Expanding the use of experiments on civic behavior: experiments with local government as research partners

Sarah Cotterill and Liz Richardson (Institute for Political and Economic Governance, University of Manchester, UK)

Paper prepared for a special issue of the Annals of Political and Social Sciences

Corresponding Author:

Liz Richardson
IPEG, Room 1.6, Humanities Bridgeford Street
University of Manchester
Oxford Road
Manchester
M13 9PL
Tel: +44(0)1612750879
Fax: +44(0)1612750793
Liz.richardson@manchester.ac.uk

Co-author: Sarah.Cotterill@manchester.ac.uk
Biographies:

Sarah Cotterill is a researcher in the Institute for Political and Economic Governance (IPEG) at the University of Manchester. Her research interests are: civic engagement, citizen participation, social networks and public sector partnerships. She is currently undertaking a number of field experiments to study civic behaviour, as part of the Rediscovering the Civic and Achieving Better Outcomes in Public Policy project.

Liz Richardson is a researcher in the Institute for Political and Economic Governance (IPEG) at the University of Manchester. Her research interests are: community engagement in decision making and user involvement in public services; civic renewal and civil society; public service delivery; neighborhood governance structures. Her research has had a strong policy and applied focus. Liz is author of DIY Community Action: neighborhood problems and community self help (Policy Press 2008).

Suggested pull quotes:

“Some harsh realities about doing experimental work with local governments are revealed, not to argue against experiments, but to try to ensure that these harsh realities can be negotiated to deliver better quality local research through the increased use of trials” (p.5).

“Joint research with local practitioners and policy makers is not just a cash cow, but offers opportunities to engage in complex specialist interventions, undertaken at a significant scale, with citizens, in field conditions” (p.5).
Abstract

Despite the acknowledged benefits of experimental methods and interest in evidence-based policy, there have been very few attempts to undertake experiments at a local level in co-operation with policy makers and public service providers. This paper discusses the challenges of implementing and delivering experiments in these contexts, drawing on four successfully completed experiments, two current experiments and referring to four proposed experiments which did not take place. The completed and on-going experiments investigated: increasing household recycling, internet deliberation, changing children’s environmental attitudes, community support to drug users, pledging to donate to charity; and the response of councilors to interest groups. The paper concludes by assessing the benefits of co-produced field experiments against the challenges, and argues that, overall, co-production has the potential to expand the scope, range and depth of possible experiments, improve access to citizens and to data, and leads to research that is more credible to policy makers.

Key words: civic behavior, field experiments, local government, difficulties of implementation
Introduction

Despite the acknowledged benefits of experimental methods and interest in evidence-based policy, there have been very few attempts to undertake experiments at a local level in cooperation with policy makers and public service providers. In the last four years researchers at the University of Manchester have done just this by using experimental methods to research civic behavior. The University was offered a number of research commissions by local governments to explore ways of encouraging greater civic behavior, with the methodology left open for discussion. Therefore, these commissions offered an opportunity to persuade local research funders to do experiments. However, while there has been a strong emphasis in the literature on how to conduct rigorous and robust experiments (e.g. see Torgerson and Torgerson 2008), it can be hard to implement these recommendations if potential partners do not fully understand the method. Help on this first stage is missing in the academic literature, which focuses on a specific set of issues around research methodology and tends to leave out the wider context and partnerships with public sector actors. There is far less material on how to persuade partners to do field experiments, how to select appropriate interventions and outcome measures, how to overcome practitioners’ practical and ethical objections, and how to manage governance arrangements with delivery partners. It is these challenges that tend to take up much energy, limit the scope for expanding the use of the method, and can even prevent experiments from taking place.
Locally commissioned work is a common source of research funding for many Universities, but is often the least methodologically advanced, dominated by poor quality, non-experimental program evaluation. Joint research with local practitioners and policy makers is not just a cash cow, but offers opportunities to engage in complex specialist interventions, undertaken at a significant scale, with citizens, in field conditions. Given the value of experimental methods, and in many cases their methodological necessity, combined with the research scope offered by locally commissioned work, this paper attempts to fill some of those gaps in the literature so that these potential obstacles do not get in the way of the use of experimental methods. Some harsh realities about doing experimental work with local governments are revealed, not to argue against experiments, but to try to ensure that these harsh realities can be negotiated to deliver better quality local research through the increased use of trials.

This discussion paper looks at six key practical issues and challenges that researchers can encounter and how they might be overcome or dealt with, drawing on the experience of commissioned and commissioning experiments at the University of Manchester. Figure 1 summarizes the four completed, and two current experiments, plus four failed proposed experiments referred to in this paper.

Figure 1. Brief summary of the experiments discussed in this paper

<table>
<thead>
<tr>
<th>Completed Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Household recycling</td>
</tr>
</tbody>
</table>
be canvassed. Half of the streets were placed in a control group and received no special attention. All households in the canvass group were visited by canvassers who were trained to promote and encourage recycling. Recycling participation rose by 5% immediately after canvassing, but three months later the effect had diminished to 2% (Cotterill, John, and Liu, 2008)

| 2. Internet deliberation | An RCT examining how randomly-selected citizens behave when given the opportunity to participate in a large-scale online consultation event. Participants in the experiment were randomly assigned to three groups, each of 2000 people. Members of the deliberation group (Treatment 1) were able to discuss policy-relevant topics with each other in a moderated on-line environment. Treatment 2 could view the site but had no opportunity to interact with other individuals. A control group had no access to the site. All participants, including the control group, undertook surveys before and after the online event. |
| 3. Changing children’s environmental attitudes | An RCT which evaluated the impact of school-based education programs on environmental attitudes and behavior of 715 primary school students and their families in 27 primary schools. The outcome measures were obtained through pre and post surveys that were completed by the students in-class and at home. The results show no significant differences between schools (Goodwin, Greasley, John & Richardson, 2009) |
| 4 Community drug support project | The community drug support design experiment was conducted in a deprived neighborhood. It provided support to ‘problem’ substance-users. The intervention focused on changing the lifestyles of users as well as reducing drug use. The intervention group was matched with a comparison group of substance users receiving only mainstream treatment. The treatment group performed better on health outcomes, relationships and substance use (Askew, John, & Liu, 2008) |

Current Experiments

| 5. Pledging to donate to charity | A RCT to establish whether making a pledge increases the likelihood that an individual will make a charitable donation of a used book, mobile phone or glasses, compared to those in a control group who are asked to donate without making a pledge. |
| 6. Response of councilors to interest groups | A clustered RCT of lobbying of local councillors in England and Wales, where the councillors randomly receive different letters from citizen interest groups recruited into the study. The two treatments are differently worded letters that are sent by a lobby group to the councillors, which are generated from a random sample of half of all councillors in a local government, then randomised into two treatment groups. There are 200 councilors in each of the two treatment groups (Richardson and John, 2009). |

Planned experiments that did not proceed

| 7. Citizen turnout at community | An RCT to investigate whether canvassing can affect turnout at a planned community consultation event. The treatment group would receive information about the event in person. The control group would }
8. Reducing car use on journeys to work & school

An RCT exploring different methods of persuading people to use alternative transport methods for school trips or to work, reducing car use.

9. Binge drinking

A RCT to test out a relatively new, targeted, motivational counseling approach called ‘brief interventions’, with young people engaged in harmful drinking but who were below the threshold for statutory support, compared to a control group receiving no special treatment.

10. Home energy use

An RCT exploring different methods to persuade households to reduce home energy use.

The first half of the paper deals with three issues relating to the use of experimental methods when conducting field experiments on civic behavior: finding appropriate outcome measures; the challenge of researching collective behavior; and the benefits of mixed methods research. The second half of the paper looks at three issues which arise when working with local partners and practitioners: the difficulty of persuading local government to commit resources to experiments and to coordinate with a range of delivery partners; moral and ethical issues objections; and the importance of good research governance.

**Outcome measurement for civic behavior**

The first challenge of using experimental methods is identifying a suitable outcome measure. Field experiments require an outcome measure that is sensitive, reliable and valid. The outcome measure should be sensitive to the important effects of the
intervention. It must be reliable in the sense that a repeated measure in unchanged circumstances will give the same results. It has to be a valid measure to accurately assess the outcome of interest (Torgerson and Torgerson 2008, 147). Our research projects on civic behavior have looked for outcome measures that fulfill these criteria, and which in addition are: clearly observable and measurable; can be identified across a large sample population; cover the widest range of civic behaviors and are relevant to public policy. This has proved to be a relatively stretching set of criteria, as outlined below.

Observable and Measureable Behaviors

Voter turn out is a behavior that can be clearly observed across a population, and in countries with established electoral rolls, administrative data can be used to establish who did or did not turn out to vote. Papers presented at the Field Experiments in Comparative Politics and Policy conference in 2008 discuss a number of experiments where voter turnout was used to measure the impact of poverty relief funds (De La O Torres 2008), public deliberation (Wantchekon 2008), public meetings for women (Green 2008) and there is a substantial body of research on the effect of canvassing on electoral turnout in various parts of the world (Green and Gerber 2008). An adaptation of this is the use of vote share between rival candidates as an outcome measure (Hyde 2008). Signing a petition is another observable behavior, and when conducted on-line, results are readily available (Margetts and John 2008).

Observable outcomes are preferable to self-reported outcomes. In the absence of observable measures it is sometimes necessary to accept self-reported outcomes from
surveys or interviews, usually administered before and after an intervention to test for change. These self-reported views have been successfully adopted in studies of citizen knowledge, attitudes and behavior on the environment (Hooghe and Vissers 2008; Mahéo and Stolle 2008), and politics (Moehler 2008). Two issues arise when using before and after surveys in experiments on civic behavior. Firstly, questionnaires asking people to self-report their civic activity are likely to lead to over-reporting, for example, recycling surveys have been found to be biased, over-reporting the extent of recycling activity (Shaw, Lyas, Maynard, and Van Vugt 2007; Timlett and Williams 2008). This may not be a problem if both treatment and control group over-report their activity, but where available, observations of behavior are a more accurate measure. Secondly, there is a danger in field experiments that the group who have been treated will be more aware of the expected response and therefore be more likely to over-report their behavior than those in the control group, leading to bias in the results.

Wanting an unbiased measure is not peculiar to experiments on civic life, and neither are interventions trying to change behavior as well as attitudes. But changes to enhance democracy need to be focused as much on what people actually do and how they behave, as what they think or feel. At its heart, enhancing civic life involves citizens in active roles in relation to decision making as well as questions of trust and other opinions. Therefore, if available, an observable behavioral civic outcome is preferable.

_A large sample population_
Small N designs for field experiments have been used successfully, for example, to test forms of mobilization to increase voter turnout (Addonizio, Green, and Glaser 2007). Nevertheless, the focus for the research outlined here was to attempt to use large N designs because this allows the research “to estimate effects with precision, to estimate reliably differential treatment effects across subgroups and varying contexts, and to map out the apparent function that links dose to response across subject- and setting-types.” (Green 2007, 1). As a result, the proposal for experiment 7, an experiment on citizen turnout at community events, was ultimately rejected by the research team because, based on existing national data, the volume of people who would attend a community event was estimated to be too small to produce adequate statistical power. Interestingly, Green supports an “ample N” (p1) as a desirable element of the study “in the absence of practical/ethical constraints” (p1), and makes a strong case for turning design problems in the real world into research opportunities, which could translate into more creative use of small N designs. Arguably the research team could look at small N designs with a closer eye in future.

A wide range of civic behaviors

The differential availability of large N observable behavioral outcome measures could create a bias towards the more easily measurable behaviors such as voter turnout, vote share, campaign contributions and group membership, with field experiments being concentrated on these behaviors rather than others that are harder to measure. This is important because the more easily measurable behaviors tend to fall within a repertoire of more conventional, mainstream or traditional forms of civic activity and collective action.
(voting, membership of groups, petitioning). These behaviors are still activities that the majority of citizens are more likely to engage in, and are crucial for democracy. However, they have been shown to be in decline. Regardless of trends in types of civic activity, they do not capture the full range of civic acts that citizens might choose to undertake. The forms of civic activity that are less easy to source or collect data on include some of the newer forms of individualized political participation, which has been shown to be a growth area in participation (e.g. see Micheletti 2003 and van Deth 2009), and therefore experiments run the risk of missing out on work on newer forms of popular participation. These new forms include: political consumerism (buying or boycotting products on political grounds); other individualized political acts such as wearing a badge or sticker about a cause; ‘pro-social’ behaviors by individuals like eating more healthily, behaving in an environmentally sustainable way, or parenting children more responsibly; participatory methods using new technology; and forms of citizen governance which involve local people in decision making.

These newer forms of civic activity present fresh challenges in finding suitable outcome measures. The experiments reported here have attempted to find suitable measures for some of these newer behaviors, with varying rates of success. Observing whether a bin has been set out is a robust measure of recycling behavior (experiment 1), but we have had less success in finding suitable measures of other environmental behaviors. One failure was Experiment 10, reducing home energy use, where the proposed outcome measure was actual gas and electricity use in the home, but where it was too difficult to find a way to obtain utility meter readings or copies of utility bills without accusations of
invasion of peoples’ privacy, and without the measurement becoming itself a form of intervention for the control group. Administrative data is potentially very useful here, but these are confidential records and often subject to access restrictions. For example, experiment 8, an experiment on reducing car use on journeys, failed to go ahead because of lack of access to data in two different settings: the University, were unwilling to provide access to data on car park usage which would have allowed a study on journeys to work; and schools would not release addresses of parents for an experiment to persuade parents to walk their children to school rather than drive. Recent research on what motivates people to complete internet petitions shows the future potential of undertaking experiments on newer forms of e-participation (Margetts and John 2008).

In trying to bring all of these criteria together – large N, different types of observable civic behaviors - the field experiments in Figure 1 indicate some successes, and some failures. One successful experiment that brought all the criteria together was Experiment 1, on curbside recycling. Observations were made of which households put out a recycling container for collection (Cotterill, John and Liu 2008). A monitor sat in the recycling vehicle while the crew were working and noted all the houses on the street that had placed recycling material outside the house boundary. The monitoring was repeated over three consecutive weeks: and any household who recycled at least once in the three week period is counted as a recycler, (WRAP 2006). Other experimental studies have used observed recycling behavior in a similar way as an outcome measure (Shaw 2008), sometimes also measuring the weight and variety of recycled materials (Lyas, Shaw and Van Vugt 2004; Schultz 1998).
Policy relevant outcome measures

While the criteria of large N designs on a range of observable conventional and non-conventional civic behaviors already presented a challenge, the selection of outcome criteria was further complicated by policy considerations. Partnering with local government meant that it was important to start with the policy problem to be tackled, not with the method per se. Other writers on field experiments have described one of the functions of experiments as being able to communicate messages to policy makers, or “whisper in the ear of princes”:

“To understand the recent emergence of experiments, we invoke economist Alvin Roth’s three purposes of experiments: to search for facts, speak to theorists, or “whisper in the ears of princes.” Using this classification scheme, we suggest that the rise of experimental designs in political science reflects both an increased demand to test causal claims that existing methodologies could not fully address and an increased capacity of experiments to simulate relevant political contexts.”

(Druckman, Green, Kuklinski and Lupia 2006, 627)

A forthcoming experiment by the University of Manchester has an objective outcome measure of a non-conventional civic intervention for individual citizens, which is also of significant policy interest by central government. This experiment, Experiment 5 described in Figure 1, tests the idea of pledging, where citizens make a written commitment to undertake certain behaviors.
Interventions to test individual and collective civic behavior

A second practical challenge when conducting experiments is to find an intervention to increase civic activity that can be applied and randomized at the individual level. Civic behavior can be undertaken either as an individual or within a group. Individual behavior includes voting, signing petitions, complaining, recycling or boycotting consumer goods on political ethical grounds. Collective behavior is where people work together by forming a neighborhood association, setting up a social enterprise or joining a political party. The importance of the distinction between individual and collective behavior is that often civic activity is not a lone activity, but is about citizens working together to effect change. This has two particular implications for field experiments. The first implication relates to outcome measurement, because it is difficult to measure collective outcomes at an individual level. An example is where a local authority takes measures to improve community cohesion by encouraging inter-faith and inter-ethnic understanding; the outcome measure is improved relations and more interaction, which is not the property of an individual, but a group level effect.

The second implication of the often collective nature of civic activity when doing experiments is that local interventions to increase civic behavior are, by their nature, often widely marketed to a general population rather than tailored towards targeted individuals. One current example is the widespread adoption in the UK of participatory
budgeting, where citizens are invited to take part in decisions on how local public funds are spent. Another example is the promotion of neighborhood agreements in which service providers and residents commit to a list of responsibilities towards the neighborhood, for example for street cleaning services to ensure dog mess is cleaned up promptly, and for residents to agree to clean up their own dog mess. The successful introduction of these collective approaches requires high profile events, activities at local schools and community centers and multiple types of publicity to work effectively. It is counterproductive to the chances of producing results to handicap the intervention by not doing widespread marketing. An individual or household level intervention could not be expected to provide enough momentum to achieve success. These initiatives which are marketed to a wider population at a collective level make randomization of individuals or households very difficult to achieve.

One option in these circumstances is to use the neighborhood or local authority as the unit of intervention and measurement, but an experiment then needs so many neighborhoods to achieve a statistically significant result that it can become unfeasible or too expensive. One way round this, as shown in an experiment measuring the effect of radio broadcasts on communal behavior (Paluck 2009), is to find pairs of communities that are matched on relevant factors and randomly allocate between treatment and control groups. Another way to deal with situations where the treatment is collective is to adopt a cluster design, where schools, neighborhoods or councils form the unit of randomization, rather than individual students, households or citizens (Torgerson and Torgerson 2008, 99–107). One example of this is current research by the University of Manchester in
Experiment 6, testing out the effectiveness of different approaches by interest groups to local councilors, where a clustered research design has been used.

Another related risk for experiments on civic behavior is the danger of contamination or spillover between the treatment and control groups. This risk can be successfully overcome by taking account of them during careful research design. In experiment 1, on household recycling, a canvassing intervention was delivered to streets rather than individual households because of concerns that if one household was persuaded to recycle, there would be a peer effect on their untreated neighbor ( Cotterill, John and Liu 2008). Even with a street based design there is a risk that households on one street might be contaminated by the influence of others on nearby streets, particularly those who live on small side roads and tests were run to rule this out. In future street based designs, “buffer” streets could be introduced to create barriers to contamination or spillover. A current experiment, experiment 6, which seeks to influence the behavior of local politicians, faces a risk of contamination if a politician receiving one treatment consults with a colleague receiving a different treatment.

Mixed methods

A third practical methodological challenge has been to fully understand how an intervention is, or is not, working. A randomized trial is a gold standard method for measuring whether or not a particular intervention works better than doing something
else or doing nothing. Clearly one of the benefits of the experimental method is that it allows for the attribution of causality. The identification of the causal mechanism is only part of the picture of if, as well as how and why, an intervention works. A theoretical framework and analysis plan specified at the research design phase can be used to generate hypotheses about how and why. The CONSORT guidelines are a valuable tool in achieving this (Moher, Schulz and Altman 2001).

A mixed methods approach, where qualitative methods such as focus groups and interviews are conducted alongside a field experiment trial, can provide valuable information on how or why a particular intervention worked the way it did. One lesson from the experiments described in this paper is that the use of mixed methods alongside or within the trials can give an added richness to the research. Experiment 1, on the impact of canvassing on recycling, was enriched by observational and interview data which suggested that the research needed to take account of changes in the quality of service delivery over the course of the experiment. The Universities of Manchester and Southampton have recently completed a large scale randomized controlled trial to test whether participation in an internet deliberation has an impact on participants (experiment 2). Pre- and post- surveys will be analyzed to see if the attitudes of participants in the deliberation change over time, compared to a control group who did not have access to the deliberation and an information only group who received the same materials but did not participate in the deliberation. This assessment of causality will be enhanced by qualitative analysis of the deliberative discussion, examining how randomly-selected citizens behave when given the opportunity to participate in a large-
scale online consultation event and the extent to which their interactions be understood as ‘deliberative’. The qualitative analysis will look at the deliberative quality including the inclusiveness of the debate, respect for the opinion of others and whether arguments are justified by evidence. Qualitative methods can also offer a useful quality check to ensure that the intervention is implemented in a similar fashion across different sites.

**Local governments’ roles in conducting experiments**

Having examined the challenges of using experimental methods when researching civic behavior, the second half of the paper will discuss the challenges of persuading local partners and practitioners to work with researchers on experiments. The fourth challenge explored is the difficulty of persuading local governments to commit resources to experiments, and of coordinating a large number of potential partners to deliver experiments. One obstacle has been persuading local governments as potential partners to commit to the additional work involved in doing an experiment. Unlike elites, local resources tend to be devoted for local benefit. It is often the case that local practitioners and policy makers feel that they would derive the maximum local benefit from devoting resources to those participating in a program, that is, ‘their’ clients or users. More robust experimental research which requires extra resources to collect data on those not in the program is frequently seen as a luxury extra which local resources should not be responsible for covering. In discussions, some practitioners wanted to benefit from experimental evidence, but did not feel they should be the ones to pay. The additional
costs needed to undertake experiments were not seen as the responsibility of any single local authority or public sector provider when this type of contribution to the evidence base was one that all local public sector organizations could benefit from. So even where people agreed with the intellectual case for the counterfactual, they ultimately chose to use non-experimental methods.

A related issue has been that doing experiments with local government means working with large numbers of different bodies, each with its own authority to say yes or no to the experiment, to impose conditions, and accept the structure of the intervention, not to mention the risks of inconsistent implementation even after significant numbers agree in principle. Attempts to conduct Experiment 9, on binge drinking, failed to get off the ground precisely because of problems in getting a number of different local agencies in different local governments to agree to participate. The idea of the experiment was to test out a relatively new, targeted, motivational counseling approach called ‘brief interventions’, with young people engaged in harmful drinking but who were below the threshold for statutory support. However, this required the co-operation in each local authority area of the referring agency which was usually the police, but could also be a local government community safety team, the strategic partnership overseeing provision of relevant services (itself made up of a number of different public sector bodies), and the commissioned service delivery body (usually a voluntary sector body with a contract to provide public services). It required those services to a) work together with new partners on referrals and provision of services, b) pilot a wholly new intervention and approach, and c) alter the criteria for which clients would receive the service (both to lower the
thresholds for support, and to randomly allocate the intervention). Not only that, but in order to generate sufficient numbers, the researchers needed to persuade at least three different local authority areas to participate. Unsurprisingly, after 18 months of negotiation, the proposal to undertake this experiment collapsed. During Experiment 3, on the impact of an environmental education program on the environmental attitudes and behavior of children and their families, several different bodies did agree to participate but there were then problems of inconsistent delivery across the partners. The delivery of the intervention was done by a single worker, to encourage consistency across a number of schools, but there was some variation in the completion of the survey as this was overseen by the schools themselves, each of which took a different approach despite attempts to standardize through thorough briefings to schools.

**Randomization, resource allocation and risk - moral and ethical issues**

The fifth challenge has been to overcome numerous moral and ethical objections made by practitioners to experimental methods. Other literature has highlighted moral and ethical barriers to the implementation of experiments, for example Ann Oakley’s work illustrates conflicts for midwives in one study between the idea of random allocation and the principles of their midwifery training (Oakley 2005, 235). In some of the attempts to create experiments there were similar obstacles from practitioners, particularly where the trials dealt with health and social care providers on high-intensity services to vulnerable individuals. These public services are areas where classic models of a gate-keeping
bureaucracy apply. As a result of limited resources to deal with many needy clients, the allocation criteria for resources and mechanisms for resource allocation are key debates for professionals. Professionals have a strong public service ethos that they are there to help the neediest. Although many acknowledge the unintended inequalities resulting from bureaucratic needs-based allocation systems, and that needs-based systems can be distorted or manipulated by those seeking help, there is nevertheless a strong reluctance to accept a truly random allocation system as an alternative rationing system. Fundamental ideas of fairness still hold most sway, and fairness based on need rather than ‘chance’.

In the aborted binge drinking experiment (Experiment 9) described above, the logic of a needs-based allocation argument meant that practitioners preferred to deny services to at-risk young people than accept that only some would receive the experimental treatment. This was not on the grounds that the treatment was unproven and potentially ineffective or even counter-productive, but on the grounds that it might be likely to be effective. Practitioners argued that their policy was not to deny clients support if available, therefore their organization would be more liable by making a treatment available than not offering it. A group of up to 150 young people had been identified by the police as at risk for alcohol-related but low-level criminality, but below the threshold for other forms of support or intervention. Although the experiment would have levered in new resources for some of those young people, ultimately none of them received treatment from the local service responsible for tackling alcohol use by young people.
Some writers have argued that random allocation is used as part of a staged or stepped design and therefore acts as a form of rationing in the same way as a waiting list (Toroyan, Roberts and Oakley 2000). Most public policy interventions involve selective roll out, or pilots in specific areas or with small numbers of clients, which could be seen as comparable to selection in experiments. This argument was successful in the environmental education experiment (3), as the delivery organization had ongoing resources to offer the program over a longer period than the experimental trial. However, in the community drug support field experiment (Experiment 4), there was no guaranteed ‘second stage’ as these were high risk pilot programs with no guarantees of further funding. With the chances of special funding at stake, the case against experimental methods was even stronger, as anyone missing out at this stage would be unlikely to get another opportunity. In this field experiment a quasi-experimental match comparison group was used, which was an inferior method as it did not adequately deal with the issues of endogeneity or selection bias, and therefore were less convincing to decision makers when allocating resources. The researchers counter-argument was that ‘denying’ treatment to some in a fully experimental method may, paradoxically, have given the opportunity to more to receive the treatments in the future if they were found to be effective, as the case for additional resources would have been stronger.

In a broad classification of civic behavior, pro-social behavior, ‘good’ parenting, and appropriate use of alcohol and drugs are all civic behaviors by individuals. However these are all areas which overlap with health and social care interventions where statutory and legal frameworks require assessment of risk and treatment. Therefore it was harder
to convince practitioners of the benefits of randomization where statutory frameworks specified that individuals at risk of harm must be treated. For other types of behavior, more traditionally viewed as civic, there were fewer objections on moral and ethical grounds. One example of this was Experiment 6, the interest group lobbying project, where ethical approval was negotiated from the University without needing to gain consent from participants. This was on the grounds that the local politicians lobbied were public figures with official obligations to respond to constituents, who could reasonably expect their responses to citizen interest groups to be shared publically with community members and other stakeholders. Here, the very public nature of civic participation, emphasis on transparency of political decision making, and concern for a more citizen-responsive democracy all worked in favor of the needs of experimental method for participants to be unaware of the treatment. Another persuasive line of reasoning was an analogy between the project and ‘mystery shopping’ of private and public sector services’ customer relations. The fact that the experiment was in the field of civic behavior helped, as studies in civic participation have regularly focused on user engagement in improving public services including user audits such as mystery shoppers. Another positive aspect in obtaining ethical approval was that this was an experiment truly in the field, with an assurance that all issues in the lobbying would be genuine local concerns.
Governance

The sixth and final challenge has been in the governance and management of projects. As with other forms of applied research, the field experiments required complex sets of governance arrangements to be created between the researchers and the practitioners. Indeed, two of the criteria set out for a design experiment to take place were: clear governance arrangements, including a clearly defined and stable governance group who own the design experiment and have the authority to implement change; jointly led-delivery, with researchers and practitioners working as a team. Partnership agreements were signed for several of the field experiments, although these were voluntary agreement with no legal force. Good governance was crucial given that the researchers were not directly controlling the implementation of the intervention, unlike other experiments where the intervention has been undertaken by canvassers or others employed by the academic institution. But governance went beyond discussion about details of delivery of the intervention. Co-production of experiments inevitably raised questions about joint ownership of the research itself – the instruments, data collection, analysis and interpretation, and results of the jointly-led work.

Where the researchers and partners did not put adequate governance arrangements in place, the projects experienced problems, for example around presentation and disputed interpretation of some negative results. The context for this was a decision by the local government organization whether to continue funding the service delivering the treatment in a context of public sector funding cuts. As an independent academic organization, we
struggled to balance the needs of the research for independence and objectivity, with the needs of the delivery partner for positive results to lobby within their own organization for funding. Where partnership arrangements were in place there has been greater consensus over results, even negative results, because delivery partners have had warning of poor outcomes at an early stage, with an opportunity to debate potential weaknesses in research instruments, data collection or other methodological issues.

More important perhaps has been that practitioners have used the experimental method as a way to reduce the risks to themselves in their own organizations when attempting a new approach. The risk of failure of an intervention is transferred from the practitioner over to the research project, which has been a positive draw for practitioners wanting to use the approach. Local government in the UK has experienced many years of modernization and improvement policy designed to make public spending more efficient. The modernization agenda has led to pressure on local government to make sure that they achieve certain outcomes, and not waste resources by getting things right first time. At the same time, there have been conflicting pressures to improve efficiency by doing things differently, innovating and risk taking. Practitioners working with the researchers have seen a benefit of field experiments in allowing them to try out new interventions without a presumption that they will definitely be successful first time around. Experiments have created spaces for practitioners to literally experiment; experiments allow them to get things wrong, fail, learn and adapt in light of the results.
Conclusions

This paper has described a series of challenges in doing field experiments on civic behavior with local governments as research partners. Suitable outcome measures can be hard to find. The collective nature of civic interventions can militate against individual randomization. Resource-intensive mixed research methods are necessary to get full value and understanding out of the projects. There are ethical and moral objections from principled public service practitioners. Significant time and skill must go into setting up and managing governance arrangements and structures. Co-production means partners want to have equal say over the research methodology, and negotiations cover both the research and the intervention. It is not enough for researchers to be convinced of the experimental method; it must also be feasible for them to implement the method. More literature on the experience of implementation may assist researchers to make informed choices, and to continue to work to overcome a series of predictable hurdles and obstacles.

These challenges are described from the perspective of the researchers. But as this paper has shown, practitioners in local governments have their views, and need to be convinced if more local field experiments are to take place. The researchers faced many thoughtful and relevant questions from partners and potential partners about the practical and ethical issues, some of which were resolved, but others that were not. The vast majority of these questions were not about the fundamental value of experimental methods in producing the most reliable results, although this debate took place. Instead there were many
practitioners committed to the idea of experiments, and keen to work with researchers using the approach, but several of the attempted experiments failed where they could not resolve very practical implementation problems.

As was argued at the beginning of this paper, the aim of documenting the complexities faced in doing the trials was to fill some of the gaps in the literature on doing experiments, so that the potential obstacles do not get in the way of the further use of experimental methods in a wider range of contexts. Its key message is to do more co-production of field experiments. The value of the method remains, and the benefits to researchers have been to vastly expand the scope of experiments that could be done, and to make any work potentially more credible to policy makers by being conducted under authentic field conditions. The experiments have been able to tap into specialist skills to deliver complex interventions around substance use and family work. The range and depth of interventions open to experimental work has been increased through the use of partners’ existing resources. Working with partners has enabled access to citizens and data. For practitioners, co-production has reduced the risks associated with innovation. Several of the experiments would simply not have been possible otherwise, and this creates rewards that go some way to balancing the costs of an often messy and difficult process of co-production of field experiments with local governments.
References


---

i During a design experiment researchers work with practitioners over a period of time to design, implement, evaluate and re-design an innovative intervention. The group receiving treatment is compared to a comparison group. The aim is to perfect the intervention over several iterative cycles until it is the best it can be. (Stoker and John, 2008)

ii The UK government is currently planning a pilot program of community Pledgebanks and commissioned the authors to complete a literature review on this topic.

iii Rubin’s stable unit treatment value assumption (SUTVA) is relevant here: for causal inference it is important that one unit’s outcomes are not affected by the assignment of another unit to a treatment or control group (Gerber and Green 2008).

iv Mystery shopping is a tool to measure service quality. Mystery shoppers pose as customers and then provide feedback on their experience. It is widely used in retail and has spread to public services.