J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

GUIDE 4: DEWORMING IN KENYA

Addressing threats to experimental integrity

This case study is based on Edward Miguel and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," Econometrica 72(1): 159-217, 2004.

J-PAL thanks the authors for allowing us to use their paper.

LEARNING OBJECTIVE

To explore how common threats to experimental integrity can be managed

SUBJECTS COVERED

Threats to experimental integrity, equivalence, comparability, compliance, spillovers or externalities, behavioral responses, intention to treat, treatment on the treated

THREATS TO EXPERIMENTAL INTEGRITY

There are three main types of threats. This case covers types 1 and 2, but we include 3 in case you are asked about it.

1. When the groups do not remain equivalent—attrition bias

Estimates may become biased if people select themselves in or out of either of the groups— join or drop out— over the course of the experiment, *and* their reasons for doing so are systematically related to the treatment. While this can be seen as a program effect, it makes it more difficult to interpret any differences in outcomes. In a sense treatment correlated attrition reintroduces selection bias. The experimental groups comprise different people at the end; they are no longer equivalent and the planned comparison may no longer be valid.

2. When the planned experimental contrast is diminished—partial compliance, alternative services, and spillovers

The planned difference in treatment rates between the groups can disappear if people assigned to the treatment group are not actually treated or if people assigned to the comparison group do in fact get treated, directly or indirectly.

Some people assigned to the treatment may in the end not get treated. For example, children assigned to an after-school tutoring program may simply not show up for tutoring. This is called partial compliance.

Some people assigned to the comparison may access program services or else get equivalent services from another provider. For example, children assigned to the after-school tutoring comparison group may get extra help from the teachers or get program materials and methods from their classmates. If this happens systematically, the treatment contrast between the groups begins to disappear and the impact comparison begins to become invalid. This is sometimes called contamination or, more benignly, diffusion of treatment to control.

Then people assigned to comparison may benefit indirectly from the treatment group getting treated. So, for example, a program that distributes insecticide treated nets may reduce malaria transmission in the community, indirectly benefiting those who themselves do not sleep under a net. Such effects are called externalities or spillovers.

3. When there are behavioral responses to the evaluation, not the treatment itself, responses that would not exist in the absence of the evaluation

When a program is being evaluated, participants may change their behavior because they are under observation; that is, they may respond to the program in ways they wouldn't if the program was not being evaluated. In such cases the impact estimates may capture not only the effects of the treatment but also the effects of the evaluation of the treatment.

People assigned to the comparison may start to compete with people in the treatment group. So, for example, in a program using contract teachers (treatment), the regular teachers (comparison) may work extra hard, harder than normal, during the course of the experiment so as not to be outdone by the contract teachers. And once the experiment is over, they may revert to their normal level of effort. Competition makes the outcomes of the comparison higher than normal, biasing any impact downwards. These effects are sometimes called John Henry Effects.[1]

People assigned to the treatment group may also change their behavior. For example, they may react positively to the novelty of the treatment. So when a school receives new inputs, morale goes up and students and the teachers temporarily perform better. Then the novelty wears off and performance drops. Or else the innovation is disruptive. Students and teachers struggle with a new way of learning

1 These effects are called John Henry effects after an American steel driver of the late nineteenth century, who worked in Virginia laying railway track with hammers. When steam drills were introduced, threatening to make steel drivers redundant, John Henry is said to have challenged the steam engine to a drilling competition, telling his captain "A man ain't nothing but a man. Before I am bitten by that steam drill, I will die with this hammer in my hand." He won the competition, but died "with the hammer in his hand" from overexertion. His story survives in American folk music.

and teaching and temporarily perform worse. Either way, if the evaluation period coincides with the adjustment period, impact estimates would also capture the effects of the novelty or disruption. Such effects are sometimes called Hawthorne Effects.[2]

THREATS TO INTEGRITY OF THE PLANNED EXPERIMENT

Discussion Topic 1

Threats to experimental integrity

(15 minutes)

1. What does it mean to say that the groups are equivalent at the start of the program?

   It means they are composed of individuals that on average have comparable characteristics.

2. Can you check if the groups are equivalent at the beginning of the program? How?

   Yes, compare the means of the groups on the characteristics that are important. Same as checking if "randomization was successful"

MANAGING ATTRITION: WHEN THE GROUPS DO NOT REMAIN EQUIVALENT

Discussion Topic 2

Managing Attrition (25 minutes)

1.

   a. At pretest, what is the average worm load of each group?

      T=2, C=2

   b. At posttest, what is the average worm load of each group?

      T=1, C=2

   c. What is the impact of the program

      Impact = -1

   d. Do you need to know preset values? Why or why not?

      No, because randomization ensures that the two groups are equivalent at the beginning of the program in expectation.

2.

   a. At posttest, what is the new average worm load for the comparison group?

      C=1.5

   b. What is the difference?

      Difference = -0.5

   c. Is this outcome difference an accurate estimate of impact of the program? Why or why not?

      NO, it is not an accurate estimate because it omits the drop-outs. The children who dropped out were worse off than the average child, and this reason for dropping out (because they had high worm loads) is correlated with the treatment.

   d. If it is not an accurate, does it overestimate or underestimate the impact?

      Underestimates by 0.5

   e. How can we get a better estimate of the program's impact?

      Follow up the whole lot of them (Intention to treat—take the average on the same people at the beginning and at the end—compare the averages of treatment and comparison based on the original assignments).

3.

   a. Would differential attrition (i.e. difference in drop-outs between treatment and comparison groups) bias either of these outcomes? How?

      Yes. Treatment can affect attendance and, through that, the test scores. Symptoms (listlessness, etc.) may also affect ability to concentrate in class and, through that, the test

2 In a study carried out at Western Electric Company's Hawthorne, USA, site in the 1930s, it was thought that workers responded to being under observation by increasing productivity. This interpretation has since been challenged but the name survives.

scores. And children who fall behind may also tend to drop out.

b. Would the impacts on these final outcome measures be underestimated or overestimated?

They would be underestimated for the same reasons since attendance and test score averages in the treatment would be based on the outcomes of the children in schools, which are better, and so attendance and test scores would be higher than they would be if all children were tracked.

4.

a. Does the threat of attrition only present itself in randomized evaluations?

No. The threats are general to all methods of estimating impact.

MANAGING PARTIAL COMPLIANCE:WHEN THE TREATMENT DOES NOT ACTUALLY GET TREATED OR THE COMPARISON GROUP GETS TREATED

Discussion Topic 3

Managing partial compliance

(25 minutes)

1.

a. Calculate the impact estimate based on the original group assignments

Impact = (25/15 -2) = -1/3 = -0.333

b. This is an unbiased measure of the effect of the program, but in what ways is it useful and in what ways is it not as useful?

This estimate provides a measure of the effect of the program as a whole, not accounting for the fact that not everyone complied with the planned intervention protocol. This is referred to as the "intention to treat" (ITT) estimate.

Ultimately, it depends what you want to learn about. ITT may relate more to how programs are actually implemented on the ground. For example, we may not be interested in the medical effect of deworming treatment, but what would happen under an actual deworming program. If students often miss school and therefore don't get the deworming medicine, the ITT estimate may actually be most useful.

To learn the impact of the treatment on those that actually receive the pill, you would need the "treatment on the treated" (TOT) estimate.

c. Five of your colleagues are passing by your desk; they all agree that you should calculate the effect of the treatment using only the 10,000 children who were treated and compare them to the comparison group. Is this advice sound? Is this advice sound? Why or why not?

This advice is not sound. The question that must be asked is, how are the children you exclude different from the average child? In this case they have above average worm loads and excluding them introduces attrition and selection bias, thereby producing non-equivalence between the treatment and comparison groups.

d. Another colleague says that it is not a good idea to drop the untreated entirely; you should use them but consider them as part of the comparison. Is this advice sound? Why or why not?

This advice is also not sound. It does not stick to the original assignments; the suggested manipulation reintroduces selection bias, by re-categorizing the high worm-load children from the treatment group into the comparison group. Again, this produces non-equivalence between the two groups.

MANAGING SPILLOVERS: WHEN THE COMPARISON, ITSELF UNTREATED, BENEFITS FROM THE TREATMENT BEING TREATED

Discussion Topic 4

Managing spillovers (25 minutes)

1.

a. If there are any spillovers, where would you expect them to show up?

In this example, spillovers would show up within the schools: girls 13 or older who are not treated but are near students who are treated would benefit.

b. Is it possible for you to capture these potential spillover effects? How?

YES, we can compare non-treated girls in treatment schools (in other words, the girls 13 and older) with non-treated girl in comparison schools.

2.

a. What is the treatment effect for boys in treatment v. comparison schools?

(10,000)(1)/10,000 = 1

((5000)(3) + (5000)(2))/10,000 = 2.5

1 – 2.5 = -1.5

b. What is the treatment effect for girls under 13 in treatment v. comparison schools?

(5000)(1)/5000 = 1

((3000)(2) + (2000)(3))/5000 = 2.4

1 – 2.4 = -1.4

c. What is the direct treatment effect among those who were treated?

-1.5(2/3) + -1.4(1/3) = -1.47

d. What is the treatment effect for girls 13 and older in treatment v. comparison schools?

((3000)(1) + (2000)(2))/5000 = 1.4

((3000)(2) + (2000)(3))/5000 = 2.4

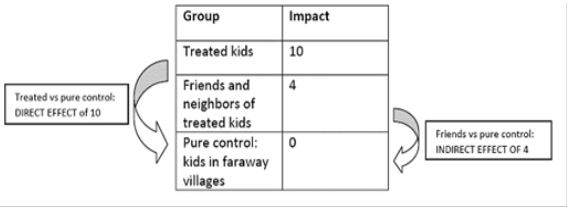
1.4 – 2.4 = -1

e. What is the indirect treatment effect due to spillovers?

-1

f. What is the total program effect?

(3/4)*(-1.47) + (1/4)*(-1) = -1.35



Total Program effect: 4+10=14

Traditional evaluations compare treated kids with others in the same village, and estimate an effect of 6. Or they compare treated kids with those in control villages, and estimate an effect of 10. Both underestimate the true impact.