MANCHESTER
1824

The University
of Manchester

Institute of Social Change

When Context Matters: An Assessment of the External Validity of Get-Out-The-Vote Experiments Using a Population Based Field Experiment

Institute for Social Change Working Paper 2012 -01

Edward Fieldhouse, David Cutts, Peter John and Paul Widdop

Whilst field experiments enjoy the advantage over laboratory experiments that the treatments are tested in realistic settings, it is generally unclear whether the results of a particular experiment can be extrapolated from the specific location and election to a generalised situation. 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 uniform. This has hugely important implications for the external validity of GOTV field studies more generally, especially where single locations are used. However, there is one important exception this: experiments carried out in high turnout locations are likely to show larger effects than those carried out in low turnout areas.

www.manchester.ac.uk/socialchange

Introduction

Get-Out-The-Vote (GOTV) field experiments have an important and long history in political science, going back to Eldersveld's 1954 study and before that to Gosnell's (1927). More recently, Gerber, Green and colleagues (Gerber and Green 2000a, b, 2001; Gerber et al. 2003; Green 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".

In spite of all that has been achieved, the key limitation of these studies is external validity. Whilst field experiments enjoy the advantage over laboratory experiments that the treatments are tested in realistic settings, it is generally unclear whether the results of a particular experiment can be extrapolated outside the specific location and election to a generalised situation. Mutz (2001) 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 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

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

for a single group of the electorate. Of course there are exceptions: some studies have carried out experiments spanning multiple areas, such as the six locations studied by Green et al (2003) for door-to-door canvassing. Researchers have pooled findings across several areas, such as Nickerson (2006) research on the effect of volunteer banks in eight areas. Bennion and Nickerson 2010 study of e-mail voter registrations took place across twenty-six locations, but this was to student populations.

Meta-studies that 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 in 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 is there is no easy fix. When comparing studies, it is difficult (if not impossible) to separate variation caused by the use of different mobilisation methods from variation caused by geographical heterogeneity. The challenge is to have a study design that can make a generalisation despite the presence of heterogeneity. For this there is the need for a large-scale study implemented in a wide range of locations. Given that we expect heterogeneity to be greater across national boundaries, we argue that a nationally representative study is the solution. However, no previous GOTV field experiment, as far as we are aware, has been based on a nationally representative sample of voters across a sample of electoral districts and across different elections. In this article, we use just such a design 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.

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. In particular, we can make comparisons between treatments administered to different sections of the population. 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 representative of all types of electoral context are capable of identifying the potentially crucial role of local electoral context. Elections are highly heterogeneous across space and it is highly 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.

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 the treatment varies across different political contexts. In particular, we have a nationally representative cross section with a mixture of areas on a number of different dimensions.

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, Arcenaux and Nickersen (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 (Arcenaux and Nickerson 2009). For example, in low saliency elections it is relatively high propensity voters who are more likely to be on the cusp of their personal voting 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 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. By contrast, in very low turnout areas it is likely that electors are less susceptible to the intervention. In these areas, the average latent propensity to vote is far lower and, given that the treatment is likely to raise this propensity by only a small amount, then in very low turnout areas the proportion that are raised above a critical threshold is likely to be low. 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” (Arcenaux 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).

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 2009; 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 interest and extensive general and 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 an established campaign measure such as party campaign spending, given that the former is far easier to replicate in other GOTV experiments.

Party Control

Generally speaking, there is no a priori reason why a non-partisan campaign may be more or less effective depending on the party of the incumbent candidate. Incumbents are likely to have a better organised and more efficient campaign infrastructure, which is designed to maximise their vote. They can readily get publicity through the local media - much of it unpaid for - through their representative work and their attendance at many local functions. And in the UK, incumbent MPs have free access to the postal service (outside of election time) that they can use to contact their constituents, either individually or through mass mailings aimed at sustaining and winning support. Put simply, they are more likely to have the visibility and resources to sustain their level of activism over a longer period, both in the run-up to the election and during the official election campaign period. Yet incumbents are already “saturated” with the sort of recognition brought about by campaigning, hence additional activism adds little to the voters’ knowledge or support (Jacobson, 1978). As a consequence, most scholarly studies have found that incumbent campaigning (spending) is less effective than challenger activism (Abramowitz 1988; Green and Krasno 1988; Jacobson 1990; Ansolabehere and Snyder 1996). The more success incumbents have enjoyed in previous elections, the more difficult it is for their campaign activities to generate votes, something that does not hold for challengers who start with much lower levels of support (Denver and Hands 1997).

Party incumbency does, however, provide a useful proxy for other possible sources of variation. For example, in any given election the nature of the background campaign may be shaped by whether the defending incumbent is from the governing party or the opposition. Also the demographic and socio-economic profile of constituencies is highly correlated with the identity of the incumbent party. In order to capture these differences and to test for possible biases among experiments carried out exclusively in government controlled or opposition controlled seats, we also 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 Arcenaux and Nickersen (2009) argue that the point of optimum efficacy of a treatment will depend on the salience of the election. Although plausible there is limited 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 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 U.K. General Election on May 6th 2010 (see Fieldhouse et al. forthcoming). In a multistage design twenty-seven local authority districts were randomly sampled and three electoral wards were randomly selected from each sampled district. Using a database based on electoral registers and telephone records, 40,000 individuals were sampled from these eighty-one wards. 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).[2]

Each sampled individual was randomly assigned to one of three treatment groups (telephone, mail, or mail and telephone). 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, 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

[2] By sampling the mail and control group from both telephone accessible and inaccessible numbers we are able to observe whether there is any difference in the propensity to vote and in the effectiveness of a (mail) treatment between those with and without telephone information. This informs the interpretation of the telephone treatment effect, which is restricted to the telephone accessible sample.

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. forthcoming.

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. forthcoming). 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.

Results

Before examining whether there was any significant variation in treatment effects between areas and across elections, we start the results section of the paper by summarising the estimated treatment effects for the different interventions from the two GOTV experiments. Table 1 shows the estimated treatment effect for each part of the experiment for both 2009 and 2010. In both elections, the mail experiment for the telephone accessible group was more likely to vote than the control group the difference was statistically significant at the 10% level. Amongst the telephone inaccessible group, the impact of the intervention had doubled in 2010 from 2009 and was significantly different from the control group at the 10% level.[3] While the telephone treatment had little or no effect in 2009, our results suggest that the intent-to-treat effect in 2010 was statistically significant (as was treatment effect accounting for contact) perhaps reflecting the context of the election and, like the mail, that the 2010 telephone intervention was a follow-up treatment (Fieldhouse et al forthcoming). The combined treatment was statistically significant in both elections, although this treatment amongst newly treated electors was slightly larger in the 2010 general election (4.0%) than where the same electors were treated with mail and telephone in both elections (3.0%). So whereas mail and telephone alone had varying effects, a combined approach, contacting electors by both telephone and mail, was far more effective at both elections. However, while there was clearly an additive effect, it was not possible to infer any synergy between the two modes of intervention (see Fieldhouse et al, forthcoming).

Table 1. 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 0.10; ** significant at 0.05 (one-tailed test)

[3] All tests based on one-tailed test of significance as effects are hypothesised to be positive. All treatment effects and ITTs are percentage point increases in turnout.

Comparing between areas and within elections

Figure 1 shows the relationship between ward turnout and the size of the treatment effect for each ward in 2010. Although each ward estimate is based on small numbers, there appears to be a very weak relationship between the turnout of the whole sample (treatment and control group) and the treatment effect. As turnout in the control group increases, the level size of the treatment effect seems to decrease very slightly, but the R-squared is less than .001. This provides prima facie evidence that there is no variation in treatment effects geographically within a single election. Figure 2 shows the equivalent chart for the telephone treatment effect, but with a very slightly positive relationship with overall turnout but with an almost zero R-squared. Similar graphs for other treatments and for 2009 show a similar pattern. In other words there is no systematic relationship between the underlying turnout level and the effectiveness of the treatment.

Figure 1. Treatment effect of mail and telephone (combi) treatment 2010 by turnout of control group by ward

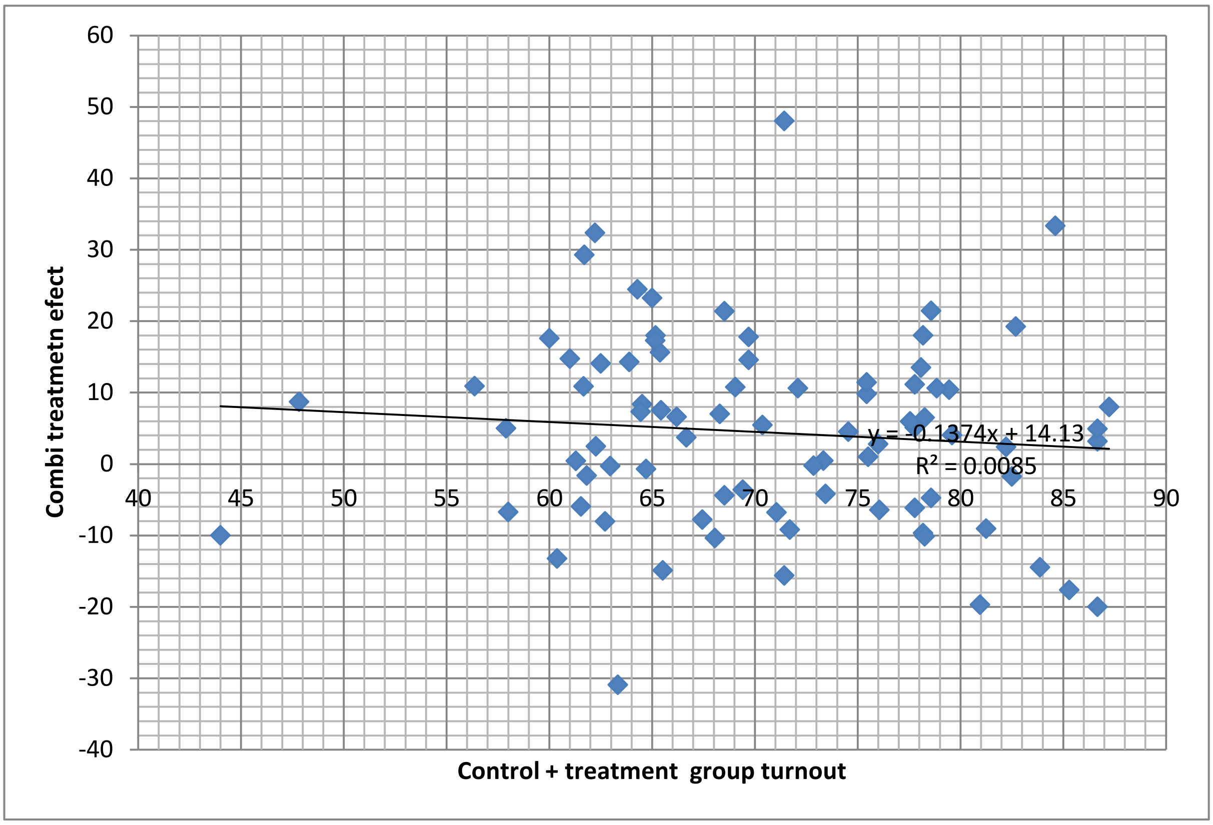
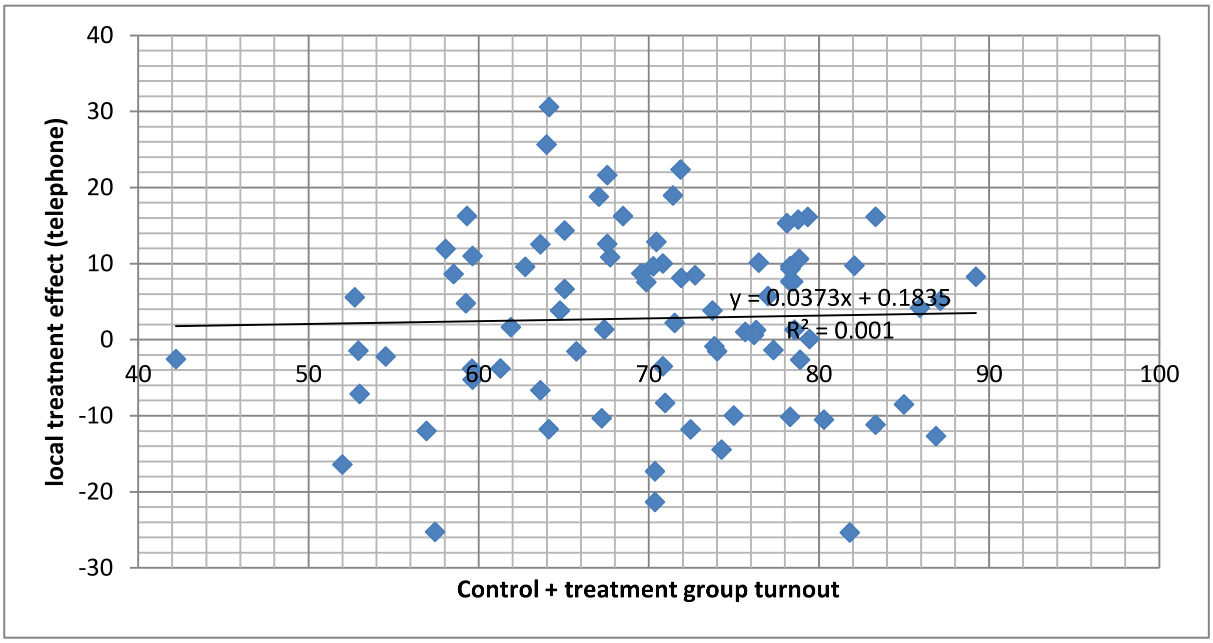


Figure 2. Treatment effect of telephone treatment 2010 by turnout of control group by ward



Modelling spatial variation in treatment effects

Above we have shown that there is no systematic relationship between the local treatment effect and the underlying level of turnout. However, although this was a large-N experiment, when broken down into wards the sample sizes are not sufficient to provide precise estimates for individual wards. Although the regression across all 81 wards should smooth out any ward level sampling errors, Figure 1a and 1b do not take into account the sampling variation around each estimate. 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). This 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 2 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 2. 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	
All Treatments#	0.062**	0.031	0.165**	0.032	0.005	0.004	0.002	0.010	20888
Combi (Mail+tel)	0.097*	0.054	0.202**	0.044	0.017	0.014	-0.036	0.023	7466
Telephone	0.029	0.046	0.178**	0.039	0.005	0.007	-0.009	0.015	8645
Mail (Tel accessible)	0.068	0.045	0.199**	0.042	0.011	0.011	-0.010	0.018	9546
Mail (Inaccessible)	0.066	0.058	0.143**	0.038	0.004	0.006	0.001	0.012	5589
All Mail#	0.068*	0.036	0.166**	0.033	0.007	0.006	0.005	0.012	15135

#Models with telephone accessible and inaccessible include 'known' as covariate.
** significant at p= <0.05; * significant at p= <0.10

Table 2 shows the summaries of model results for six different configurations of the experiment. The 'All Treatments' experiment represents a comparison of 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 same approach is used for 'all mail' which simply pools the 'mail accessible' and 'mail inaccessible' experiments. The 'telephone', 'mail accessible' and 'combination' (those receiving both) treatment groups are compared with the telephone accessible control group only and the 'mail inaccessible' with the telephone inaccessible control group only. Only the overall treatment effects were statistically significant at the 5% level, though 'all mail' and the combination treatment were significant at the 10% level. Looking at the random effects, whilst turnout varies by ward (the intercept variance) there is no significant variance in the slope (the treatment effect) in any of the models. There is also no significant covariance between the intercept and the slope, confirming that there is no systematic relationship between the local treatment effect and the level of turnout. We therefore cannot reject $H_{0(1)}$ or $H_{0(2)}$.

Table 3 gives the equivalent results for 2010. Here we find more significant treatment effects, with only the mail proving ineffective (Fieldhouse et al. forthcoming). Again we find significant variation in turnout but no significant variance in either the slope or the slope/intercept covariance in any of the experiments. Interestingly, in 2010 there was more variation by mail than by telephone despite the fact that ward variation in telephone effects could potentially have been inflated by interviewer effects, whereas other things being equal, mail effects might be expected to be uniform.

Table 3. 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	
All Treatments#	0.116**	0.046	0.117**	0.031	0.019	0.012	0.001	0.016	17117
Combi (Mail+tel)	0.189**	0.073	0.103**	0.040	0.058	0.040	-0.035	0.032	4374
Double Combi	0.149**	0.070	0.166**	0.045	0.006	0.011	0.009	0.015	4261
All Combi 09 +10	0.176**	0.058	0.136**	0.037	0.003	0.004	0.002	0.010	6330
Telephone	0.161**	0.062	0.127**	0.036	0.005	0.007	0.003	0.012	5234
Mail (Tel accessible)	0.069	0.067	0.103**	0.039	0.078	0.044	-0.025	0.033	6010
Mail (Inaccessible)	0.065	0.070	0.132**	0.047	0.025	0.021	-0.042	0.026	4153
All Mail#	0.075	0.050	0.113**	0.033	0.035	0.022	-0.019	0.022	10163

#Models with telephone accessible and inaccessible include 'known' as covariate.
** significant at p= <0.05; * significant at p= <0.10

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, Best, and van der Linde 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 without (random intercept only) and with (random slopes) the two level two parameters suggests there was no difference between the models for all treatments at both elections (see Appendix Table A1 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.

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). We test this by fitting fixed effect logit models with interactions between treatment effects and indicators for each of the four dimensions. More specifically, we examine whether the treatment effects vary along the three dimensions outlined above: electoral competitiveness of the seat (marginality), party control of the seat (Labour incumbency) and prior turnout (high, medium and low).

Table 5 shows the effects of treatment interventions, marginality and the interaction between treatment effects and marginality on turnout in the 2010 General Election.[4] Of the interventions, the overall treatment and combination treatments, as well as the telephone and combination interventions in 2010 were 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. We find no evidence that it does by party control either (see Table 5). Again, a number of the treatment interventions have a significant effect on turnout, while those living in seats where there is a Labour incumbent are less likely to turn out. This is hardly surprising given the socio-economic characteristics of many of these constituencies and the general electoral context, where Labour as the governing party were facing a backlash from voters. 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. Given these findings, we therefore cannot reject $H_{0(3)}$ and $H_{0(4)}$.

[4] 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. Given the advantages of using margin (as outlined earlier), we decided to report these models rather than the spending analyses although these models are available on request from the authors.

Table 4. Logit Model of Treatment Effects and Electoral Competitiveness (Margin) on Turnout in the 2010 General Election

Treatment	Treatment		Margin		T*Margin		LL	Cases
	β	Odds	β	Odds	β	Odds		
All Treatments#	0.14*	1.15	0.27*	1.31	-0.11	0.90	-10406.66	17117
Combi (Mail+tel)	0.18*	1.20	0.26*	1.30	-0.10	0.90	-2624.92	4374
Double Combi	0.14	1.15	0.26*	1.30	-0.03	0.97	-2576.69	4261
All Combi 09 +10	0.22*	1.25	0.26*	1.30	-0.07	0.93	-3782.16	6330
Telephone	0.18*	1.20	0.27*	1.31	-0.07	0.93	-3141.81	5234
Mail (Tel accessible)	0.09	1.09	0.26*	1.30	-0.07	0.93	-3658.16	6010
Mail (Inaccessible)	0.15	1.16	0.28*	1.32	-0.26	0.77	-2660.59	4153
All Mail#	0.12*	1.13	0.27*	1.31	-0.15	0.86	-6319.81	10163

#Models with telephone accessible and inaccessible include ‘known’ as covariate.
* significant p= <0.05. Robust standard errors clustered by constituency (n=47). LL = Log Likelihood.

Table 5. Logit Model of Treatment Effects and Labour Incumbency on Turnout in the 2010 General Election

Treatment	Treatment		Labour Incumbent		T*Incumbent		LL	Cases
	β	Odds	β	Odds	β	Odds		
All Treatments#	0.19*	1.21	-0.21	0.81	-0.12	0.89	-10381.63	17117
Combi (Mail+tel)	0.20	1.22	-0.24*	0.79	-0.01	0.99	-2623.05	4374
Double Combi	0.29*	1.34	-0.24	0.79	-0.23	0.79	-2569.12	4261
All Combi 09 +10	0.24*	1.27	-0.24*	0.79	-0.12	0.89	-3773.47	6330
Telephone	0.29*	1.34	-0.24*	0.79	-0.19	0.83	-3131.63	5234
Mail (Tel accessible)	0.10	1.11	-0.24*	0.79	-0.05	0.95	-3653.39	6010
Mail (Inaccessible)	0.12	1.13	-0.17	0.84	-0.06	0.94	-2659.04	4153
All Mail#	0.11	1.12	-0.21	0.81	-0.06	0.94	-6312.71	10163

#Models with telephone accessible and inaccessible include ‘known’ as covariate.
* significant p= <0.05. Robust standard errors clustered by constituency (n=47). LL = Log Likelihood.

Table 6 shows the results of whether the treatment is related to prevailing turnout - in a curvilinear fashion - through the splitting of the sample according to whether the overall level of turnout in the area is high, medium or low. We used previous local election turnout for the 2009 model[5] and prior turnout in the 2009 European elections (from our 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[6]. 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 categorisations of turnout in areas. There is some evidence—in both elections—that some treatment have a significantly greater effect in higher turnout areas. Indeed, none of the experiments show a significant treatment effect in low turnout areas (the reference category) as represented by the main effect. However, in a number of cases the interaction effects were significant, indicating significant treatment effects in high turnout areas. For instance, in 2010, both the overall treatment, double combination treatment and the overall combination treatment show a significant impact on turnout in higher turnout areas. A similar finding was found for the telephone treatment in 2009 although not in 2010, albeit the intervention was only marginally insignificant at the 5% level.

It should be noted here that at the 2009 European election the overall turnout was much lower than in the general election a year later. In 2009 'high turnout' areas saw turnout levels of around 50%, which suggests the results were consistent with the (contingent) curvilinear argument which predicts highest treatment effects where the average propensity to vote is around 0.5 in a mid-salience election (Arcenaux and Nickerson 2009). In 2010 however, turnout rates were considerably higher (around 70%), yet it was still the high turnout areas which saw the largest effects. This is more consistent with the individual level

[5] We are grateful to Professor Michael Thrasher (University of Plymouth) for providing these data

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

equivalent of maximum treatment effects for high propensity voters even in high salience elections (Gerber and Green, 2004).

Overall, for a number of interventions in both elections, there was some evidence that the treatment effects varied by the prevailing level of turnout in the area, with the treatments being more effective where turnout was already high.

Table 6. Logit Model of Treatment Effects and Prior Turnout on Turnout in the 2009 European Elections and the 2010 General Election

Treatment 2009	***Treatment***		***High Turnout***		***Mid Turnout***		***T*High T***		***T*Mid T***		***LL***	***Cases***
	β	***Odds***	***β***	***Odds***	***β***	***Odds***	***β***	***Odds***	***β***	***Odds***		
All Treatments#	0.08	1.08	0.62*	1.86	0.33*	1.39	0.00	1.00	-0.05	0.95	-13775.22	20888
Combi (Mail+tel)	0.18	1.20	0.56*	1.75	0.33*	1.39	-0.04	0.96	-0.14	0.87	-5001.80	7466
Telephone	-0.05	0.95	1.06*	2.89	0.60*	1.82	0.24*	1.27	0.03	1.03	-5677.67	8645
Mail (Tel accessible)	0.08	1.08	0.56*	1.75	0.33*	1.39	0.01	1.01	-0.04	0.96	-6397.19	9546
Mail (Inaccessible)	0.07	1.07	0.74*	2.10	0.34*	1.40	-0.05	0.95	0.01	1.01	-3515.79	5589
All Mail#	0.07	1.07	0.62*	1.86	0.33*	1.39	0.00	1.00	-0.02	0.98	-9914.33	15135
Treatment 2010	***Treatment***		***High Turnout***		***Mid Turnout***		***T*High T***		***T*Mid T***		***LL***	***Cases***
	β	***Odds***	***β***	***Odds***	***β***	***Odds***	***β***	***Odds***	***β***	***Odds***		
All Treatments#	-0.02	0.98	0.42*	1.52	0.24*	1.27	0.24*	1.27	0.15	1.15	-10339.62	17117
Combi (Mail+tel)	0.18	1.20	0.45*	1.57	0.32*	1.38	0.25	1.28	-0.07	0.93	-2611.58	4374
Double Combi	-0.02	0.98	0.45*	1.57	0.32*	1.38	0.36*	1.43	-0.02	0.98	-2561.25	4261
All Combi 09 +10	0.08	1.08	0.45*	1.57	0.32*	1.38	0.31*	1.36	0.03	1.03	-3757.01	6330
Telephone	0.01	1.01	0.45*	1.57	0.32*	1.38	0.30	1.35	0.15	1.16	-3121.84	5234
Mail (Tel accessible)	0.01	1.01	0.45*	1.57	0.32*	1.38	0.09	1.09	0.07	1.07	-3643.90	6010
Mail (Inaccessible)	-0.11	0.90	0.41*	1.51	0.14	1.15	0.21	1.23	0.27	1.31	-2645.53	4153
All Mail#	-0.04	0.96	0.43*	1.54	0.24*	1.27	0.15	1.16	0.16	1.17	-6290.30	10163

#Models with telephone accessible and inaccessible include ‘known’ as covariate.

* significant p= <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). LL = Log Likelihood.

Comparing between elections

So far we have focussed on variation across space within elections. Table 7 compares the effectiveness of the combination treatment across two different elections. We focus here on the combination treatment because the mail and telephone separate treatments are not strictly comparable between elections because 2009 respondents were re-contacted on 2010 (Fieldhouse et al. forthcoming). 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. Whilst there is reason to suppose there may be a curvilinear relationship between salience and the efficacy to GOTV treatments (Arcenaux and Nickerson, 2009) 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 treatment relative to 2010. Second, those receiving telephone or mail treatments in 2010 had also received the equivalent treatment in 2009 which may have led to a reinforcement of the message. Notwithstanding this, 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 of the result do indicate that the treatment may have been more effective at the 2010 high salience election.

Table 7. Treatment effects for combined experiment, compared for 2009 and 2010)

	2009	2010
N treatment	2287	2120
N Control	5179	2352
Voted (treatment)	957	1545
Voted (control)	2058	1620
Estimated Intent-to-treat Effect (standard error)	2.11 (1.23)	4.00 (1.36**)**
Difference in TE=1.89 t-statistic =1.03		

Conclusions

The nationally representative sample allowed us to explore geographical variations in the effect of the treatment. 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. As a consequence, we examined one 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 treated 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 by the marginality of the seat or by party control. However, there was a notable exception to this pattern of consistency. There were a number of interventions—in both elections—where treatment effects were more likely to have a significant impact on turnout in higher turnout areas. Just as some previous research has shown treatments may be more effective amongst regular previous voters (Gerber and Green 2004), at the aggregate level GOTV treatments do appear to be more effective in higher turnout areas. Regardless of the salience of the election, it was in areas with a higher prevailing rate of turnout that the treatment proved most effective, suggesting it may be easier to nudge those already likely to vote than to change the mind of ardent non-voters.

Overall it seems, taking the geography of treatment effects as a whole, it does not much matter where an experiment is conducted: the treatment effects are to all intents and purposes uniform. This has hugely important implications for the external validity of GOTV field studies more generally, especially where single locations are used. It is possible to use these findings to generalise that the effects of single or multi-locations GOTV can be extended to a wide range of locations. However, there is one important exception this. 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 addition 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 many unanswered questions, not least whether larger samples or different electoral contexts might throw up different patterns of variation. Future work based on meta-data could test that 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 in the research programme and we would expect it to conform to the results presented here.

References

Abramowitz, A. (1988), ‘Explaining Senate election outcomes’, *American Political Science Review,* 82: 385-403.

Ansolabehere, S., & Snyder, J. M., Jr. (1996). *Money, elections and candidate quality*. Unpublished manuscript, Massachusetts Institute of Technology.

Arceneaux, K. (2005) ‘Using cluster randomized field experiments to study voting behavior’, *The ANNALS of the American Academy of Political and Social Science,* 601(1): 169-179.

Arceneaux, K. and 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. and Nickerson, D. W. (2010), ‘The cost of convenience: An experiment showing e-mail outreach decreases voter registration’, *Political Research Quarterly,* 64: 858-869.

Browne, William. (2002) *MCMC Estimation in MLwiN.* London: London Institute of Education.

Browne, William, S. V. Subramanian, Kelvyn Jones, and Harvey Goldstein. 2005. ‘Variance partitioning in multilevel logistic models that exhibit over dispersion’ *Journal of Statistical Society A* 168(3): 599–613.

Cutts, D. (2006) ‘Continuous activism and Electoral Outcomes: The Liberal Democrats in Bath, *Political Geography,* 25: 72-88.

Crombie and Davies (2009), *What is Meta-analysis?* Hayward Medical Communications.

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

Rebecca DerSimonian, Nan Laird, Meta-analysis in clinical trials, *Controlled Clinical Trials,* Volume 7, Issue 3, September 1986, Pages 177-188,

Eldersveld, Samuel J. “Experimental Propaganda Techniques and Voting Behavior,” *American Political Science Review,* vol. 50 (March 1956): 154–65;

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.

Franklin, Mark N. 2004. *Voter Turnout and the Dynamics of Electoral Competition*. Cambridge: Cambridge University Press.

Gerber, A and Green, D (2000a) 'The effects of canvassing, direct mail, and telephone contact on voter turnout: A field experiment', *American Political Science Review,* 94(3): 653-63.

Gerber, A and Green, D. (2000b) 'The effect of a nonpartisan Get-Out-the-Vote drive: an experimental study of leafleting', *Journal of Politics,* 62(3), 846-57.

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

Gerber, A. and Green, D. (2005) 'Do phone calls increase voter turnout?: An update', The *ANNALS of the American Academy of Political and Social Science,* 601 (1): 142-154.

Gerber, A., Green, D. and Green, M. (2003) 'The effects of partisan direct mail on voter turnout', *Electoral Studies,* 22, 563-79.

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. and Gerber, A. (2008) *Get Out The Vote: How to Increase Voter Turnout.*2nd edition Brookings Institution Press, Washington.

Green, D., Gerber, A. and Nickerson, D. (2003) 'Getting out the vote in local elections: Results from six door-to-door canvassing experiments', *Journal of Politics,* 65(4): 1083-96.

Green, D., Aronow, P. and McGrath M. (2012) Field Experiments and the Study of Voter Turnout. *Journal of Elections, Public Opinion and Parties*, Vol. 22 No.4.

Green, D. and Krasno, J. S. (1988). Salvation for the spendthrift incumbent. *American Journal of Political Science, 32,* 884-907.

Hillygus, D. Sunshine. (2005). “Campaign Effects and the Dynamics of Turnout Intention in Election 2000.” *Journal of Politics* 67(February): 50–68

Johnston, R. (1987) *Money and Votes: Constituency Campaign Spending and Election Results* (London: Croom Helm)

Jacobson, G. C. (1985). Money and votes reconsidered: Congressional elections, 1972-1982. *Public Choice, 47*(1), 7-62.

Jacobson, G. C. (1990). The effects of campaign spending in House elections: New evidence for old arguments. *American Journal of Political Science, 34*(May), 334-362.

Johnston, R. J. and Pattie, C. J., (2006) *Putting Voters in their Place: Geography and Elections in Great Britain,* (Oxford: Oxford University Press).

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.

Marsh, Michael. 2002. Electoral Context. *Electoral Studies*, Vol. 21(2): 207-217.

Mutz, Diana 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.

Nickerson, D. W. (2007a) ‘Does email boost turnout?’, *Quarterly Journal of Political Science* 2(4): 369-379.

Nickerson, D. W. (2007b), ‘Quality is job one: professional and volunteer voter mobilization calls’, *American Journal of Political Science,* 51(2): 269–282.

Niven, David. 2001. “The Limits of Mobilization: Turnout Evidence from State House Primaries.” *Political Behavior* 23(December): 335–50.

Snijders, T. A. B., Bosker, R. 2011. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage Publication.

Appendix

Table A1: Comparison of Deviance Information Criterion for each Treatment in 2009 and 2010: Random Intercepts Models and Random Slope Models

2009 Treatment Groups	**Random Intercepts Model Only**	**Random Slope**
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	**Random Intercepts Model Only**	**Random Slope**
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