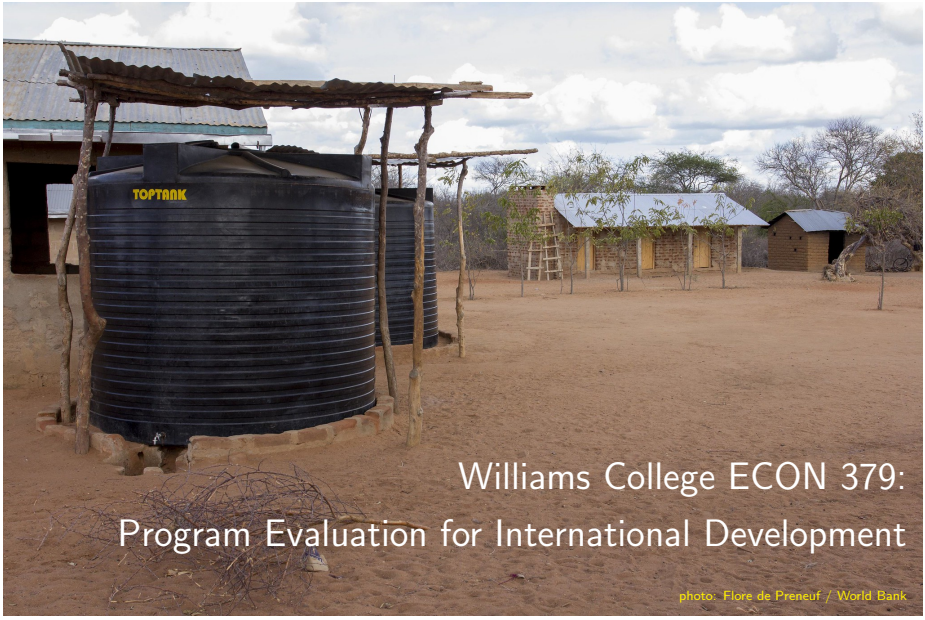
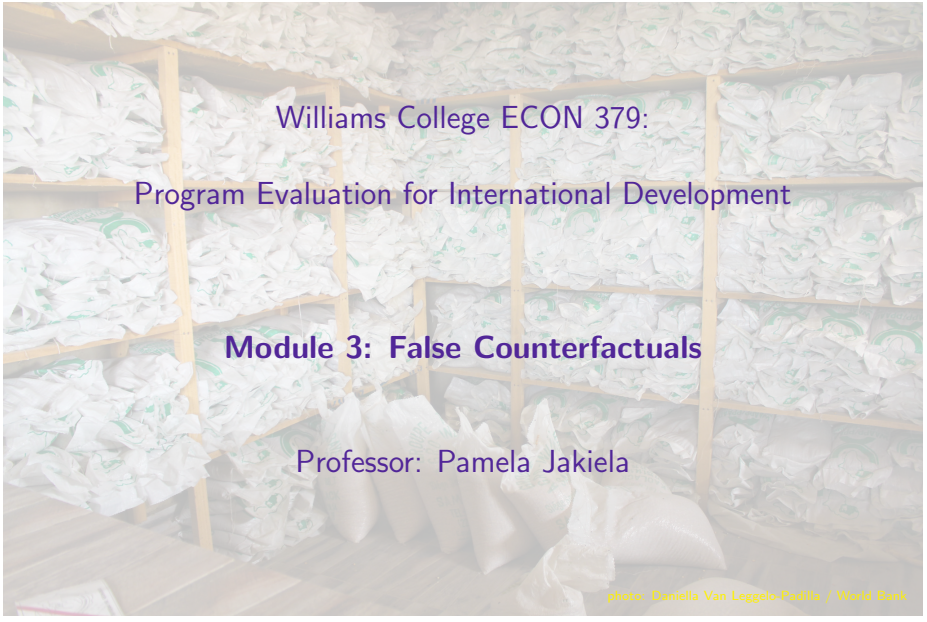


Williams College ECON 379:
Program Evaluation for International Development

photo: Flore de Preneuf / World Bank



Williams College ECON 379:

Program Evaluation for International Development

Module 3: False Counterfactuals

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

False Counterfactuals

False Counterfactual #1

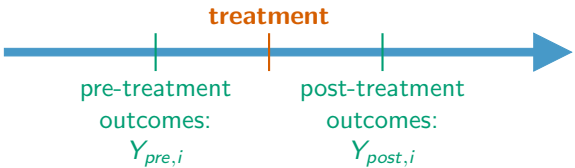
Participant vs. Non-Participant Comparisons:

- Compares individuals/communities/units/etc that chose to participate in a program to those that did not
- Treatment effect conflated with . . . ?

The Experimental Ideal:

- Treatment is randomly assigned
- Treatment, control groups similar in terms of observable, unobservable characteristics (at least in expectation)
- Difference in means = unbiased estimate of treatment effect

False Counterfactual #2: Before/After Comparison



Three Approaches to Estimating Treatment Effects

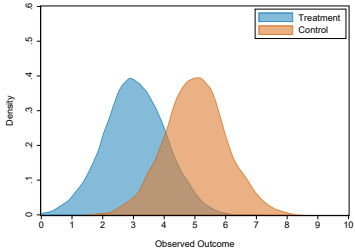
Any of these **can** be credible (all require assumptions):

1. The experimental ideal: randomized treatment vs. control
2. Participant vs. non-participant comparisons $\rightarrow$ selection bias?
3. Pre vs. post comparisons $\rightarrow$ changes over time?

In each case, we ask:

1. Is there a difference (in means) between groups?
2. Is that difference likely to have occurred by chance?
3. Should we interpret difference as causal effect of treatment?

Testing $H_0 : \bar{Y}_T = \bar{Y}_C$



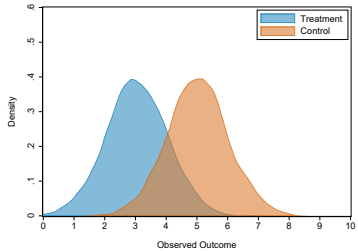
When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$SE_{\bar{Y}_T} = \sqrt{\frac{s^2_T}{n_T}} = \sqrt{\frac{\sum_{i \in T}\left(Y_i - \bar{Y}\right)^2}{n_T(n_T - 1)}}$$

where n_T is treatment observations, and $\sum_{i \in T}$ sums over treated i

Testing $H_0 : \bar{Y}_T = \bar{Y}_C$



When $\bar{Y}_T$ and $\bar{Y}_C$ are independent:

$$SE\left(\bar{Y}_T - \bar{Y}_C\right) = \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$$\Rightarrow t = \left(\bar{Y}_T - \bar{Y}_C\right) / \sqrt{SE^2_{\bar{Y}_T} + SE^2_{\bar{Y}_C}}$$

$\Rightarrow$ p-value – which is what, again?

When $\bar{Y}_T$ and $\bar{Y}_C$ are NOT independent, calculate the change, and test null hypothesis that the mean (change in Y) is equal to 0

Testing $H_0 : \bar{Y}_T = \bar{Y}_C$ in Stata

```
. ttest y, by(t)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      50     3.95408    .1318144    .9320683    3.689189    4.218971
       1 |      50    5.423295    .1445738    1.022291    5.132763    5.713827
---------+--------------------------------------------------------------------
combined |     100    4.688688    .1221617    1.221617    4.446292    4.931083
---------+--------------------------------------------------------------------
    diff |           -1.469215    .1956441               -1.857465   -1.080966
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -7.5096
Ho: diff = 0                     degrees of freedom =               98

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

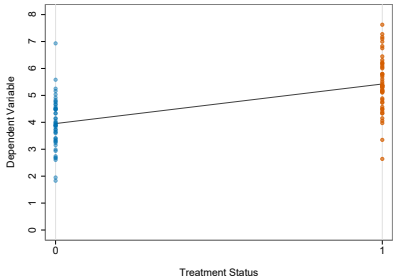
$$t = \frac{\bar{Y}_T - \bar{Y}_C}{SE\left(\bar{Y}_T - \bar{Y}_C\right)}$$

$$= \frac{-1.4692}{0.1956}$$

$$= 7.5096$$

To get p-value (in Stata): `display 2*(1 - abs(ttail(98,-7.5096)))`

Testing $H_0 : \bar{Y}_T = \bar{Y}_C$ in a Regression



OLS regression on a binary independent variable: $Y_i = \alpha + \beta D_i + \varepsilon_i$

- Only two possible predicted values of $\hat{Y}_i$: α and $\alpha + \beta$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} \\ &= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}\end{aligned}$$

Notice that the numerator can be re-organized:

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_i X_i Y_i - \sum_i \bar{X} Y_i - \sum_i \bar{Y} X_i + \sum_i \bar{X}\bar{Y}$$

$$= \bar{Y} \sum_i X_i = \bar{Y} N \bar{X} \qquad = N \bar{X} \bar{Y}$$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} \\ &= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}\end{aligned}$$

Notice that the numerator can be re-organized:

$$\begin{aligned}\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_i X_i Y_i - \sum_i \bar{X} Y_i \\ &= \sum_i [Y_i (X_i - \bar{X})]\end{aligned}$$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\hat{\beta}_{OLS} = \frac{COV(X,Y)}{VAR(X)}$$

$$= \frac{\sum_i [Y_i(X_i - \bar{X})]}{\sum_i (X_i - \bar{X})^2}$$

When independent variable is binary:

$\bar{X} = \frac{n_T}{N}$ (n_T is # of treated observations)

Assume observations are ordered:

$$\{\underbrace{Y_1, Y_2, \ldots, Y_{n_{T-1}}, Y_{n_T}}_{\text{treatment group}}, \underbrace{Y_{n_{T+1}}, Y_{n_{T+2}}, \ldots, Y_N}_{\text{control group}}\}$$

$X_i = 1$

$\Rightarrow X_i - \bar{X} = 1 - \bar{X}$

$X_i = 0$

$\Rightarrow X_i - \bar{X} = -\bar{X}$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\hat{\beta}_{OLS} = \frac{COV(X,Y)}{VAR(X)} = \frac{\sum_i \left[Y_i\left(X_i - \bar{X}\right)\right]}{\sum_i \left(X_i - \bar{X}\right)^2}$$

When independent variable is binary:

$\bar{X} = \frac{n_T}{N}$ (n_T is # of treated observations)

Assume observations are ordered:

$$\{\underbrace{Y_1, Y_2, \ldots, Y_{n_T-1}, Y_{n_T}}_{\text{treatment group}}, \underbrace{Y_{n_T+1}, Y_{n_T+2}, \ldots, Y_N}_{\text{control group}}\}$$

Re-write denominator:

$$\begin{aligned}\sum_i \left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n_T} \left(1 - \bar{X}\right)^2 + \sum_{i=n_T+1}^{N} \left(-\bar{X}\right)^2 \\ &= n_T \left(1 - \bar{X}\right)^2 + \left(N - n_T\right)\left(-\bar{X}\right)^2 \\ &= \ldots = n_T - n_T\bar{X} = N\bar{X}\left(1 - \bar{X}\right) \quad (\text{because } n_t = N\bar{X})\end{aligned}$$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} \\ &= \frac{\sum_i \left[Y_i\left(X_i-\bar{X}\right)\right]}{\sum_i \left(X_i-\bar{X}\right)^2} \\ &= \frac{\sum_i \left[Y_i\left(X_i-\bar{X}\right)\right]}{N\bar{X}\left(1-\bar{X}\right)}\end{aligned}$$

Re-write denominator:

$$\begin{aligned}\sum_i \left(X_i-\bar{X}\right)^2 &= \sum_{i=1}^{n_T}\left(1-\bar{X}\right)^2 + \sum_{i=n_{T+1}}^{N}\left(-\bar{X}\right)^2 \\ &= n_T\left(1-\bar{X}\right)^2 + \left(N-n_T\right)\left(-\bar{X}\right)^2 \\ &= \ldots = n_T - n_T\bar{X} = N\bar{X}\left(1-\bar{X}\right)\end{aligned}$$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} \\ &= \frac{\sum_i \left[Y_i\left(X_i-\bar{X}\right)\right]}{\sum_i \left(X_i-\bar{X}\right)^2} \\ &= \frac{\sum_i \left[Y_i\left(X_i-\bar{X}\right)\right]}{N\bar{X}\left(1-\bar{X}\right)}\end{aligned}$$

$$\sum_i \left[Y_i \left(X_i - \bar{X}\right)\right]$$

↑

"like a weighted average"

(weights sum to 0, not 1)

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$
\begin{aligned}
\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} & & \sum_i \left[Y_i \left(X_i - \bar{X} \right) \right] \\
&= \frac{\sum_i \left[Y_i \left(X_i - \bar{X} \right) \right]}{\sum_i \left(X_i - \bar{X} \right)^2} & &= \sum_{i=1}^{n_T} \left[Y_i \left(1 - \bar{X} \right) \right] + \sum_{i=n_{T+1}}^{N} \left[Y_i \left(-\bar{X} \right) \right] \\
&= \frac{\sum_i \left[Y_i \left(X_i - \bar{X} \right) \right]}{N\bar{X}\left(1 - \bar{X} \right)} & &= \sum_{i=1}^{n_T} Y_i - \sum_{i=1}^{N} \left[Y_i \left(\bar{X} \right) \right] \\
& & &= n_T \bar{Y}_T - \bar{X} \left(N\bar{Y} \right) \\
& & &= N\bar{X}\bar{Y}_T - N\bar{X} \left[\bar{X}\bar{Y}_T + \left(1 - \bar{X} \right) \bar{Y}_C \right] \\
& & &= N\bar{X} \left(1 - \bar{X} \right) \left(\bar{Y}_T - \bar{Y}_C \right)
\end{aligned}
$$

OLS Regression on a Binary Independent Variable

You may or may not remember that in a bivariate regression:

$$
\begin{aligned}
\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} && \sum_i \left[Y_i \left(X_i - \bar{X} \right) \right] \\
&= \frac{\sum_i \left[Y_i \left(X_i - \bar{X} \right) \right]}{\sum_i \left(X_i - \bar{X} \right)^2} && = \sum_{i=1}^{n_T} \left[Y_i \left(1 - \bar{X} \right) \right] + \sum_{i=n_{T+1}}^{N} \left[Y_i \left(-\bar{X} \right) \right] \\
&= \frac{\sum_i \left[Y_i \left(X_i - \bar{X} \right) \right]}{N\bar{X}\left(1 - \bar{X} \right)} && = \sum_{i=1}^{n_T} Y_i - \sum_{i=1}^{N} \left[Y_i \left(\bar{X} \right) \right] \\
&= \frac{N\bar{X}\left(1 - \bar{X} \right)\left(\bar{Y}_T - \bar{Y}_C \right)}{N\bar{X}\left(1 - \bar{X} \right)} && = n_T \bar{Y}_T - \bar{X} \left(N\bar{Y} \right) \\
& && = N\bar{X}\bar{Y}_T - N\bar{X} \left[\bar{X}\bar{Y}_T + \left(1 - \bar{X} \right) \bar{Y}_C \right] \\
& && = N\bar{X} \left(1 - \bar{X} \right) \left(\bar{Y}_T - \bar{Y}_C \right)
\end{aligned}
$$

OLS Regression on a Binary Independent Variable

$$
\begin{aligned}
\hat{\beta}_{OLS} &= \frac{COV(X,Y)}{VAR(X)} \\
&= \frac{\sum_i \left[Y_i\left(X_i - \bar{X}\right)\right]}{\sum_i \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum_i \left[Y_i\left(X_i - \bar{X}\right)\right]}{N\bar{X}\left(1-\bar{X}\right)} \\
&= \frac{N\bar{X}\left(1-\bar{X}\right)\left(\bar{Y}_T - \bar{Y}_C\right)}{N\bar{X}\left(1-\bar{X}\right)} \\
&= \bar{Y}_T - \bar{Y}_C
\end{aligned}
$$

OLS Regression on a Binary Independent Variable

When we regress Y_i on (only) a dummy variable:

$$\hat{\beta}_{OLS} = \bar{Y}_T - \bar{Y}_C$$

- Estimated constant $\hat{\alpha}_{OLS}$ is control group mean, also $\hat{Y}_i$
- Predicted $\hat{Y}_i$ for treated individuals/units is $\hat{\alpha}_{OLS} + \hat{\beta}_{OLS}$

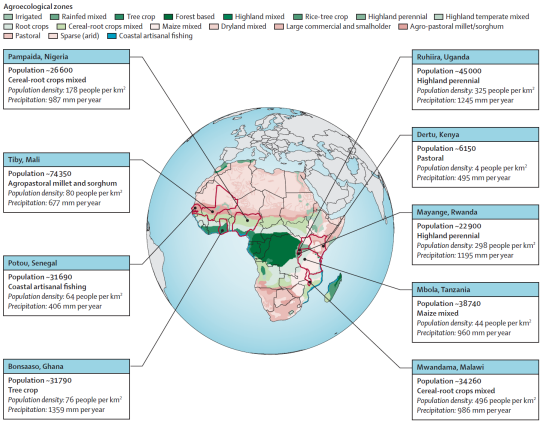
Why Did We Do This, Again?

"And the end of all our exploring will be to arrive where we started
and know the place for the first time."

–T.S. Eliot in *Four Quartets*

Millennium Villages

The Millennium Villages Project



source: Pronyk et al. (2012)

The Impacts of the Millennium Villages Project?

	Millennium Village sites (N=9)	Comparison village sites (N=9)
Village characteristics (at Year 0)		
Land area (km²)	133·2 (102·2 to 164·1)	128·2 (97·2 to 159·1)
Number of sites with electricity	0·0%*	0·0%*
Number of sites with cellular coverage	78% (39 to 95)	78% (39 to 95)
Distance to nearest main town (km)	11·9 (8 to 15·8)	12·6 (8·7 to 16·5)
Distance from centre of village to nearest paved road (km)	14·8 (0·8 to 28·7)	14·5 (0·5 to 28·4)
Number of months road not accessible to vehicles	2·3 (2·0 to 2·7)	2·5 (2·2 to 2·8)
Distance to clinic (km)	5·6 (1·8 to 9·5)	10·2 (6·3 to 14·1)
Number of NGOs or partners per site	1·3 (0·8 to 1·9)	1·4 (0·9 to 2)
Number of facilities per 10 000 people		
Markets	0·7 (–0·4 to 1·7)	1·4 (0·4 to 2·5)
Primary schools	5·6 (–0·4 to 11·5)	8·6 (2·6 to 14·5)
Secondary schools	0·0*	0·0*
Clinics	0·7 (–0·8 to 2·1)	1·3 (–0·1 to 2·7)
Number of sites that have no irrigation of cultivatable land	33·3% (10·0 to 69·1)	33·3% (10·0 to 69·1)
Religion (% of population that is Christian)	47% (32·7 to 61·4)	38% (23·4 to 52·1)

Data are mean (95% CI). Information on village infrastructure is from the village matching checklist. The characteristics of households are from the year-3 household survey. Baseline outcomes are calculated on the basis of reproductive and pregnancy histories collected from women at year 3. The asset-based wealth index is scaled to have a mean of 50 (SD 25). *Interval has zero width because there is no variance in this characteristic across sites.

Table 1: Characteristics of Millennium Villages and comparison villages

source: Pronyk et al. (2012)

	Millennium Village sites (N=9)	Comparison village sites (N=9)
Characteristics of households (at year 3)		
Household head has no primary education	87·1% (83·1 to 90·3)	87·9% (84·1 to 90·9)
Household head is a woman	14·3% (10·2 to 19·7)	11·3% (7·9 to 16)
Household head's main livelihood strategy is farming	81·9% (77·2 to 85·9)	85·1% (80·9 to 88·5)
Household size	7·1 (5·7 to 8·6)	5·9 (4·5 to 7·3)
Dependency ratio	138·2 (132·6 to 143·7)	131·9 (126·3 to 137·4)
Age of adult female household members	33·0 (32·3 to 33·8)	31·9 (31·1 to 32·7)
Baseline outcomes (at year 0)		
Asset-based wealth index	41·0 (38·3 to 43·7)	39·0 (36·4 to 41·7)
Skilled birth attendance	32·6% (26·6 to 39·1)	25·9% (20·7 to 31·8)
Access to antenatal care	45·3% (29 to 62·8)	46·0% (29·5 to 63·4)
Mortality rate in children younger 5 years of age	113 (99 to 128)	90 (77 to 103)

Data are mean (95% CI). Information on village infrastructure is from the village matching checklist. The characteristics of households are from the year-3 household survey. Baseline outcomes are calculated on the basis of reproductive and pregnancy histories collected from women at year 3. The asset-based wealth index is scaled to have a mean of 50 (SD 25). *Interval has zero width because there is no variance in this characteristic across sites.

Table 1: Characteristics of Millennium Villages and comparison villages

source: Pronyk et al. (2012)

The Impacts of the Millennium Villages Project?

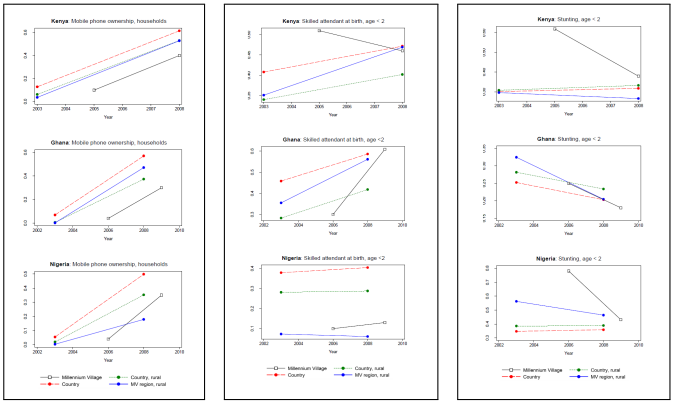
	Observational unit	Millennium Village sites (N=9)				Comparison village sites (N=9)				Millennium Villages vs comparison villages in year 3	
		Year 0 (number)	Year 3 (number)	Absolute change (95% CI)	p value	Year 0 (number)	Year 3 (number)	Absolute change (95% CI)	p value	Absolute difference (95% CI)	p value
Wasting	Children younger than 2 years of age‡	6·4% (271)	5·5% (644)	−0·9% (−4·1 to 2·4)	0·591	..	6·7% (776)	..	..	−1·2% (−6·5 to 4·2)§	0·630
Underweight	Children younger than 2 years of age‡	13·1% (279)	14·3% (660)	1·2% (−4·2 to 6·6)	0·669	..	16·1% (803)	..	..	−1·8% (−8·9 to 5·4)§	0·584
Stunting	Children younger than 2 years of age‡	36·0% (255)	28·2% (709)	−7·9% (−15·6 to −0·2)	0·045	..	35·7% (784)	..	..	−7·5% (−20·0 to 5·0)§	0·205
Mortality rate in children younger than 5 years of age (deaths per 1000 births)	Children younger than 5 years of age*	113·3 (5336)	88·7 (4905)	−24·6 (−44·5 to −4·8)	0·015	90·3 (4093)	96·2 (3933)	5·9 (−13·8 to 25·7)	0·556	−30·5 (−58·5 to −2·5)†	0·033

source: Pronyk et al. (2012)

Critiques of the MVP Evaluation

Bump, Clemens, Demombynes, and Haddad (2012) raise three issues:

The Impacts of the Millennium Villages Project?



source: Clemens and Demombynes. (2010)

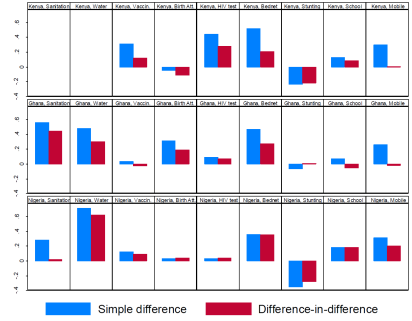
The Impacts of the Millennium Villages Project?

Indicator	All Kenya		Rural Kenya		Millennium Village Region		Millennium Village		Simple difference within Millennium Village
	2003	2008	2003	2008	2003	2008	2005	2008	
Percentage of households with access to improved sanitation facilities	0.37	0.50	0.28	0.37	0.20	0.27			
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.04)			
Percentage of households with access to an improved source of drinking water	0.53	0.63	0.45	0.54	0.38	0.49			
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.05)			
Percentage of children aged 1 who have received the measles vaccination prior to survey	0.73	0.85	0.70	0.83	0.44	0.75	0.67	0.98	0.31
	(0.02)	(0.02)	(0.02)	(0.02)	(0.08)	(0.05)			
Percentage of births in the last two years for which a skilled attendant was present	0.41	0.47	0.34	0.40	0.35	0.47	0.51	0.46	-0.05
	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)			
Percentage of men and women aged 15 - 49 who were tested for HIV in the last 12 months	0.08	0.27	0.06	0.25	0.08	0.35	0.14	0.58	0.44
	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)			
Percentage of children under 5 who slept under an ITN bednet last night	0.06	0.47	0.05	0.44	0.08	0.59	0.10	0.62	0.52
	(0.00)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)			
Percentage of children under 2 who are 2 standard deviations below the median height for their age	0.30	0.32	0.31	0.33	0.30	0.27	0.62	0.38	-0.24
	(0.01)	(0.02)	(0.01)	(0.02)	(0.04)	(0.02)			
Gross Primary Attendance Ratio	1.11	1.13	1.13	1.14	1.18	1.27	1.10	1.23	0.13
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)			
Percentage of households with at least one mobile phone	0.13	0.61	0.06	0.53	0.04	0.53	0.10	0.40	0.30
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)			

source: Clemens and Demombynes. (2010)

The Impacts of the Millennium Villages Project?

Figure 4: Summary comparison of trends within the Millennium Villages against the same trends relative to other rural households in the surrounding region



source: Clemens and Demombynes. (2010)

The end!