**Evidence For Policy Design**

****

****

**Case 4: Evaluating Evaluations**

**Millennium Villages & Deworming**

## Key Vocabulary

|  |
| --- |
| 1. **Internal validity**: a study is internally valid if the estimated impact can be interpreted convincingly as the direct causal impact of the program.
2. **External validity:** a study is externally valid if the results can be replicated in different populations and settings.
3. **Statistical significance**: a result is statistically significant at the 5% level if there is less than a 5% probability that the result would be observed purely due to sampling variation.
4. **Equivalence:** groups are identical on all baseline characteristics, both observable and unobservable. Ensured by randomization.
5. **Attrition:** the process of individuals joining or dropping out of either the treatment or comparison group over the course of the study.
 |

# Case 1: Millennium Villages

Please read the following excerpts from an article published in May 2012 presenting an impact evaluation of the Millennium Villages project.[[1]](#footnote-1) The article describes a study conducted by the project implementers to assess its impact of the project on nine rural villages in nine different countries: Nigeria, Mali, Senegal, Ghana, Uganda, Kenya, Rwanda, Tanzania, and Malawi. The main outcome of interest is the mortality rate of children younger than five years of age.

*“The Millennium Villages project is a 10-year initiative supporting the integrated delivery of a package of scientifically proven interventions with the central aim of achieving the Millennium Development Goals (MDGs[[2]](#footnote-2)) across diverse sub-Saharan African sites. Local partnerships between the project, communities and governments coordinate activities across many sectors including health, agriculture, the environment, business development, education, and infrastructure....*

*“...A core set of interventions for achieving the MDGs have been identified by the UN Millennium Project. These interventions were adapted and flexibly implemented in response to local conditions after consultation with governments and local communities....*

*“...In the health sector, basic services were often unavailable at baseline, requiring major investments in infrastructure and staffing. Governments were core partners and remained responsible for employing local professional staff and managing facilities and supply chains. To reduce access barriers, free primary health care was made available at nearly all sites as even modest co-payments can restrict access among the poorest. An evidence-based package of maternal-child health interventions was introduced in line with national and World Health Organization guidelines.*

*In agriculture, improved seeds and fertilizers were subsidized to support high-yielding crop varieties alongside farmer training on best agronomic practices. Interventions in education included upgrading buildings and classrooms, making learning materials available, recruiting qualified teachers, and providing school meals. Finally, these efforts were combined with investments in basic infrastructure to enhance access to improved drinking water and sanitation, upgrade local roads, promote partnerships to expand mobile-phone coverage, and improve facility access to grid and solar electricity.”*

**Section 1: Mortality Trends**

The first result cited in the paper is presented in the following table:

|  |
| --- |
| **Child mortality trend in nine Millennium Villages**2006-2009 |
| **Indicator** | **Observational unit** | **Year 0\*** | **Year 3** | **Absolute change** |
| Mortality rate in children younger than five years of age (deaths per 1000 births) | Children younger than five years of age | 113.3 | 88.7 | -24.6 |

1. What evaluation method is being used in this example?

|  |
| --- |
| **ANSWER:**  |

1. What represents the counterfactual in this evaluation method?

|  |
| --- |
| **ANSWER:**  |

1. What is the estimated impact of the program on child mortality, using the data in the table?

|  |
| --- |
| **ANSWER:**  |

1. What assumption is needed in order for this evaluation method to produce a valid causal estimate of program impact (internal validity)?

|  |
| --- |
| **ANSWER:**  |

1. Is this assumption likely to be met? Why or why not?

|  |
| --- |
| **ANSWER:**  |

# Section 2: Comparing Trends

The results presented in the table above imply an annual rate of decline in child mortality of 7.8% in Millennium Villages from 2006 to 2009. In the original study, this rate is then compared with national declines in child mortality across rural areas between 2001 and 2010. The average national rate of decline over this period was 2.6%. (The difference between the two rates is not statistically significant.)

1. Does the national trend from 2001-2010 represent a good counterfactual for the rate of decline in Millennium Villages? Why or why not?

|  |
| --- |
| **ANSWER:**  |

1. What would make a better comparison for the rate of decline in Millennium Villages?

|  |
| --- |
| **ANSWER:**  |

In order to strengthen the estimate of the project’s effect on child mortality, the evaluation team decided to compare the villages that participated in the program to nine similar villages from the same nine countries. First, researchers matched participating villages with up to three non-participating villages based on similarity in factors thought to influence child mortality and other Millennium Development Goals. Then, one comparison village was randomly selected for each intervention village from among the potential matches. The matching was done at year 3.

1. How does this study design differ from a randomized evaluation?

|  |
| --- |
| **ANSWER:**  |

The following table contains data on baseline outcomes for both Millennium Village sites and the comparison sites. The data was collected in a household survey three years after the intervention began (2009), but the questions referred to outcomes three years prior (2006). The households included in the survey were selected through a random sample of 300 households from each Millennium Village and comparison village.

|  |
| --- |
| **Baseline outcomes**(2006, measured in 2009) |
| **Baseline outcome** | **Millennium Village sites**(95% confidence interval) | **Comparison village sites**(95% confidence interval) | **Difference statistically significant at 5% level?** |
| Asset-based wealth index | 41.0(38.3 to 43.7) | 39.0(36.4 to 41.7) | No |
| Skilled birth attendance | 32.6%(26.6 to 39.1) | 25.9%(20.7 to 31.8) | No |
| Access to antenatal care | 45.3%(29.0 to 62.8) | 46.0%(29.5 to 63.4) | No |
| Under-five mortality rate | 113(99 to 128) | 90(77 to 103) | Yes |

1. Do you see any problem with collecting 2006 baseline data in 2009, when the intervention had already been implemented for three years?

|  |
| --- |
| **ANSWER:**  |

1. What factors could account for the difference in baseline under-five mortality rates between Millennium Village sites and comparison sites?

|  |
| --- |
| **ANSWER:**  |

1. Are these baseline differences between the treatment and comparison villages a threat to internal validity?

|  |
| --- |
| **ANSWER:**  |

Using the nine selected comparison village sites, the researchers provide the following information on the main outcome of interest – under-five mortality rates.

|  |
| --- |
| **Mortality rate in children younger than five years of age**(deaths per 1000 births) |
|  | **Comparison village sites** | **Millennium Village sites** | **Difference** |
| Baseline (2006) | 90.3 | 113.3 | -23.0 |
| Endline (2009) | 96.2 | 88.7 | 7.5 |
| Difference | -5.9 | 24.6 | -30.5\* |
| \*Statistically significant at the 5% level |

1. What is the evaluation method being used in this example?

|  |
| --- |
| **ANSWER:**  |

1. What represents the counterfactual in this evaluation method?

|  |
| --- |
| **ANSWER:**  |

1. What is the estimated impact of the program on child mortality, using the data in the table?

|  |
| --- |
| **ANSWER:**  |

1. What assumption is needed in order for this evaluation method to produce to a valid causal estimate of program impact (internal validity)?

|  |
| --- |
| **ANSWER:**  |

1. Is this assumption likely to hold in this case?

|  |
| --- |
| **ANSWER:**  |

1. What other evidence could be used to convince you that the required assumption holds?

|  |
| --- |
| **ANSWER:**  |

# Section 3: Program Theory & External Validity

Suppose you are interested not only in the magnitude of the program’s impact, but also in why the program worked and whether it could work elsewhere. For the moment, leave aside any doubts you may have about the evaluation’s internal validity, and assume that the impact estimates are valid.

1. Can this evaluation answer the question of which specific interventions (whether in health, education, infrastructure, or agriculture) were necessary to achieve reductions in child mortality? Why or why not?

|  |
| --- |
| **ANSWER:**  |

1. Discuss whether you think that this type of program would have a similar impact in your home country? That is, does this study have external validity when applied to communities in your country?

# Case 2: Deworming in Kenya[[3]](#footnote-3)

# Introduction

This evaluation investigates the effect of a school-based mass deworming program conducted by the NGO International Child Support Africa. The program treated 30,000 pupils enrolled at 75 Kenyan schools for worms – hookworm, roundworm, whipworm, and schistosomiasis.

Worm infections account for over 40 percent of the global tropical disease burden. Infections are common in areas with poor sanitation. More than 2 billion people are affected. Children, still learning good sanitary habits, are particularly vulnerable: 400 million school-age children are chronically infected with intestinal worms.

Worms affect more than the health of children. Symptoms include listlessness, diarrhea, abdominal pain, and anemia. Beyond their effects on health and nutrition, heavy worm infections can impair children’s physical and mental development and lower their attendance and performance in school.

Treatment kills worms in the body, but does not prevent reinfection. Oral medication that can kill 99 percent of worms in the body is available and costs less than 20 US cents per patient. The drugs have very few and minor side effects. Schools with hookworm, whipworm, and roundworm prevalence over 50 percent should be mass treated with albendazole every six months, and schools with schistosomiasis prevalence over 30 percent should be mass treated with praziquantel once a year.

# The Primary School Deworming Program

International Child Support Africa (ICS) implemented the Primary School Deworming Program in the Busia District in western Kenya, a densely settled region with high worm prevalence. The medicine was administered in schools by public health nurses from the Ministry of Health. The program was expected to affect health, nutrition, and education. To measure impact, ICS collected data on a series of outcomes: prevalence of worm infection, severity of worm infection, self-reported illness, and school participation rates and test scores.

Because of administrative and financial constraints, the program could not be implemented in all schools immediately. Instead, the 75 schools were randomly divided into three groups of 25 schools and phased-in over three years. Group 1 schools were treated starting in both 1998 and 1999, Group 2 schools in 1999, and Group 3 starting in 2001. Group 1 schools were the treatment group in 1998, while schools Group 2 and Group 3 were the comparison. In 1999 Group 1 and Group 2 schools were the treatment and Group 3 schools the comparison. The following table illustrates the experimental design.

|  |
| --- |
| **Primary School Deworming Program:****Treatment Timeline** |
|  | **1998** | **1999** | **2001** |
| Group 1 | **Treatment** | **Treatment** | **Treatment** |
| Group 2 | Comparison | **Treatment** | **Treatment** |
| Group 3 | Comparison | Comparison | **Treatment** |

# Section 1: Evaluation Method & Internal Validity

1. What evaluation method is being used in this example?

|  |
| --- |
| **ANSWER:**  |

1. What represents the counterfactual in this evaluation method?

|  |
| --- |
| **ANSWER:**  |

1. What assumption is needed in order for this evaluation method to produce to a valid causal estimate of program impact (internal validity)?

|  |
| --- |
| **ANSWER:**  |

1. Can you test whether this assumption is met? If so, how?

|  |
| --- |
| **ANSWER:**  |

# Section 2: Assessing Equivalence

Randomization ensures that the groups are on average equivalent, and therefore comparable, at the beginning of the program. The impact is then estimated as the difference in the average outcome of the treatment group and the average outcome of the comparison group, both at the end of the program. To be able to say that the program caused the impact, you need to be able to say that the program was the only difference between the treatment and comparison groups over the course of the evaluation.

1. Using the information in the following table, check for equivalence in the treatment and control groups. Does randomization appear to have worked?

|  |
| --- |
| **1998 Average Pupil and School Characteristics, Pre-Treatment** |
|  | **Group 1****(25 schools)** | **Group 2****(25 schools)** | **Group 3****(25 schools)** |  | **Group 1 – Group 3** | **Group 2 – Group 3** |
| Male | 0.53 | 0.51 | 0.52 |  | 0.01 | -0.01 |
| Year of birth | 1986.2 | 1986.5 | 1985.5 |  | 0.4\* | 0.8\* |
| Attendance recorded in school registers, prior four weeks | 0.973 | 0.963 | 0.969 |  | 0.003 | -0.006 |
| Access to latrines at home | 0.82 | 0.81 | 0.82 |  | 0.00 | -0.01 |
| School latrines per pupil | 0.007 | 0.006 | 0.007 |  | 0.001 | -0.000 |
| Weight-for-age Z-score (low means more undernourished) | -1.39 | -1.40 | -1.44 |  | 0.05 | 0.04 |
| Sick often (self-reported) | 0.10 | 0.10 | 0.08 |  | 0.02\* | 0.02\* |
| Pupil population of school | 392.7 | 408.3 | 375.9 |  | 16.8 | 27.9 |
| \*Statistically significant at the 5% level |

|  |
| --- |
| **ANSWER:**  |

1. Other than the program’s direct and indirect impacts, what could happen over the course of the evaluation (after conducting the random assignment) to make the groups non-equivalent?

|  |
| --- |
| **ANSWER:**  |

1. How does non-equivalence at the end threaten the integrity of the experiment?

|  |
| --- |
| **ANSWER:**  |

# Section 3: External validity

1. In Kenya, it was found that a 100 US dollar investment in school-based deworming could achieve 13.9 additional years of student participation. A team from J-PAL also assessed the impact of deworming in India. They found a smaller but still positive impact on participation – in India, a 100 US dollar investment in deworming as well as iron fortification in preschools bout 2.7 additional years of student participation. Why do you think the magnitude of the impact is so different in India?

|  |
| --- |
| **ANSWER:**  |

1. Do you think that this evaluation would have external validity when applied to your country?

**Section 4: Attrition**

Attrition is when people join or drop out of the sample – both treatment and comparison groups – over the course of the experiment. One common example in clinical trials is when people die; so common indeed that attrition is sometimes called experimental mortality.

You are looking at the health effects of deworming. In particular you are looking at the worm load (severity of worm infection). Worm loads are scaled as follows:

Heavy worm infections = score of 3

Medium worm infections = score of 2

Light infections = score of 1

Suppose there are 30,000 children: 15,000 in treatment schools and 15,000 in comparison schools. After you randomize, the treatment and comparison groups are equivalent, meaning children from each of the three categories are equally represented in both groups.

Assume that protocol compliance is 100 percent: all children who are in the treatment get treated and none of the children in the comparison are treated. Children that were dewormed at the beginning of the school year (that is, children in the treatment group) end up with a worm load of 1 at the end of the year because of re-infection. Children who have a worm load of 3 only attend half the time and drop out of school if they are not treated. The number of children in each worm-load category is shown for both the pre-test and post-test.

|  |
| --- |
| **World Loads Pre and Post, by Treatment Group** |
|  | **Pre-test** | **Post-test** |
| Worm Load | Treatment | Comparison | Treatment | Comparison |
| 3 | 5,000 | 5,000 | 0 | **Dropped out** |
| 2 | 5,000 | 5,000 | 0 | 5,000 |
| 1 | 5,000 | 5,000 | 15,000 | 5,000 |
| Total children tested at school | 15,000 | 15,000 | 15,000 | 10,000 |

1. Please answer the following questions:
	* 1. At post-test, what is the average worm load for the treatment group?

|  |
| --- |
| **ANSWER:**  |

* + 1. At post-test, what is the average worm load for the comparison group?

|  |
| --- |
| **ANSWER:**  |

* + 1. What is the difference?

|  |
| --- |
| **ANSWER:**  |

* + 1. Is this outcome difference an accurate estimate of the impact of the program? Why or why not?

|  |
| --- |
| **ANSWER:**  |

* + 1. If it is not accurate, does it overestimate or underestimate the impact?

|  |
| --- |
| **ANSWER:**  |

* + 1. How can we get a better estimate of the program’s impact?

|  |
| --- |
| **ANSWER:**  |

1. You have learned about other methods to estimate program impact, such as pre-post, simple difference, differences-in-differences, and multivariate regression. Does the threat of attrition only present itself in randomized evaluations?

|  |
| --- |
| **ANSWER:**  |

1. Pronyk PM, Muniz M, Nemser B, et al, for the Millennium Villages Study Group (2012). “The effect of an integrated multisector model for achieving the Millennium Development Goals and improving child survival in rural sub-Saharan Africa: a non-randomized controlled assessment.” *Lancet*, published online May 8. DOI:10.1016/S0140-6736(12)60207-4. [↑](#footnote-ref-1)
2. The Millennium Development Goals are eight international development goals agreed upon by UN member states. Defined in 2000, to be achieved by 2015, the goals are: 1) end poverty and hunger, 2) achieve universal primary education, 3) promote gender equality and empower women, 4) reduce child mortality, 5) improve maternal health, 6) combat HIV/AIDS, malaria, and other diseases, 7) ensure environmental sustainability, and 8) develop a global partnership for development. [↑](#footnote-ref-2)
3. This case was adapted from the original J-PAL case study “Deworming in Kenya: Addressing threats to experimental integrity”, and is based on Miguel E & Kremer M (2004).“Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities,”*Econometrica*, 72(1): 159-217. [↑](#footnote-ref-3)