30 October 2014

Evaluating democracy & governance projects with randomized control trials: J-PAL and the state of the art

Paper prepared for FHI360

Harry Blair (harry.blair@yale.edu)

In the present decade, as USAID and other donors have increasingly demanded rigorous quantitative evaluations of democracy and governance (DG) programs, the randomized control trial (RCT) methodology has come to be the "gold standard" for conducting evaluations, and the Abdul Latif Jameel Poverty Action Lab (J-PAL) has become widely recognized as RCT's leading practitioner. The present paper will explore the following themes in this connection:

- The RCT methodology.
- The ascendancy of the RCT as the "gold standard" for DG evaluations.
- J-PAL's track record in DG evaluations.
- Using RCT techniques in civil society evaluations.

RCT as an evaluation methodology

The randomized control trial was first introduced in medical research in the late 19th century and then spread to agriculture in the early 20th century. In medicine, the clinical trial of new drugs and therapies has long been the standard for determining efficacy. The typical test for a new drug is a familiar one today: recruit a randomly selected sample of patients, who would then be divided into two groups matched by medical condition, age, gender, ethnicity and such other characteristics as desired, with one group becoming the "treatment" patients and the other becoming the "control." The treatment group receives the experimental drug, while the control group takes a placebo. None of the patients know which group they are in, nor do the medical personnel implementing the test and observing them (the "double-blind" aspect of the trial). Before the treatment and again at its end, all the patients are examined to determine in what ways if at all the treatment group now differs from the control.

In the social sciences, RCTs have been used for many decades, for example in education (how would addition of a second teacher in the classroom impact reading skills in a third grade setting?) or marketing (how would the addition of a children's menu affect sales in a fast food restaurant chain?). Of course, the "double-blind" component would no longer be possible (everyone in the treatment group would know it, as would the project implementers and monitors). But it is possible to separate the groups from contact with each other during the trial (use different schools and different restaurant locations). In economics, RCT experiments have long been important research tools (Levitt and List 2009), but in political science they have come into significant use only recently, as quantitative analysis and "large *n*" datasets have come into more prominence.

Conceptually, the RCT offers a combination of simplicity and elegance, as shown in Table 1. If the samples have been selected with proper randomization between the treatment and control groups, the baseline data gathered at the outset of the experiment (often called an "intervention" for the

treatment group) will show that the test variables measured for the two groups (math skills, total restaurant sales) have an equivalent profile in terms of averages, standard deviations, and the like (or in other words A1 = B1 in Table 1). After the intervention, the variables are again measured, and the differences (A3 vs. B3 in Table 1) are analyzed.

RCT's ascendancy in DG evaluations

Up until quite recently, virtually all project monitoring and evaluation (M&E) in the DG sector were based on the traditional methodology of document review, key informant interviews, focus group sessions, and field visits. But as demand within USAID and other donor agencies has increased for evaluations focusing not just on project outputs but also outcomes and then impacts, evaluators have responded with increasingly quantitative indicators.

A key development here was the 2008 publication of a lengthy (337 pages) report commissioned by USAID with the National Research Council (NRC 2008), titled *Improving Democracy Assistance: Building Knowledge Through Evaluations and Research*. The report argued for more rigorous quantitative evaluations in general and for randomized studies with controls in particular. Despite considerable doubt about how amenable DG initiatives could be to RCT evaluations (e.g., Kumar 2013: 87-115), the methodology gained adherence within the DCHA/DRG Center, which launched several RCT studies in the ensuing years. Within several years, many began referring to RCT as the "gold standard" of project evaluation. But to conduct a thorough RCT evaluation of a project takes considerable time, and as of summer 2014, only one thorough study had been completed (Baldwin and Muyengwa 2014).

Fortunately, the J-PAL group, which formed at MIT in 2003 as an economics research unit, had already undertaken a wide range of quantitative analyses, most of which involved RCT approaches.[1] By late 2014, the J-PAL center had completed more than 560 evaluations, including more than 100 in what it called its "Political Economy and Governance" theme, as shown in Table 2. Although about 20% of the evaluations were conducted in Europe of North America, the vast majority focus on Africa, LAC, and South and Southeast Asia. Table 3 provides a breakdown of the Political Economy and Governance evaluations by location and topic. Almost a third dealt with developed countries (35 of the 117), and of the remaining 82, something over half (46 from Table 2) took place in South and Southeast Asia, with a majority of the rest (20) in Africa, leaving only 12 in LAC.

Table 3, which breaks down J-PAL studies in the Political Economy and Governance sector, shows a wide variety of topics. About a third (41 of 117) were concerned with elections, most (26) of them in advanced countries, typically short-term interventions aiming to increase voter turnout). Roughly one-fifth dealt with public service delivery – mostly education and health, but some welfare programs, housing, roads, etc. Next came various initiatives promoting community participation in public decision-making and revenue generation, followed by local governance, women leaders, corruption and local fund raising.[2]

[1] Two of J-PAL's principal founders have provided an account of the organization's history and achievements in Banerjee and Duflo (2011).

[2] I created this taxonomy among the 117 Political Economy and Governance evaluations. Some categories were obvious (elections, revenue generation), but others proved difficult (e.g., service delivery vs. community participation, local governance vs. women leaders), so the resulting table should be regarded as less than totally exact.

Other individuals and organizations have also conducted RCT evaluations, though J-PAL remains the leader in the field. The Innovations for Poverty Action (IPA) program at Yale, dating from 2002, has conducted scores of development projects and evaluations, though the vast bulk of their work has taken place outside the DG sector.[3] The International Initiative for Impact Evaluations (widely known as 3ie) has collected a repository of more than 2400 evaluations of developing country interventions, from among which a search for "governance" reports revealed some 196 dating back to 1995. These covered a very wide range, as is clear in Table 4, and appeared to include any experiment involving a state agency (e.g., a scholarship program to reduce school drop-outs in Indonesia).[4] Yet another list of randomized DG evaluations has been provided in Moehler's 2010 article (Moehler 2010). But the present exploration will be confined to J-PAL's work. At some future point, a thorough search of IPA's and 3ie's evaluations would produce much of interest to the present exploration.

J-PAL's track record in DG evaluations

Four patterns resonated through all the J-PAL evaluations (and likely those of IPA and 3ie as well). First, they have focused on the local level. Some centered on villages, while others might take on larger jurisdictions up to a district , but few have been national or even state/provincial in scope. Micro-level rather than macro-level work has been the approach, which is not surprising, because while a medical RCT evaluation might survey thousands of subjects over an entire country (or even several countries), DG evaluations necessarily focus on specific projects covering fairly small areas, and of course a nationwide project would preclude using a control group.

Second, the evaluations have mostly addressed short-term interventions, often a couple of months on elections, and generally not more than a year or two for other experiments, though there have been some lasting several years. Third, J-PAL has most often used RCT surveys, but some evaluations have used other techniques like "difference in differences" analysis, community score cards, or "natural experiments" in which exogenous factors led to conditions establishing what amounted to randomized treatment and control groups.[5]

Fourth, although the J-PAL inventory includes a number of studies looking at civil society organizations broadly defined, all of them were concerned with service delivery. J-PAL has conducted no evaluations that I could find that focused on civil society advocacy. A search through the entire 3ie repository for "civil society" turned up only three reports, of which two evaluated community development councils charged with selecting/implementing/managing local projects and the other analyzed CSOs disseminating local government efficacy reports. In short, a quick computer search of 3ie's 2400+ evaluations found nothing on civil society advocacy. And nothing turned up among the Yale IPA studies or those cited by Moehler (2010).

---

[3] IPA's founder offers an account of its work in a recent book (Karlan and Appel 2011). Much of that work has been done in collaboration with J-PAL and so appears in the reports of both groups.

[4] To be included in 3ie's repository, an evaluation had to have been published in some fashion (articles, working papers, etc.), to have used an experimental estimation strategy, and to have occurred in a developing country. Many of the 196 had been conducted by J-PAL, but since many of the J-PAL studies have not (yet) been published, many did not make the 3ie listing. Like J-PAL and IPA, 3ie has also sponsored many evaluations of its own.

[5] "Difference in differences" analysis is similar to the RCT approach; see the note to Table 1. A good summary of the approach can be found in Wikepedia. Community score cards, also known as citizen report cards, is a technique pioneered by the Public Affairs Centre in Bangalore, India. For an example, see Paul (2006).

Among the 117 evaluations J-PAL classified under its Political Economy and Governance theme, some 15 seemed at first glance to possibly have a civil society component, as tallied in the rightmost column of Table 3. On closer examination, though, these 15 turned out to be at most concerned with service delivery, not advocacy.

J-PAL is currently undertaking a more concentrated exploration in the governance sector, with its Governance Initiative (GI), launched in 2011 and building on the principal-agent concepts set forth in the 2004 World Development Report (World Bank 2003). Building on an initial literature review (Olken and Rohini 2013), the program has commissioned a series of RCT evaluations, focusing on interventions designed to enhance voter control over political actors and to discourage corruption in the public sector. By the end of 2013, the GI had commissioned 19 evaluations, but none appear to have been completed as of this date. To judge from the capsule write-ups on the GI webpage, neither did any of them seem to focus on civil society advocacy (J-PAL 2014).

Using RCT approaches in civil society evaluations

RCT evaluations offer some real strengths in conducting DG evaluations. If properly done, they can show in measurable ways project impact (or its absence) exclusive of confounding exogenous variables. This is unquestionably a genuine advance over the traditional evaluation based on documents, key informants, site visits, and (more recently) focus groups.

The one RCT report that has thus far emerged from USAID's DRG Center illustrates the strength of the technique nicely in assessing an Agency-sponsored alternative dispute resolution project in Zimbabwe. Using two treatment groups of villages and a control group, the evaluation looked programs designed to enable traditional village leaders to better mitigate village disputes. The year-long project actually had a civil society (generously defined) component in that in one treatment group of community leaders (including women's CSO leaders) were invited to the training sessions in the hope that they would afterwards put pressure on the village leaders to use their training in their dispute resolution work. The other group received training only. The study found that some measureable behavioral change had occurred in the training-plus-pressure villages but not in the training-only villages as compared with the control group. In that first treatment group, however, the evaluators found some evidence of decreased social trust, which they explored in a qualitative study using focus groups (Baldwin and Muyengwa 2014).

The RCT technique also presents some significant weaknesses, as should be expected with any evaluation methodology:[6]

- RCT cannot be employed with macro-level programs, where an entire country is the treatment group. It is perforce a sub-national approach to evaluation.
- RCT requires a quantitatively measureable index of impact, a high bar in much DG work, particularly civil society advocacy.
- Projects evaluated are assumed to have proceeded as designed, i.e., without the serious mid-term changes of direction that so often occur in USAID-supported initiatives.
- "Selection bias" can affect the treatment group, making it different from the control group at the outset (though strict observation of randomization protocols should prevent this).

[6] Many of these points are taken from Kumar (2013: 96-102), which should be required reading for anyone engaged in RCT research.

- Spillover contamination can easily occur between the two groups, as individuals/associations/villages in the control group discover what assistance their counterparts in the treatment group are receiving.
- Exogenous factors can affect the control group (e.g., assistance from another donor) or both groups (e.g., natural disasters, new roads) in ways that overwhelm the treatment effects.
- The statistical tests (e.g., logistical regressions) employed in RCT evaluations can become too esoteric to be understood by anyone except those conducting the study.
- Opinion surveys tend to be expensive (especially if there are more than two of them), taking an overly large bite out of program budgets.
- While RCT studies can provide insight into *what* happened during a project, they shed little light on *why* it happened, which must be understood if the intervention is to be replicated on a wider scale. Qualitative analysis developed through more traditional means such as key informant interviews and focus groups are needed to understand the *why* factors.

Using RCT to evaluate civil society initiatives

Can RCT be used to evaluate civil society programs? One wonders why so few RCT studies thus far have looked at them, and those few focused on civil society's service delivery function rather than advocacy. Surely the answer in significant measure lies in the difficulties that would be encountered in assessing advocacy programs. Service delivery projects could be could be evaluated by gauging the impact of the services being provided at the end of the evaluation period (e.g., incidence of malaria after an NGO-managed eradication project, crime rates after a community association's effort to lower drug or alcohol abuse), or citizen opinion about either program,[7] but how to measure the impact of civil society advocacy?

Donor-sponsored civil society advocacy efforts generally can be thought of as endeavoring to influence public policy in one (or more) of three main areas:

- Enhancing human rights (e.g., gender, ethnic or racial or religious minorities, LGBT issues, disabled persons, HIV-AIDS victims).
- Promoting public goods (e.g. free speech, environment).
- Supporting weak economic actors (e.g., labor unions, small businesses, small farmers).

Assuming that the many constraints to RCT research outlined above can be overcome, it should be possible to get both objective and subjective data on a project's success in meeting these three public policy domains through opinion surveys or citizen report cards:

- Objective measures: Have public policy changes enabled women to enjoy more rights? Have toxic wastes in rivers diminished? Can small farmers now sell their produce at the weekly market?
- Subjective measures: Are minorities feeling less harassed by the police? Do citizens feel freer in expressing views? Do small business owners experience fewer bureaucratic obstacles in getting licenses?

The real problem will come with *attribution*: When a public policy change has occurred, how can we tell who or what brought about the change? Was it only the evaluated CSO's work? Other CSOs

[7] On citizen assessment of service delivery, again see Paul (2006).

supported by other donors? A growing economy in some treatment areas? Political officeholders independently changing their own agendas?

Here is where qualitative assessment work, as was carried out in the Zimbabwe evaluation noted above, can usefully supplement the quantitative RCT studies that USAID and other donors have begun to demand.

Table 1.

Randomized Control Trial template

	Treatment group (individuals, villages, etc.)	Control group (individuals, villages, etc.)
Baseline data (Time 1)	**A1** Survey, Community score card	**B1** Survey, Community score card
Post-intervention data (Time 2)	**A2** Survey, Community score card	**B2** Survey, Community score card
Difference between Time 1 data and Time 2 data	**A2-A1=A3**	**B2-B1=B3**

If samples are properly drawn, A1-B1 = 0
If A1-B1=0, the A3-B3 will show the effect of the intervention.
If A1-B1=some positive or negative number, this can be taken into account in the A3-B3 calculation, and the technique is known as "Difference in differences." Here, the final calculation would be (A2-B2)-(A1-B1).

http://www.povertyactionlab.org/search/apachesolr_search?view=grid&filters=type:

Table 2. J-PAL Evaluations, 2003-2014

NUMBER OF EVALUATIONS BY REGION AND THEME

Click on a number below to see all evaluations in that region and/or theme

		REGION					
		Africa	Europe	Latin America & the Caribbean	North America	South Asia	Southeast Asia
THEME		174	35	103	79	124	35
Education	138	26	7	39	23	33	3
Finance & Microfinance	178	62	4	48	12	30	16
Environment & Energy	28	5	0	6	6	10	1
Health	128	63	2	17	7	25	5
Political Economy & Governance	117	20	4	12	31	34	12
Labor Markets	64	12	20	16	8	4	2
Agriculture	56	34	0	4	0	16	2

NOTE: Many evaluations are counted under more than one theme. Thus many of the "Political Economy and Governance" evaluations are also included under the "Education" theme, "Health" theme, etc. The grand total of evaluations as of 25 October 2014 was 567.

Source: http://www.povertyactionlab.org/search/apachesolr_search?view=grid&filters=type:evaluation (accessed 25Oct14)

Table 3.
J-PAL evaluations in the Political Economy and Governance Sector

Evaluation topic	Evaluations completed	Evaluations ongoing	N.America & Europe	Africa, Asia, LAC	Total evaluations	Possible civil society content?
Elections	37	4	26	15	41	0
Service delivery	20	5	0	25	25	2
Community participation	9	2	1	10	11	8
Revenue generation	4	4	3	5	8	0
Local governance	4	2	0	6	6	4
Women leaders	4	1	0	5	5	0
Corruption	2	1	0	3	3	1
Fund raising	3	0	0	3	3	0
Other topics	10	5	5	10	15	0
Total	**93**	**24**	**35**	**82**	**117**	**15**

Source:
http://www.povertyactionlab.org/search/apachesolr_search?view=&filters=type%3Aevaluation%20sm_cck_field_themes%3A73&viewall=all (accessed 25 October 2014).

3ie impact evaluations search results for "governance" as of 21 Oct 2014

Evaluation topic	**Total**
Public health	56
Education	37
Poverty alleviation (incl. employment generation)	29
Agriculture	16
Women's issues	10
Elections	9
Local governance	7
Environment	5
Microcredit	3
Small & medium enterprises	3
Community-driven development	2
Post-conflict	2
Other (1 each)	17
Total	**196**

Source: www.3ieimpact.org/en/evidence/impact-evaluations/impact-evaluation-repository/?q=governance&title=&author=&published_from=&published_to=&publication_status=All+completed+Impact+Evaluations (accessed 24 October 2014).

References

Baldwin, Kate, and Shylock Muyengwa. 2014. "Impact evaluation of supporting traditional leaders and local structures to mitigate community-level conflict in Zimbabwe." Final Report (Arlington, VA: Social Impact for USAID).

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*.

J-PAL. 2014. J-PAL Governance Initiative, web page, accessed 25 October 2014 at <http://www.povertyactionlab.org/GI>.

Karlan, Dean, and Jacob Appel. 2011. *More Than Good Intentions: How a New Economics Is Helping to Solve Global Poverty* (New York: Dutton).

Kumar, Krishna. 2013. *Evaluating Democracy Assistance* (Boulder, CO: Lynne Rienner).

Levitt, Steven D., and John A. List. 2009. "Field experiments in economics: The past, the present and the future," *European Economic Review* 53, 1-18.

Moehler, Devfa C. 2010. "Democracy, governance and randomized development assistance," *Annals of the American Academy of Political and Social Science* 628, 1 (March), 30-46.

NRC (National Research Council of the National Academies). 2008. *Improving Democracy Assistance: Building Knowledge Through Evaluations and Research*. (Washington: National Academies Press).

Olken, Benjamin A., and Rohini Pande. 2013. "Governance review paper: J-PAL Governance Initiative" (J-PAL, October).

Paul, Samuel. 2006. "Public spending, outcomes and accountability: Citizen report card as a catalyst for public action," *Economic and Political Weekly* 41, 4 (28 January), 333-338.

World Bank. 2003. *World Development Report 2004: Making Services Work for Poor People* (Washington: World Bank).