39*(2), 367–388.

Fieldhouse, E., Cutts, D., Widdop, P., & John, P. (2013). Do impersonal mobilisation methods work? Evidence from a nationwide get -out -the -vote experiment in England. *Electoral Studies, 32*(1), 113–123.

Fieldhouse, E., Tranmer, M., & Russell, A. (2007). Something about young people or something about elections? Electoral participation of young people in Europe: Evidence from a multilevel analysis of the European social survey. *European Journal of Political Research, 46*(6), 797–822.

Franklin, M. N. (2004). *Voter turnout and the dynamics of electoral competition*. Cambridge: Cambridge University Press.

Freedman, D. A. (2005). *Statistical models: Theory and practice*. New York: Cambridge University Press.

Gerber, A., & Green, D. (2000a). The effects of canvassing, direct mail, and telephone contact on voter turnout: A field experiment. *American Political Science Review, 94*(3), 653–663.

Gerber, A., & Green, D. (2000b). The effect of a nonpartisan get-out-the-vote drive: An experimental study of leafleting. *Journal of Politics, 62*(3), 846–857.

Gerber, A., & Green, D. (2001). Do phone calls increase voter turnout?: A field experiment. *Public Opinion Quarterly, 65*, 75–85.

Gerber, A., Green, D., & Green, M. (2003). The effects of partisan direct mail on voter turnout. *Electoral Studies, 22*, 563–579.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A, 159*(3), 505–513.

Gosnell, H. F. (1926). An experiment in the stimulation of voting. *The American Political Science Review, 20*(4), 869–874.

Green, D. (2004). Mobilizing African–Americans using direct mail and commercial phone banks: A field experiment. *Political Research Quarterly,57*(2), 245–255.

Green, D. P., Aronow, P. M., & McGrath, M. C. (2010). *Making sense of 200+ field experiments on voter mobilization, part I*. Paper presented at American Political Science Association meetings, Washington, DC.

Green, D. (2009). Regression adjustments to experimental data: Do David Freedman's concerns apply to political science, Annual Meeting of the Society for Political Methodology, Yale University.

Green, D., Aronow, P., & McGrath, M. (2012). Field experiments and the study of voter turnout. *Journal of Elections, Public Opinion and Parties,22*(4), 431.

Green, D., & Gerber, A. (2008). *Get out the vote: How to increase voter turnout* (2nd ed.). Washington: Brookings Institution Press.

Green, D., Gerber, A., & Nickerson, D. (2003). Getting out the vote in local elections: Results from six door-to-door canvassing experiments. *Journal of Politics, 65*(4), 1083–1096.

Green, D., & Kern, H. (2012). Modelling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly, 76*, 491–511.

Hillygus, D. S. (2005). Campaign effects and the dynamics of turnout intention in election 2000. *Journal of Politics, 67*, 50–68.

Imai, Kosuke, & Strauss, Aaron. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis, 19*, 1–19.

Johnston, R. J., & Pattie, C. J. (2006). *Putting voters in their place: Geography and elections in Great Britain*. Oxford: Oxford University Press.

Marsh, M. (2002). Electoral context. *Electoral Studies, 21*(2), 207–217.

McNulty, J. (2005). Phone-based GOTV—what's on the line? Field experiments with varied partisan components, 2002–2003. *The ANNALS of the American Academy of Political and Social Science, 601*(1), 41–65.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Nickerson, D. (2006). Volunteer phone calls can increase turnout: Evidence from eight field experiments. *American Politics Research, 34*(3), 271–292.

Niven, D. (2001). The limits of mobilization: Turnout evidence from state house primaries. *Political Behavior, 23*, 335–350.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A, 158*(1), 78–89.

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publication.

Spiegelhalter, D. J., Best, N., & van der Linde, A. (2002). Bayesian models of complexity and fit. *Journal of the Royal Statistical Society: Series B, 64*(4), 583–639.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica, 59*(1), 45–56.