**RANDOMIZED CONTROL TRAILS**

**References**

[DGK] Duflo, Esther, Glennerster, Rachel and Kremer, Michael. 2006. Using Randomization in Development Economics Research: A tool kit. Accessible at <http://economics.mit.edu/files/806>

Bloom, H. S. (1995): Minimum detectable effects: A simple way to report the statistical power of experimental designs." Evaluation Review, 19, 547-56.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): “Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," American Economic Review, 92(5), 1535-1558.

Banerjee, A., E. Duflo, S. Cole, and L. Linden (2007): “Remedying Education: Evidence from Two Randomized Experiments in India," Quarterly Journal of Economics, [Volume 122, Issue 3](http://qje.oxfordjournals.org/content/122/3.toc) , Pp. 1235-1264

Ashraf, N., D. S. Karlan, and W. Yin (2006): “Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," Quarterly Journal of Economics, [Volume 121, Issue 2](http://qje.oxfordjournals.org/content/121/2.toc), Pp. 635-672.

Karlan, and J. Zinman (2006): “Expanding Credit Access: Using Randomized Supply Decisions To Estimate the Impacts," mimeo, Yale University.

Miguel, E., and M. Kremer (2003) “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," Econometrica, 72(1), 159-218.

Holland, P. W. (1988): “Causal Inference, Path Analysis, and Recursive Structural Equations Models," Sociological Methodology, 18, 449-484.

1. **Why Randomized Control Trials (RCTs) ?**

Let’s say is health status, and is binary variable of “going to hospital or not”. To obtain the effect of on let’s define potential outcomes. denote the health status of an individual who did not go to hospital, irrespective whether he/she actually did. Similarly, denote the health status of an individual who went to hospital, irrespective whether he/she did.

---------------(1)

Causal effect for each person i is simply:

-----------------------------------(2)

It is clear from this expression that, to find the effect of going to hospital on health status, we need to answer counterfactual questions. What will be the health status of individuals who went to the hospital, had they not gone to the hospital? Or more generally, how would individuals who participated in a program have fared in the absence of the program? The difficulty with these questions is immediate. At a given point in time, an individual is either exposed to the program or not. We actually do not observe the counterfactual. In the parlance of equation (1), we do not know if D=1.

*How do we deal with this absence of real world counterfactuals?*

Rather than induvial causal effect like in equation 2, we can obtain the *average* impact of D on a group of individuals by comparing them to a similar group of individuals who were not exposed to the program.

This will simply be:

Does the simple average difference yield causal effect? No, because it has two components:

1. First term is the treatment effect on the treated.
2. Second term is selection bias. It shows how if not treated, the outcome of those who end up getting treatment and those who do not are different. Compare this to: how health status if not going to the hospital is different between those who *actually* go to hospital and those *actually* who do not. In other words, the control group is usually not comparable to the treatment group in the absence of treatment. For example,

* Hospital goers are different from non-goers.
* Families choose whether to send girls to school.
* Households participating in the MNREGS program are different from those who do not.

This selection bias prevents us from taking simple differences of outcome between treatment and control groups, to obtain causal effects. How do we eliminate the selection bias?

*Again, how do we deal with this selection bias?*

Make independent. That is, make “going to hospital” independent of someone’s health status. This is exactly what we call “randomizing” a treatment. Randomization solves the selection bias. How? From the above equation,

β = E

The third equality is because are independent. The final expression tells us that if indeed are independent, or in other words, if we randomize treatment, we can then take simple differences of outcomes between treatment and control status, and obtain causal effects without any selection bias.

The regression counterpart to obtain causal effect after a randomization β is as follows:

D is a dummy for assignment to the treatment group. Equation (2) can be estimated with Ordinary Least Squares (OLS), and we already know that. Note that the estimated is only the average treatment effect (ATT), and this estimation does not estimate individual specific treatment effect or assume a constant treatment effect across individuals.

1. **How to implement an RCT?**

*Pilot projects*

A natural window to introduce randomization is before the program is scaled up, during the pilot phase.

*Implementation Partners*

* Introducing randomization in real-world programs almost requires working with a partner who is in charge of actually implementing the program.
* Governments are possible partners. Governments usually run pilot programs before a public program is scaled up, researchers can collaborate with governments.
* Example: *Progresa*[[1]](#footnote-1). The program offers grants, distributed to women, conditional on children's school attendance and preventative health measures (nutrition supplementation, health care visits, and participation in health education programs).
* In 1998, since budgetary constraints made it impossible to reach the 50,000 potential beneficiary communities of PROGRESA all at once, they started with a randomized pilot program in 506 communities.
* Half of those were randomly selected to receive the program, and baseline and subsequent data were collected in the remaining communities. [[2]](#footnote-2)
* With the success of this, many other Latin American countries implemented programs like the PROGRESA[[3]](#footnote-3), which had a pilot component that was randomized. Since then, many government sponsored pilot programs have been implemented. In many cases governments work with researchers in designing the RCTs.
* *But governments are hard to partner* because many of these decisions are political.

Other alternatives:

* NGOs can be good alternatives in many cases. They can be open to innovative questions which may or may not be immediately politically feasible, allowing for greater input from researchers into the design of programs, especially in the pilot stage
* For-profit firms, banks, micro-finance organizations

*RCTs versus field experiments*

Early studies simply compared treatment group to control groups to understand if a program worked or not. For example, the *Progresa* pilot program implemented the program in the treatment villages and did not introduce it in the comparison villages. The evaluation is only able to say whether, taken together, all the components of PROGRESA are effective in increasing health and education outcomes. The shortcoming of this is that we cannot disentangle the various mechanisms at play without further assumptions.

The new trend is the so called “field experiments”, which test various mechanisms at play beyond simply testing if the program works or not. They have one control group, and different variants of program as different treatment arms. Example: Ashraf, Karlan, and Yin (2006) designed a commitment savings product for a small rural bank in the Philippines. The rural bank was interested in participating in a program that had the potential to increase savings. Individuals could restrict the access to the funds they deposited in the accounts until either a given maturity date or a given amount of money was saved. Relative to standard accounts, the accounts carried no advantage other than this feature. The product was offered to a randomly selected half of 1,700 former clients of the bank. The other half of the individuals were assigned either to a pure comparison group or to a group who was visited and given a speech reminding them of the importance of savings. This group allowed them to compare two mechanisms; whether the simple fact of discussing savings is what encourages clients to save, rather than the availability of a time-commitment device.

*How to randomize?*

1. Randomizer or Lottery

Karlan and Zinman (2010) evaluated the impact of expanded consumer credit in South Africa by working with a lender who randomly approved some marginal loan applications that would normally have been rejected. All applicants who would normally have been approved received loans, and those who were well below the cut-off were rejected.

Angrist, Bettinger, and Kremer(2006) to evaluate the long term impact of a Colombian voucher program on latent learning. Secondary school vouchers were allocated by lottery among a set of applicants.

1. Phase-in

Provide treatment to groups in a random order staggered over time. The Primary School Deworming Project provides an example of this type of randomized phase-in trial (Miguel and Kremer 2004). This program provided medical treatment for intestinal worms (helminths) and schistosomiasis as well as worm-prevention health education lessons to children in 75 primary schools in rural Busia district, Kenya during 1998-2002. The program randomly divided the schools into three groups, each consisting of 25 primary schools. Treatment in the schools was done as follows: 25 Group 1 schools began receiving treatment in 1998; 25 Group 2 schools began receiving treatment in 1999; and 25 Group 3 schools began receiving treatment in 2000. The impact of the program on the health, nutrition, and education of the children was evaluated by comparing the results from group 1 schools in 1998 with group 2 and 3 acting as comparisons and the results of group 1 and 2 schools in 1999 with group 3 schools acting as a comparison. The researchers found that deworming led to improved health and increased school participation.

Potential pitfalls:

* If a randomized phase-in is too rapid relative to the time it takes for program effects to materialize, it will be impossible to detect treatment effects at all. For example, one would be unlikely to detect the effect of a microcredit program that was phased-in to control villages only six months after it was introduced to the treatment group. When planning a phase-in design, the time between phases should be sufficient to encompass any treatment lag.
* Randomization becomes problematic when the comparison group is affected by the expectation of future treatment.

1. *Within group randomization*

Even a randomized phase-in may not spread the benefits sufficiently smoothly across the whole group to ensure good cooperation with the study. For example, schools that do not receive a treatment may refuse to let researchers collect test scores on their students. In this case, it is still possible to introduce an element of randomization by providing the program to some subgroups.

The evaluation of the balsakhi program, a remedial education assistance in poor urban schools in India provided by Pratham, an Indian education NGO (Banerjee, Duflo, Cole, and Linden 2007) provides an example. The program was designed to provide those children falling behind in school the basic skills they need to learn effectively. Pratham hires and trains tutors, referred to as balsakhi or “child's friend," to give remedial math and reading comprehension instruction to children. To ensure cooperation from school authorities, every school in the study received a balsakhi in every year. However, to ensure full cooperation, schools were divided into two groups (A and B). Schools in group A, where the balsakhi was assigned in grade three in the year 2001-2002 were now assigned a balsakhi in grade four. Schools in group B, where the balsakhi was assigned to grade four in year 1 received balsakhi assistance for grade three in year 2. This design was deemed fair by school teachers, since all schools received the same assistance. Furthermore, since the NGO could make a credible case that they could not provide more than one balsakhi per school, there was no expectation that all children in a school should benefit from the program.

Given this design, in each year, children in grade three in schools that received the program for grade four form the comparison group for children that receive the program for grade three, and vice versa.

Problems:

The drawback of such designs is that they increase the likelihood that the comparison group is contaminated. For example, in the balsakhi program, one may have been worried that head-masters reallocated resources from grade 3 to grade 4 if grade 3 got a balsakhi but grade 4 did not. In this particular application, such contamination was unlikely because schools have a fixed number of teachers per grade and few other resources to reallocate. But this risk needs to be considered when deciding whether or not to adopt such a design.

1. *Encouragement design*

Encouragement designs allow researchers to evaluate the impact of a program that is available in the entire study area but whose take up is not universal. They are particularly useful for evaluating programs over which randomization of access is not feasible for ethical or practical reasons. Rather than randomize over the treatment itself, researchers randomly assign subjects an encouragement to receive the treatment.

One of the early encouragement designs was a study of whether studying for the GRE could lead to an increase in test scores (Holland 1988). While studying is available to everyone, researchers increased the number of students who studied for it by mailing out free materials to a randomly selected set of GRE candidates. Another encouragement design is providing farmers with assistance or information on best farming practices.

1. **Sample size, design, and the power of experiments**

Suppose we provide treatment T (=1) randomly to a sample of participants. Let P be the proportion of sample of size N, which is treated. The following regression is run to estimate the impact of treatment. is the average treatment effect.

The difference in sample means for two groups (treatment and control) is the OLS coefficient of D in the regression. We are generally interested in testing the hypothesis H0, that the effect of the treatment is equal to zero against the alternative that it is not.

H0:

HA:

When we say , we must also have a sense of what “should be”. This is called expected size effect. How do we determine the size effect? Various ways: Context, past studies, experience, cost-benefit analysis. Suppose an experiment intends to provide relevant information to treatment farmers on best agricultural practices. The control framers do not get any information. If as a result of this simple information provision, the policy makers at least expect a minimum 10% increase in yield. Then 10% is your expected effect size.

But can 10% be used in other contexts? It could be much easier to increase attendance of students as a result of provision of mid-day meals, rather than increase their test score as a result. Equally if not harder will be increasing child weight as a result. So if you want to provide mid-day meals to randomly chosen schools, expected effect sizes differs across each of these outcomes. The effect size also obviously depends on the cost of the treatment. If we undertake an expensive treatment (providing cash transfers to households), then we expect a size effect that is “large enough” so that the benefits exceed the cost of the treatment provision. But if the treatment is inexpensive (providing daily price information to farmers), then the effect size expected to cover the cost of the program would also be small.

Note that the above is simply based on literature and cost/benefit. It had nothing to do with the design of our experiment. But given that we want to be able to get this effect with statistical significance, our objective as researchers is to design an experiment which will give us statistically significant . There is a parameter called the Minimum Detectable Effect (MDE) size, which determines the minimum that can be detected statistically different from zero, as a result of treatment, for a given sample size (N), level of significance (α), and power of the study (τ). Before we formally define what MDE is, let’s first see how to interpret it. Suppose the expected effect size of information provision to farmers (based on the cost and other characteristics of the intervention) is a 10% increase in yield of paddy, which is 2.6 quintals/acre. Suppose based on the proposed experiment’s N, α, and τ, the MDE is 15%, or 3.9 quintals/acre. But suppose the “true” effect is 3.2 quintal/acre, something above what we want our to be equal to. This 3.2 will not be statistically different from zero. Only above the MDE of 3.9 will be statistically significant based on our design.

How do we improve our design such that 3.2 is statistically significant? For that we need to understand hypothesis testing deeper.



In the figure above, the left bell-shaped curve is the true distribution of when the null hypothesis is true (H0: ). Now since our expected effect size is 2.6. We want to be able to significantly obtain that as . Let us then set HA:.

Usually, we set significant level (α) for a hypothesis test. The significance level represents the probability of type I error, i.e., the probability we reject the hypothesis when it is in fact true. Usually, we set α=.05. For a one-sided test, as in the figure we reject the null hypothesis if > , or >

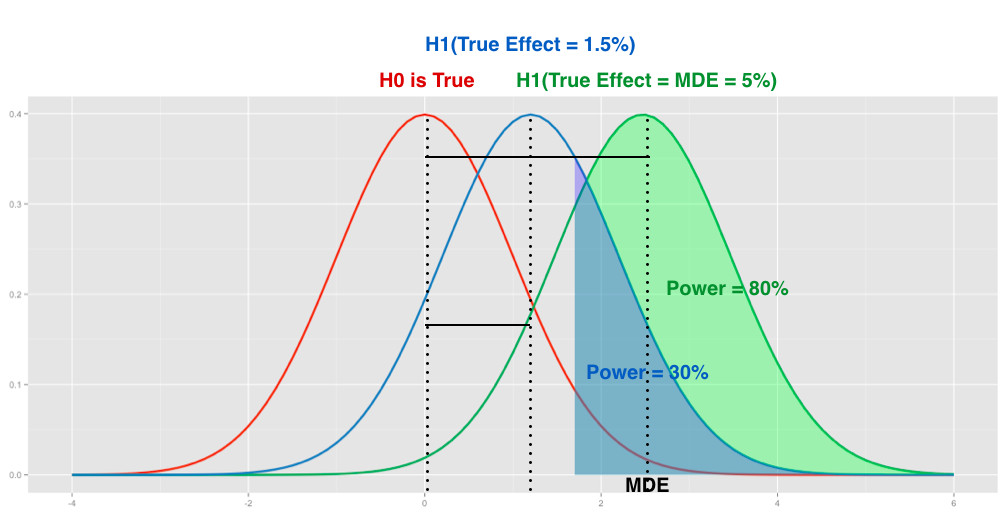
The right bell-shaped curve is the true distribution of when the alternate hypothesis is true (HA: . What is type-II error? Probability of accepting a null hypothesis that is false. 1 – type-II error = power of the test, κ, which is the probability of rejecting the null hypothesis when it is in fact false. Power κ is the area to the right of the critical value in the right bell-shaped curve. From the figure above, note that to achieve a power κ, then,

>

The minimum detectable effect size for a given power (κ), significance level (α), sample size (N), and portion of subjects allocated to treatment group (P) is therefore given by:

or a single sided test (tα is replaced by tα/2 for a two-sided test). Alternatively, the above equation implicitly denotes the sample size N required to achieve a given power, given the effect size that is posited and the level of significance chosen. There is a trade-off between power and sample size.

Graphically:



*Intuition*

Larger the sample size, lower the MDE for the same power and significance, and vice versa.

A larger MDE size gives higher power

A larger sample size and/or small variance gives higher power

Coming back to our paddy example. Suppose MDE was 3.0. Now how can we increase the power of detecting an effect size of .?

1. We can increase the sample size
2. Decrease by adding more covariates

*Group randomization*

Many of the designs we discussed above (Balsakhi program, *progresa* program) involve randomizing over groups rather than individuals. In such cases, researchers nevertheless often have access to individual data. For example, in the *progresa* program, the village was the unit of randomization, but individual data were available. Then we would run the following regression to obtain treatment effects:

Suppose there are J clusters (groups) of identical size n, is i.i.d with variance , is i.i.d with variance . The OLS estimator is still unbiased, and its standard error is:

1. **Covariates and stratification**

*Controlling for covariates*

In a simple randomized experiment, controlling for baseline values of covariates likely to influence or predict the outcome does not affect the expected value of an estimator[[4]](#footnote-4), but it can reduce its variance. But note also that if we control for covariates affected by the treatment itself that would bias the estimate of the treatment effect by capturing part of its impact. Information on covariates should therefore be collected in the baseline surveys. A special case of a covariate of interest is the pre-treatment value of the outcome.

This is a reason why baseline surveys can greatly reduce sample size requirement when the outcome variables are persistent. For example, controlling for baseline test scores in evaluations of education interventions greatly improves the precision of estimates, which reduces the cost of these evaluations when a baseline test can be conducted. Note, however, that controlling for variables that explain little or none of the variation in the outcome will increase standard errors by reducing degrees of freedom.

Stratification

The randomization ensures that treatment and control groups will be similar in expectation. But what if you an RCT only to find that control and treatment groups were not actually randomly allocated. How can you prevent this issue? Stratification is the answer. In stratification, we pick groups and randomize control and treatment within groups, so that we ensure balance on that group.

For example, in the balsakhi program, researchers stratified according to class size, language of instruction, and school gender (boys, girls, or coed) as well as according to pre-test scores for schools in the Mumbai area. A block is constituted of all the schools that share the same language of instruction, the same school gender, and fall in the same “bin" of pre-test scores. An extreme version of this design is the pairwise matched design where pairs of units are constituted, and in each pair, one unit is randomly assigned to the treatment and one unit is randomly assigned to the control. In the final estimation equation, if we add block dummies for each strata, the standard error is much lower than when we estimate the same stratified sample but without block dummies (refer to DGK, p38).

1. **Randomization can avoid publication bias**

*Publication and citation bias*

* While using non-experimental methods, we may tend to mine various regression specifications for one that produces statistically significant results.
* Even without mining: if we get expected results, we may just stop there without exploring other specifications.
* Even if a researcher introduces no bias, the selection by journals of papers with significant results introduces another level of publication bias.
* Researchers cite papers with extreme results by advocates on one side or another of policy debates is likely to compound publication bias with citation bias.
* Result: If we find null effects, no one wants to publish it. Only strongly positive and/or strongly negative estimates are likely to be published and widely cited. [[5]](#footnote-5)

*How can RCTs help?*

1. Little ex-post discretion on modelling

If a randomized evaluation is correctly implemented, there can be no question that the results, whatever they are, give us the impact of the particular intervention that was tested (subject to a known degree of sampling error). If results are unexpected (contrary to intuition or theory), they have less chance of being considered the result of specification error and discarded.

Why?

* Treatment and comparison groups are determined before a researcher knows how these choices will affect the results, giving limiting room for ex post discretion.
* There is usually still some flexibility ex-post including what variables to control for, how to handle subgroups, and how to deal with large numbers of possible outcome variables which may lead to “cherry-picking" amongst the results.
* However, unlike omitted variable bias which can become arbitrarily large, this ex post discretion is bounded by the ex-ante design choices.

Consider Miguel and Kremer (2003) evaluation of the impact of network and peer effects on the take up of deworming medicine. Ex ante, the researchers probably expected children with more links to students in treatment schools to increase their uptake of deworming treatment in subsequent rounds of the program as they learned about the medicine's benefits. Instead, their experimental findings showed a significant effect in the opposite direction. Had they obtained these results in a retrospective study, most researchers would have assumed there was a problem with their data or specification and explored alternative specifications. The experimental design leaves little doubt that peer effects are in fact negative.

1. Part of the publication bias can be overcome

Results are usually documented even if they suggest insignificant effects. Even when unpublished, they are typically circulated and often discussed in systematic reviews. In addition, funding agencies typically require a report of how their money was spent regardless of outcomes.

1. **Practical Design and Implementation Issue**
2. Level of randomization:

An important practical design choice is whether to randomize the intervention at the level of the individual, the family, the village, the district, etc.

1. Early randomization of deworming medicines were carried out at the individual level within schools (Dickson and Garner 2000), while Miguel and Kremer (2004) look at similar programs by randomly phasing in the program at the school level.
2. Interventions such as input provisions in classrooms could be carried out at the school level (for example, schools get selected to receive textbooks)
3. The level of individuals students (in the Tennessee Star experiment, students within schools were randomly assigned to either a large class, a small class, or a class with a teacher aid (Krueger and Whitmore 2002)

When there is some flexibility in the level of randomization, several factors need to be considered:

1. The larger the groups that are randomized, the larger the total sample size needed to achieve a given power.
2. Second, however, spill overs from treatment to comparison groups can bias the estimation of treatment effects
3. Third, randomization at the group level may sometimes be much easier from the implementation point of view, even if it requires larger sample sizes
4. Individual-level randomization of a program perceived as desirable in a village or a neighbourhood may create resentment towards the implementation organization.
5. Cross cutting designs

In cross-cutting designs several different treatments are tested simultaneously with randomizations being conducted so that treatments are orthogonal to each other. From cross cutting designs, we can check which combination of treatments work.

Progresa is a combination of: A cash transfer, a redistribution of resources towards women, and an incentive component. The Mexican government may be interested in checking if the entire program works or not in improving education and health. But researchers may want to see what among these three components does that job, if at all. May be those components can scaled even bigger.

More generally, if a researcher is cross-cutting interventions A and B, each of which has a comparison group, she obtains four groups: no interventions (pure control); A only; B only; and A and B together (full intervention). These can be costly because the sample sizes must be sufficiently large in each of these four groups.

In the initial study, intermediate variables likely to be affected by one intervention but not the other can be used to shed light on which part of the intervention was effective. For example, in the deworming pilot mentioned above (Miguel and Kremer 2004), two programs were combined: deworming pills were distributed, and children were given advice about preventive behaviour (wearing shoes, washing hands, etc.). Researchers collected variables on behaviour which suggested that no behaviour changed in the treatment schools. This strongly suggests that the component of the intervention that made the difference was the provision of the deworming pill.

Another advantage: without much additional cost, you can test many hypotheses. For example, Banerjee, Duflo, Cole, and Linden (2007) tested in the same sample (the municipal schools in Vadodara, India) the effect of remedial education and the effects of Computer Assisted Learning.

1. Conducting baseline surveys:

* A baseline survey generates control variables that will reduce the variability in final outcomes and therefore reduces sample size requirements. When the intervention is expensive and data collection is relatively cheap, conducting a baseline will save money. When the intervention is cheap but data collection is expensive, it may be more cost effective to run a larger experiment without conducting a baseline.
* They make it possible to examine interactions between initial conditions and the impact of the program.
* A baseline survey provides an opportunity to check that the randomization was conducted appropriately.
* Collecting baseline data offers an opportunity to test and refine data collection procedures

1. Attrition

* Attrition refers to the failure to collect outcome data from some individuals who were part of the original sample.
* Random attrition will only reduce a study's statistical power
* Attrition that is correlated with the treatment being evaluated may bias estimates.
* While randomization ensures independence of potential outcomes in the initial treatment and comparison groups, it does not hold after non-random attrition.
* A first step in the analysis of an evaluation must always be to report attrition levels in the treatment and comparison groups and to compare attritors with non-attritors using baseline data (when available) to see if they differ systematically, at least along observable dimensions.
* If attrition remains a problem, statistical techniques are available to identify and adjust for the bias. These techniques can be parametric (see Hausman and Wise (1979), Wooldridge (2002) or Grasdal (2001)) or nonparametric.

1. Hawthorne and John Henry effects

Another limitation of prospective evaluations is that the evaluation itself may cause the treatment or comparison group to change its behaviour.

* Changes in behaviour among the comparison group are called John Henry effects. The comparison group may feel offended to be a comparison group and react by also altering their behaviour (for example, teachers in the comparison group for an evaluation may “compete" with the treatment teachers or, on the contrary, decide to slack off
* Changes in behaviour among the treatment group are called Hawthorne effects. The treatment group may be grateful to receive a treatment and conscious of being observed, which may induce them to alter their behaviour for the duration of the experiment (for example, working harder to make it a success). Behaviour can change for those who receive program benefits just because of high morale. For example, provision of school inputs could temporarily increase morale among students and teachers, which could improve performance in the short run. Such effects would create problems for fixed-effects, difference-in-differences and regression discontinuity estimates as well as randomized evaluations. What makes an experiment different or special is that individuals may know they are part of an evaluation and may thus react to the very fact of being evaluated, not only to the inputs received.

*How to check if there are no effects such as these?*

One way to disentangle Hawthorne or John Henry effects from long run impacts of the program which would obtain outside of an evaluation is to collect longer run data. For example, Duflo and Hanna (2006) initiated a financial incentive program to reduce teacher absenteeism. Teachers were given a camera with a tamper-proof date and time function, along with instructions to have one of the children photograph the teacher and other students at the beginning and end of the school day. The time and date stamps on the photographs were used to track teacher attendance. A teacher's salary was a direct function of his attendance. This dramatically decreased teacher absenteeism, but was it due to Hawthorne effect?

To check this, the authors continued to monitor the impact of the camera program over a year after the official “experiment" was over (but the NGO decided to continue to implement the program as a permanent program). The fact that the results are similar when the program is not being officially evaluated anymore and at the beginning of the evaluation period suggest that the initial results on presence where not due to Hawthorne effects.

**Some questions to ponder**

More generally, an issue that often comes up with randomized evaluations is the extent to which the results are replicable or generalizable to other contexts. Sure, a specific program worked in one community in Western Kenya, but can we extrapolate that it will work elsewhere? Was its success linked to a specific NGO? Would a similar program, but with minor variations, have the same impact?

1. *Progresa* is now called *Oportunidades*, and also discussed in Todd, 2006 and Parker, Rubalcava, and Teruel, 2005. [↑](#footnote-ref-1)
2. The task of evaluating the program was given to academic researchers through the International Food Policy Research Institute. The evaluations showed that it was effective in improving health and education. [↑](#footnote-ref-2)
3. Family Allowance Program (PRAF) in Honduras, a conditional cash transfer program in Ecuador (Schady and Araujo 2006). [↑](#footnote-ref-3)
4. It does not affect the expected value of the estimator by the very definition of randomization. If you add more covariates that are not correlated to the existing ones, it does not affect the estimate of the existing coefficients (refer to lecture notes 1 and 2) [↑](#footnote-ref-4)
5. There is formal evidence on such bias in Economics. Hedges (1992) proposed a formal model of publication bias where tests that yield lower p-values are more likely to be observed. Ashenfelter, Harmon, and Oosterbeek (1999) apply Hedges's analytic framework to the literature on rates of return to education and find strong evidence that publication bias exists for instrumental variables (IV) estimates on the rate of return to education, which suggests that the often cited results that IV estimates of returns to education are larger than OLS estimates may just be an artefact of publication bias. Likewise, Card and Krueger (1995) find evidence of significant publication bias in the time-series minimum wage literature, leading to an over-reporting of significant results. [↑](#footnote-ref-5)