Routledge
Taylor & Francis Group

Journal of Development Effectiveness

ISSN: 1943-9342 (Print) 1943-9407 (Online) Journal homepage: http://www.tandfonline.com/loi/rjde20

When does rigorous impact evaluation make a difference? The case of the Millennium Villages

Michael A. Clemens & Gabriel Demombynes

To cite this article: Michael A. Clemens & Gabriel Demombynes (2011) When does rigorous impact evaluation make a difference? The case of the Millennium Villages, Journal of Development Effectiveness, 3:3, 305-339, DOI: 10.1080/19439342.2011.587017

To link to this article: https://doi.org/10.1080/19439342.2011.587017

Journal of Development Effectiveness
Vol. 3, No. 3, September 2011, 305–339

Routledge
Taylor & Francis Group

When does rigorous impact evaluation make a difference? The case of the Millennium Villages

Michael A. Clemens[a*] and Gabriel Demombynes[b]

[a]*Center for Global Development, 1800 Massachusetts Avenue NW, 3rd Floor, Washington, DC 20036, USA;* [b]*World Bank, Nairobi, Kenya*

When is the rigorous impact evaluation of development projects a luxury, and when a necessity? The authors study one high-profile case: the Millennium Villages Project (MVP), an experimental and intensive package intervention to spark sustained local economic development in rural Africa. They illustrate the benefits of rigorous impact evaluation in this setting by showing that estimates of the project's effects depend heavily on the evaluation method. Comparing trends at the MVP intervention sites in Kenya, Ghana, and Nigeria with trends in the surrounding areas yields much more modest estimates of the project's effects than the before-versus-after comparisons published thus far by the MVP. Neither approach constitutes a rigorous impact evaluation of the MVP, which is impossible to perform due to weaknesses in the evaluation design of the project's initial phase. These weaknesses include the subjective choice of intervention sites, the subjective choice of comparison sites, the lack of baseline data on comparison sites, the small sample size, and the short time horizon. The authors describe one of many ways that the next wave of the intervention could be designed to allow proper evaluation of the MVP's impact at little additional cost.

Keywords: evaluation; village; package; Millennium Development Goals

1. Introduction

An intense debate is underway among those who seek policies to improve living conditions in developing countries. This debate concerns how much evidence is necessary to demonstrate that policies to improve those living conditions have the desired effect. Advocates of more rigorous impact evaluation argue that it can improve incentives for development agencies by increasing transparency and can avoid the waste of scarce resources on attractive but ineffective projects. Advocates of more indirect and heuristic impact evaluation methods argue that high demands for rigour are often better suited to academics than practitioners, focus inordinate attention on easily quantifiable outcomes, take too long to yield results, and divert scarce resources away from interventions already known to work well.

Different methods to estimate impacts are clearly suitable in different settings. But there is sharp disagreement about exactly when high standards of rigour in impact estimation are a luxury and when they are a necessity.

In this paper we dissect the case for and against rigorous impact evaluation in one concrete and high-profile setting, the Millennium Villages Project (MVP) – a large, experimental intervention that aims to spark local economic development in 14 village

*Corresponding author. Email: mclemens@cgdev.org

ISSN 1943-9342 print/ISSN 1943-9407 online

http://dx.doi.org/10.1080/19439342.2011.587017
http://www.tandfonline.com

clusters across Africa – and show that 'people in the poorest regions of rural Africa can lift themselves out of extreme poverty in five year's time' (MVP 2007a). We focus on the MVP because of its extraordinary prominence in development policy – it is the principal single policy initiative to emerge from the largest-ever gathering of heads of state, the Millennium Summit – and because the project's impact evaluation exemplifies so many common problems at once.

First, we show how initial estimates of the project's effects change substantially if more rigorous impact evaluation methods than those used in the project's mid-term evaluation report are employed. We highlight, however, that while our estimates are superior to the before-versus-after comparisons published thus far by MVP, they do not constitute a rigorous evaluation of the project's impact; the design of the initial phase of the project makes it impossible to make definitive statements about the project's effects. Second, we describe how weaknesses in the future evaluation design also limit its prospects of producing credible estimates of the project's effects. These weaknesses include the subjective choice of treated villages, the subjective choice of comparison villages, the lack of baseline data on the comparison villages, the small sample size, and the short time horizon. Finally, we discuss how the initial phase of the project could have been designed – and future phases could be designed – to permit greater accuracy and clarity in assessing its impact.

Our contribution is to illustrate one important case where, although a less rigorous impact evaluation approach was chosen for practical reasons, a more careful impact evaluation method would bring important benefits at relatively low cost. We demonstrate that these benefits would be large by showing that in the absence of a rigorous impact evaluation design for the MVP, the project's effects are more uncertain: we find large differences between effects estimated by a naive before-versus-after evaluation method compared with those estimated using a better but still inadequate differences-in-differences approach. We demonstrate the relatively low cost by showing that a more rigorous approach would not require a disproportionate amount of money or time in this case. We see value in an extensive consideration of this type for a high-profile case; clearly the results of a similar analysis of another project could be very different. The MVP offers an important chance to learn that should not be missed.

2. The costs and benefits of careful impact measurement

Over the past two decades, several influential researchers in empirical economics, and in development economics in particular, have pressed for more rigorous evaluation of the effects of policy interventions. As we use the terms here, 'rigorous' or 'careful' impact evaluation is the measurement of a policy's effect with great attention to scientifically distinguishing true causal relationships from correlations that may or may not reflect causal relationships, using 'well-controlled comparisons and/or natural quasi-experiments' (Angrist and Pischke 2009, p. xii). We are concerned here exclusively with *impact* evaluation and do not discuss forms of evaluation such as qualitative process evaluation, which have different exigencies and methods.

These impact evaluation methods put a heavy burden on researchers to prove that what they call the 'effect', 'result', or 'impact' of the policy truly represents the difference between what happened with the policy and what would have happened if the policy had not been applied. The 'impact' of a project compares what happened with that project compared with a plausible estimate of what would have happened without that project – known as the counterfactual. The counterfactual for an impact evaluation of the MVP as a project includes many interventions similar to those made by the MVP – such as building

schools and vaccinating children – that would have been carried out by governments, civil society, international agencies, and local communities themselves if the MVP did not exist.

Evaluating the effects of a teacher training programme on students' test scores could be done, for example, simply by comparing test scores before and after the programme. Such an evaluation could yield useful information to programme implementers. But it could not be considered a rigorous evaluation of the effects of the programme if there are good reasons to believe that teacher training might have occurred by other means even without the programme, that scores might have changed even without the programme, or that the programme was applied selectively to classes likely to show improvement in scores. The more rigorous the impact evaluation design, the fewer assumptions needed to interpret the resulting estimates as credible estimates of effects. In the school example, impact evaluation by comparing with classes that did not receive the programme requires the assumption that those classes are equivalent in all other ways to classes that did receive the intervention. A more rigorous impact evaluation design gives strong evidence for this equivalence rather than assuming it.

Economists are now engaged in a spirited debate about the rising use of such methods, particularly in development research. Pointed exchanges on this topic fill the spring 2010 issue of the *Journal of Economic Perspectives* and the June 2010 issue of the *Journal of Economic Literature*, both flagship publications of the American Economic Association. Proponents of the broader use of rigorous impact evaluation methods argue that they increase the policy relevance of research and the transparency of conclusions about the true effects of policies, improving the incentives for development practitioners (Duflo *et al.* 2008, Angrist and Pischke 2010, Imbens 2010). Advocates of less emphasis on rigorous methods argue that high demands for rigour are frequently misplaced. Some argue that rigorous methods often focus inordinate attention on easily quantifiable outcomes, miss important information about the heterogeneity of impacts, take too long to yield results, miss long-run equilibrium effects, and divert scarce resources away from interventions already known to work well (Woolcock 2009, Acemoğlu 2010, Deaton 2010).

Neither of these groups argues for or against the use of rigorous impact evaluation in all cases. Rather, they differ on how important it is to use some methods in particular settings. Most participants in the debate would probably agree that, like any other methods, the more rigorous methods in an impact evaluator's toolkit should be deployed when their net benefits are high relative to the alternatives. This is likely to happen when: the cost of collecting highly rigorous evidence is relatively low; a policy decision is not needed more quickly than rigorous analysis can be performed; the consequences of applying the wrong policy are particularly bad; resource constraints make it impossible for everyone to receive the policy at once; strong interests on different sides of a policy decision necessitate an objective criterion of 'success'; and the carefully evaluated setting is similar enough to the scaled-up intervention that external validity is highly plausible.

To make this abstract discussion concrete, here we discuss the use of rigorous impact evaluation methods in one specific and important policy intervention. We argue that in this setting the benefits of additional rigour in impact evaluation substantially exceed the costs. We begin by putting that intervention in historical context.

3. Past model villages and the Millennium Villages Project

The core purpose of making impact evaluation rigorous is persuasion: changing people's minds about what is possible. Changing people's minds requires taking their prior beliefs into account. The need for rigour in impact evaluation thus depends on the level and

variance of prior beliefs among the audience for the evaluation. No case for greater rigour in impact evaluation can be made, then, without a consideration of reasonable priors about the impact of the intervention. What are reasonable prior beliefs about whether or not the MVP can achieve its stated goals?

Development agencies and governments have created numerous village-level package development interventions around the world over the past few decades. Such model village projects seek to demonstrate that a combination of intensive interventions can lastingly improve living standards in a rural area of a developing country, usually with a view toward scaling up the intervention across large areas. While the MVP differs in important respects from the most prominent cases, the generally poor track record of past model village interventions makes it critical to rigorously measure the effects of new projects in this genre.

3.1. Model village package interventions

The particular package of interventions has varied across model village projects. Among the interventions applied have been improvements to local infrastructure such as roads, electricity networks, and communications; improvements to agricultural technology such as provision of fertiliser and new seed varieties; improvements to human capital such as the construction and staffing of schools; improvements to public health such as the creation of free clinics and improvements to water and sanitation facilities; and improvements to financial access such as the introduction of microloan and microsavings instruments. The trait that defines the genre is the simultaneous introduction of a broad package of intense and expensive interventions in a limited area of a poor, rural region over a few years, in order to spark lasting economic development once the interventions end.

Prominent examples of model village interventions include the following:

- The *mofan* (model) villages of rural China from the 1930s through the 1970s, in the Jiangxi-Fujian Soviet and throughout the Mao Zedong era (Heilmann 2008).
- The Ford Foundation's five-year model village package interventions in India, starting in the 1960s under the Intensive Agricultural District and Community Development Programs (IADPs) (Unger 2007).
- 'Integrated Rural Development' (IRD) programmes promoted by the World Bank and other development agencies starting in the 1970s, which bundled village-level interventions in agriculture, infrastructure, human capital, health, finance, and communications.
- Various planned development packages for resettlement villages in Africa from the late 1960s to early 1980s. These included Julius Nyerere's pilot schemes for 'Operation Planned Villages' across Tanzania and Mengistu Haile Mariam's related agricultural villages in Ethiopia (Scott 1998, pp. 229–252), as well as Houari Boumedienne's 430 'agricultural villages of the agrarian revolution' in Algeria (Sutton 1984).
- The Southwest Project in China, a five-year village-level package intervention executed in 1800 rural villages in the late 1990s (Chen *et al.* 2009).
- A range of experimental model villages across rural Great Britain, colourfully reviewed by Darley (2007).

Many analysts have harshly criticised these past efforts. Diamond (1983) argues that China's *mofan* villages 'show that rural development can be done, if from somewhere

there are loans, new technology, and scientific information, and new marketing alternatives for what is produced', but do 'little or nothing' to promote sustained and diversified economic development, functioning primarily as 'good places to bring foreign visitors to'. Barker and Herdt (1985, p. 244) report that dissatisfaction with India's IADP villages led the package intervention to be ended within a few years. de Janvry *et al.* (2002) find that various IRD package inventions were 'generally not successful beyond the level of pilot projects'. By one account, Tanzania's and Ethiopia's model villages succeeded primarily in creating 'an alienated, skeptical, demoralized, and uncooperative peasantry' (Scott 1998, p. 237). Sutton (1984) concludes that Algeria's agricultural villages amounted to 'technocrats "doling out" modernity' but leaving their goals of sustained economic development 'unachieved'.

One example of a project in this genre is Kenya's Second Integrated Agricultural Development Project (IADP II), initiated in 1979 with technical support and $22 million from the World Bank and the International Fund for Agricultural Development. The five-year package intervention targeted 40,000 households in villages across Kenya (including in the Nyanza region, where the first Millennium Village [MV] was to begin 25 years later). The IADP II package included fertiliser provision, microcredit, transportation infrastructure, soil conservation, reforestation, domestic water supply, irrigation, livestock and dairy development, and training programmes. Five years after IADP II ended, the World Bank flatly concluded: 'The project had no sustained impact on farmers'. The economic rate of return was 'zero or negative'. This failure was attributed to broader weaknesses of governance and administration in Kenya at the time, which were untouched by the intervention. The World Bank's post-mortem on the project portrays it as a relic of a bygone era:

> The project was formulated at a time when the Bank was supporting integrated rural development as the key to agricultural production growth. The experience of such projects has not been good and Bank operations have since veered away from this concept. (World Bank 1990, p. v)

Cabral *et al.* note the 'striking similarities between the MVP and past rural development initiatives, which, for various reasons, proved to be ineffective in *sustaining* rural development' (2006, p. 3; original emphasis). Former World Bank agricultural development expert Hans Binswanger-Mkhize (2011) describes the history of IRD in strong terms:

> The last time a major institution banked on ex-ante and planned integration for poverty reduction, it was the World Bank's approach to integrated rural development (IRD) in the form of area development programs. . . . IRD failed spectacularly and was completely discredited in the early 1990s. It lived only for 20 years, and was already shown to have failed by Uma Lele and others within 10 years of its launching. Billions and billions of dollars went into hundreds of such programs without hardly any sustainable impact on poverty or institutions anywhere. (2011)[1]

In some respects, the circumstances and form of the MVP interventions differ from those of the model villages approaches attempted under the banner of IRD and other cases outlined above. The MVP (Sanchez *et al*. 2007) describe what they say are differences between the MVP and IRD: that MVP targets are quantitative and time-bound; that IRD projects were 'based on insufficient experience with local agricultural systems'; that '5- to 10-year commitment of the MVP is longer than the 2–3 year duration of IRD projects'; the 'decentralization and devolution of authority to local government'; the fact that the pool of available development aid is much larger now; and the fact that there have been advances in

agriculture, health, and information technology. These differences might potentially offer some basis for optimism that the MVP may have better prospects than past model village programmes, and it important to recognise that past experiences may not be a reliable guide to the expected impact of the MVP. Nonetheless, the troubled history of model village programmes invites a degree of scepticism that future village-level package interventions can spark sustained development. This makes it particularly important that the impact of projects of this type be subject to rigorous evaluation.

Few model village package interventions have had their effects rigorously measured. An exception is the Southwest Project in China, which has been carefully evaluated by Chen *et al.* (2009). This five-year intervention sought to permanently reverse the fortunes of poor villages with a broad-based package including roads, piped water, power lines, upgrading schools and clinics, training of teachers and healthcare workers, microcredit, and initiatives for raising crop yields, animal husbandry, and horticulture. Chen *et al.* show that income and savings grew significantly in treated villages compared with untreated comparison villages by the end of the project. But five years after the project ended, living standards had improved just as much for the average village that had not been touched by the large, costly intervention. Ignoring the trend since baseline in the comparison villages would result in a large overstatement of the intervention's effects.

3.2. The Millennium Villages Project

The most recent and prominent initiative in the model village tradition is the MVP.[2] The origins of the MVP lie in the Millennium Summit, the largest gathering of heads of state and heads of government in modern history, in New York City in 2000, at which 147 presidents, prime ministers, and monarchs pledged to meet eight general development targets – the Millennium Development Goals (MDGs) – such as the halving of global poverty and achieving universal primary school completion by 2015.

In 2002, the Secretary-General of the United Nations commissioned the Millennium Project to devise a global plan of action to achieve the MDGs. From this effort, Columbia University Professor Jeffrey Sachs, one of the world's most prominent economists and the Secretary-General's Special Advisor on the MDGs, spun off the MVP. The MVP is a large, experimental intervention to promote economic development in 14 clusters of small and very poor rural villages across Africa. It began in Sauri, Kenya in 2004. Today the MVP is a joint project of the United Nations Development Program, Columbia University's Earth Institute, and Millennium Promise, a non-governmental organisation founded in 2005.

The MVP deploys a broad package of interventions in each village, including distribution of fertiliser and insecticide-treated bednets, school construction, HIV testing, microfinance, electric lines, road construction, piped water and irrigation lines, and several others. The precise mix of interventions differs in each village cluster. Its goal is to break the villages out of poverty traps, and 'make the investments in human capital and infrastructure required to achieve self-sustaining economic growth. . . . Over a 5-year period, community committees and local governments build capacity to continue these initiatives and develop a solid foundation for sustainable growth' (MVP 2010a).[3] The total cost of the MVP intervention is US$150 per villager per year over five years, measured in 2009 dollars (MVP 2008a, p. 57). This is the same order of magnitude as income per capita in the treated areas; in other words, the MVP intervention is roughly the size of the entire local economy (see Appendix 1).

The project has, prior to the results of any impact evaluation, called itself 'a solution to extreme poverty' (MVP 2007a) and recommended that its intervention be massively

scaled up across Africa. For example, when the MVP was just beginning, Sachs (2005, p. 236) called for foreign aid to Kenya to increase 15-fold in order to provide $1.5 billion per year for MV interventions in that country alone. Before any evaluation had been published, the MVP (2008b) applauded plans to expand the MVP in several countries and concluded that 'The MVP has therefore created a powerful pressure to expand as a result of its notable successes'. At the same time, the MVP has received criticism similar to that of earlier model villages: Carr (2008) decries 'absence of critical thought' (p. 333) in the design of the short-term intervention and fears that 'human well-being in the Millennium Villages is . . . likely to rely on a constant flow of aid money in the foreseeable future' (p. 337).

4. Estimated effects depend critically on the evaluation method

We begin by taking a critical look at the mid-term evaluation results reported by the MVP for five village clusters. The design of the project makes it impossible to carry out a truly rigorous assessment of the project's effects. Our goal here is not to perform such an assessment, but to demonstrate the importance of careful evaluation by showing that the estimated effects of the project depend crucially on the point of comparison for the experience of the treated village clusters.

4.1. The mid-term MVP evaluation report

In June 2010, the project released its first public evaluation of the effects of the intervention on the MV (MVP 2010c). The report describes the interventions and compares baseline values with those three years into the MVP for several indicators, using surveys from five MV sites: Sauri, Kenya; Ruhiira, Uganda; Pampaida, Nigeria; Bonsaaso, Ghana; and Mwandama, Malawi. This before-and-after comparison follows the evaluation design described by the MVP three years earlier in the *Proceedings of the National Academy of Sciences*: 'project impact is assessed by rigorous before-and-after comparisons and detailed studies by sector' (Sanchez *et al.* 2007). The MVP (2010d) portrays changes over time at these sites as having been caused by the project, and describes these changes as evidence that their intervention can make 'villages across rural Africa . . . achieve the MDGs and escape the poverty trap'. The June 2010 report indicates that analysis with a set of comparison sites will be published later in 2010.

This report uses the before-and-after comparison in the MV to attribute changes in the villages to the effects of the project. The report lists for each site a series of 'biggest impacts,' such as 'Proportion of households that own a mobile phone increased fourfold' in Sauri, Kenya (MVP 2010c, p. 75). Changes in skilled birth attendance over time are called 'effects' of the project (MVP 2010c, p. 5). Other changes in the villages over time are labelled as 'quick wins' (MVP 2010c, p. 3). The report states that further research, to be published later, will allow 'more definitive statements' about 'whether the observed changes were due to the MVP intervention package or were instead a consequence of secular change' (MVP 2010c, p. 102). But even this wording suggests that the mid-term evaluation report is a statement about the effects of the project.

In the mid-term MVP evaluation report, the treated villages are compared with the same villages before the intervention. This has the advantage of simplicity but the major disadvantage of leaving unknown what might have happened in the villages if the project had not occurred. The before-versus-after evaluation approach requires the very strong assumption that the indicators of interest would not have changed in the absence of the

MVP. It attributes any observed changes to the intervention, when in fact some or all of those changes might have occurred in the absence of the MVP.[4] Without both baseline and post-treatment information on a credible comparison group, it is impossible to know whether this is true.[5]

The alternative method we explore here is simple: we compare trends in development indicators for each of three MV with trends in the same indicators for the same country overall, rural regions of the same country, and rural areas of the province or region where the MV is located. We use this approach because changes in the comparison areas – in particular, in the rural area of an MVP site's province or region – constitute a plausible estimate of the counterfactual; that is, what would have happened at the MVP site in the absence of the MVP. This does not constitute a rigorous measurement of the MVP's effects on the treated villages because there could be pre-existing differences between the people in the treated villages and the people in the comparator regions. Nonetheless, because it contains information about people who live in places that did not receive the intervention, it is more informative about the effects of the project than simply comparing the treated villages before and after the treatment, as the mid-term MVP evaluation does.

We conduct this analysis for three of the initial 14 village clusters. We selected these cases with two criteria: each village cluster must be covered in the MVP (2010c) mid-term evaluation of June 2010, which reports before-and-after data for five village clusters, and must be located in a country for which publicly-available Demographic and Health Survey (DHS) data can be used to establish broad trends between two recent time points in most of the development indicators reported in the MVP mid-term evaluation. We use DHS data because, compared with other data sources, they are high frequency (repeated every few years), highly standardised and comparable across countries, highly comparable with the MVP data, and publicly accessible.

This yields three village clusters: Sauri in Kenya, Bonsaaso in Ghana, and Pampaida in Nigeria.[6] The other two village clusters covered by the MVP mid-term evaluation – Ruhiira in Uganda and Mwandama in Malawi – are in countries where, at the time of writing, data from only one recent DHS survey are publicly available.[7]

We compare trends inside and outside the treated villages for two classes of the MVP's indicators, which we call 'inputs' and 'outputs'. We make this separation because some indicators might naturally be expected to be more responsive in the short run than other indicators. As used here, an 'input' is an indicator whose value could be changed overnight with sufficient effort by people who do not live in the village. For example, everyone in any village could be given access to a ventilated improved pit latrine in a matter of hours if a sufficient number of outsiders descended upon a village to install such latrines. An 'output' is an indicator whose value does not depend entirely on the intensity of effort by outsiders and depends on a range of complex decisions by villagers, such as school attendance or child stunting.

The results are presented in two formats. For three selected indicators, three figures show the trends by country for the MV, for the country overall, for rural areas of the same country, and for rural areas of the province or region where the MV is located (Figures 1–3).[8] For each country, we also present the complete results in a single table, showing standard errors for the DHS-based statistics. The tables also indicate the trend for each indicator in the MV, and the difference between this trend and the trend in the surrounding area (Tables 1–3).[9] In the figures we assume a linear trend in all indicators. This assumption is not highly problematic, given that under the most straightforward hypothetical deviation from linearity – exponential growth – the assumption of linearity provides a conservatively optimistic view of the intervention's impact.[10] Finally, for all indicators

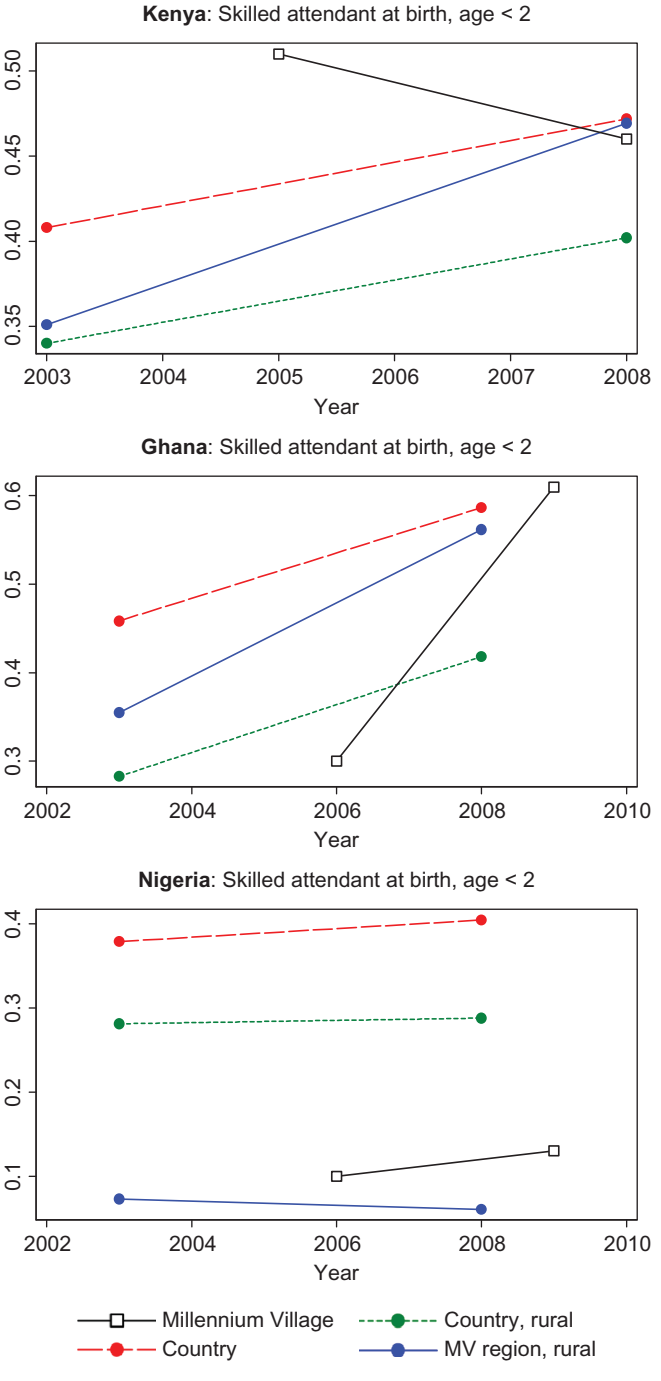

Figure 1. Fraction of live births in previous two years delivered by skilled personnel.

and all countries, we graphically compare our differences-in-differences estimates with the MVP's simple differences in Figure 4.

Here we summarise the information contained in Tables 1–3 and Figure 4.

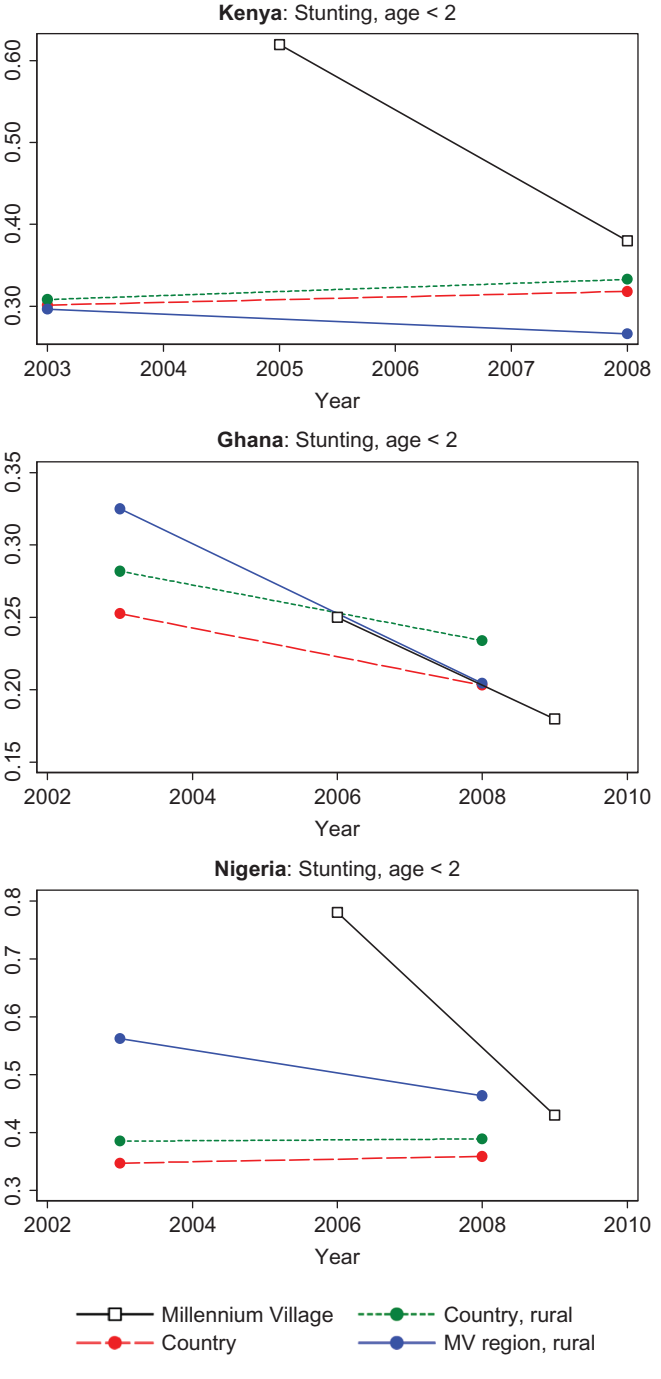

Figure 2. Stunting (chronic malnutrition) among children under age two.

4.2. Inputs: relative trends inside and outside Millennium Villages

4.2.1. Access to improved sanitation. Access to improved sanitation improved markedly across all three countries, as well as in the rural areas of the regions in which the MV in

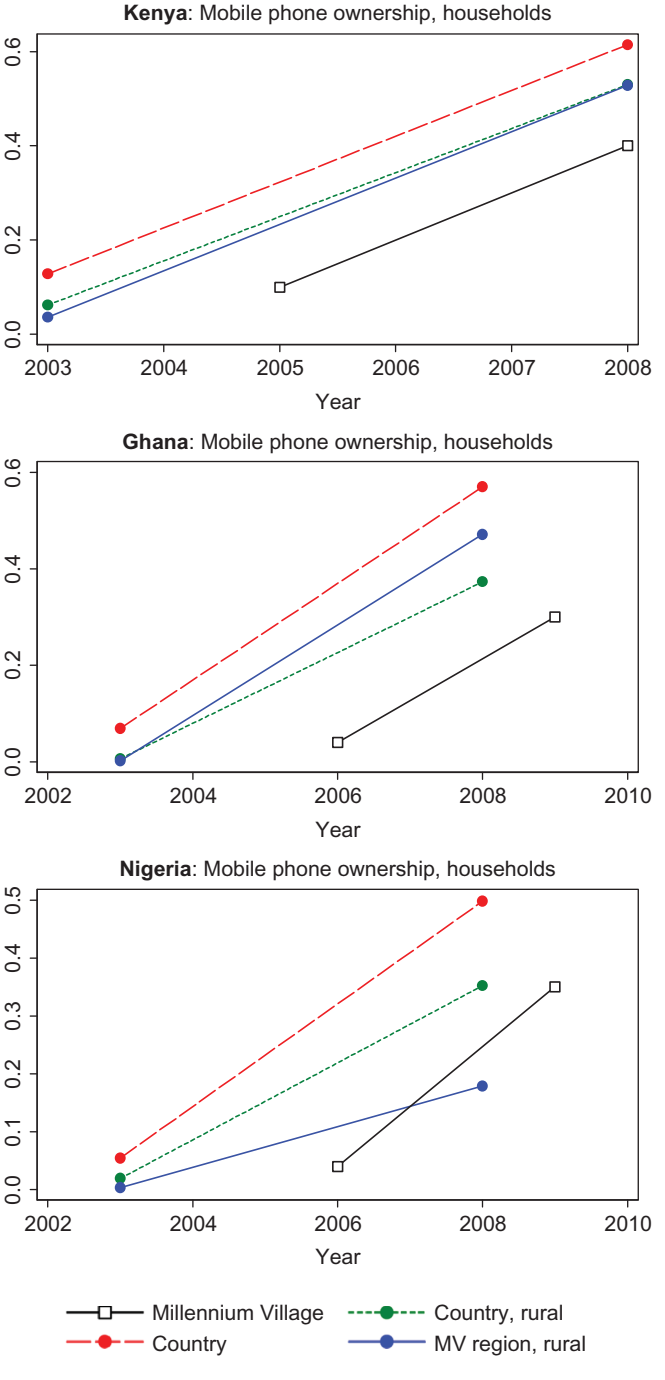

Figure 3. Fraction of households that own a mobile phone.
Note: 2003 numbers include both mobile and landline phones, and are thus an upper bound on mobile phone ownership. All other years are mobiles only.

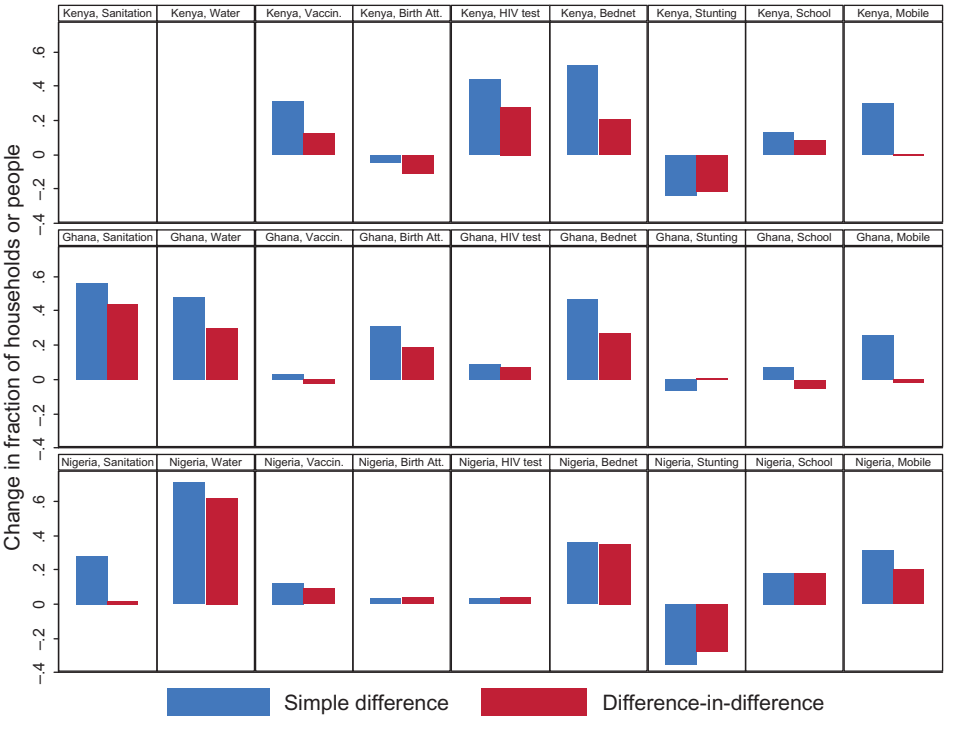

Figure 4. Summary comparison of trends within the Millennium Villages against the same trends relative to other rural households in the surrounding region.
Note: 'Simple difference' is the reported three-year change within each MV. 'Difference-in-difference' is the change in the MV relative to the change in the same indicator for average rural households in the region where the MV is located during the same period.

each country is located. In Nigeria, access at the MV intervention site rose at approximately the same rate as access in the surrounding region. In Ghana, access at the MV intervention site rose more rapidly than in the surrounding region. (MVP [2010c] does not report this statistic for the MV in Kenya.) For this and all subsequent statistics, the exact numbers and standard errors are presented in Tables 1–3.

4.2.2. Access to improved drinking water sources. Access to improved drinking water also rose notably across all three countries during this period. In Ghana and Nigeria, access at the MV intervention site rose more quickly than the surrounding area, but areas outside the intervention site also experienced substantial gains. (Again, MVP 2010c does not report this statistic for the MV in Kenya.)

4.2.3. Measles immunisation. The fraction of children immunised against measles rose in all three countries during the period. In Ghana, vaccination rates rose at the MV intervention site less than in the surrounding area. In Kenya and Nigeria, vaccination rates rose at the MV intervention site more than they rose in the surrounding area.

Table 1. Estimates and standard errors for Kenya.

Indicator	All Kenya		Rural Kenya		Millennium Village Region		Millennium Village		Simple difference within Millennium Village	Difference-in-difference
	2003	2008	2003	2008	2003	2008	2005	2008		
Percentage of households with access to improved sanitation facilities	0.37 (0.01)	0.50 (0.02)	0.28 (0.01)	0.37 (0.02)	0.20 (0.01)	0.27 (0.04)				
Percentage of households with access to an improved source of drinking water	0.53 (0.02)	0.63 (0.02)	0.45 (0.02)	0.54 (0.02)	0.38 (0.04)	0.49 (0.05)				
Percentage of children aged one who have received the measles vaccination prior to survey	0.73 (0.02)	0.85 (0.02)	0.70 (0.02)	0.83 (0.02)	0.44 (0.08)	0.75 (0.05)	0.67	0.98	0.31	0.12
Percentage of births in the past two years for which a skilled attendant was present	0.41 (0.02)	0.47 (0.02)	0.34 (0.02)	0.40 (0.02)	0.35 (0.05)	0.47 (0.04)	0.51	0.46	−0.05	−0.12
Percentage of men and women aged 15–49 who were tested for HIV in the past 12 months	0.08 (0.00)	0.27 (0.01)	0.06 (0.00)	0.25 (0.01)	0.08 (0.01)	0.35 (0.02)	0.14	0.58	0.44	0.28
Percentage of children under five who slept under an ITN bednet last night	0.06 (0.00)	0.47 (0.02)	0.05 (0.01)	0.44 (0.02)	0.08 (0.02)	0.59 (0.03)	0.10	0.62	0.52	0.21
Percentage of children under two who are two B2standard deviations below the median height for their age	0.30 (0.01)	0.32 (0.02)	0.31 (0.01)	0.33 (0.02)	0.30 (0.04)	0.27 (0.02)	0.62	0.38	−0.24	−0.22
Gross Primary Attendance Ratio	1.11 (0.01)	1.13 (0.02)	1.13 (0.01)	1.14 (0.02)	1.18 (0.02)	1.27 (0.02)	1.10	1.23	0.13	0.08
Percentage of households with at least one mobile phone	0.13 (0.01)	0.61 (0.01)	0.06 (0.01)	0.53 (0.01)	0.04 (0.01)	0.53 (0.03)	0.10	0.40	0.30	0.00

Note: Difference-in-difference is the difference between two values: the change between 2005 and 2008 in the MV; and the change between 2005 and 2008 in the rural portion of the MV region, where the value for 2005 is linearly interpolated between the 2003 and 2008 values. MVP (2010c) does not report values for access to improved drinking water and improved sanitation in Sauri, Kenya. Standard errors account for two-stage cluster sampling in DHS. Standard errors for Gross Primary Attendance Ratio (only) are bootstrapped with 1000 repetitions.

Table 2. Estimates and standard errors for Ghana.

Indicator	All Ghana		Rural Ghana		Millennium Village Region		Millennium Village		Simple difference within Millennium Village	Difference-in-difference
	2003	2008	2003	2008	2003	2008	2006	2009		
Percentage of households with access to improved sanitation facilities	0.52 (0.01)	0.71 (0.01)	0.38 (0.02)	0.56 (0.02)	0.56 (0.04)	0.76 (0.05)	0.04	0.60	0.56	0.44
Percentage of households with access to an improved source of drinking water	0.67 (0.02)	0.77 (0.01)	0.54 (0.03)	0.76 (0.02)	0.61 (0.07)	0.91 (0.04)	0.41	0.89	0.48	0.30
Percentage of children aged one who have received the measles vaccination prior to survey	0.83 (0.02)	0.90 (0.01)	0.82 (0.02)	0.88 (0.02)	0.82 (0.05)	0.91 (0.04)	0.83	0.86	0.03	−0.03
Percentage of births in the past two years for which a skilled attendant was present	0.46 (0.02)	0.59 (0.02)	0.28 (0.02)	0.42 (0.02)	0.35 (0.06)	0.56 (0.07)	0.30	0.61	0.31	0.19
Percentage of men and women aged 15–49 who were tested for HIV in the past 12 months	0.03 (0.00)	0.06 (0.00)	0.02 (0.00)	0.05 (0.00)	0.03 (0.01)	0.06 (0.01)	0.04	0.13	0.09	0.07
Percentage of children under five who slept under an ITN bednet last night	0.04 (0.00)	0.28 (0.01)	0.04 (0.01)	0.32 (0.02)	0.01 (0.01)	0.33 (0.05)	0.09	0.56	0.47	0.27
Percentage of children under two who are two standard deviations below the median height for their age	0.25 (0.01)	0.20 (0.01)	0.28 (0.01)	0.23 (0.02)	0.33 (0.04)	0.20 (0.04)	0.25	0.18	−0.07	0.00
Gross Primary Attendance Ratio	0.96 (0.01)	1.11 (0.02)	0.90 (0.02)	1.10 (0.02)	1.02 (0.04)	1.22 (0.04)	1.08	1.15	0.07	−0.05
Percentage of households with at least one mobile phone	0.07 (0.01)	0.57 (0.01)	0.01 (0.00)	0.37 (0.01)	0.00 (0.00)	0.47 (0.04)	0.04	0.30	0.26	−0.02

Note: Difference-in-difference is the difference between two values: the change between 2006 and 2009 in the MV; and the change between 2006 and 2009 in the rural portion of the MV region, where the value for 2006 is linearly interpolated between the 2003 and 2008 values, and the value for 2009 is linearly extrapolated from the 2003 and 2008 values. Standard errors in parentheses account for two-stage cluster sampling in DHS. Standard errors for Gross Primary Attendance Ratio (only) are bootstrapped with 1000 repetitions.

Table 3. Estimates and standard errors for Nigeria.

Indicator	All Nigeria		Rural Nigeria		Millennium Village Region		Millennium Village		Simple difference within Millennium Village	Difference-in-difference
	2003	2008	2003	2008	2003	2008	2006	2009		
Percentage of households with access to improved sanitation facilities	0.31 (0.02)	0.53 (0.01)	0.20 (0.02)	0.41 (0.01)	0.10 (0.01)	0.54 (0.03)	0.00	0.28	0.28	0.02
Percentage of households with access to an improved source of drinking water	0.43 (0.02)	0.56 (0.01)	0.31 (0.03)	0.45 (0.02)	0.28 (0.04)	0.44 (0.03)	0.00	0.71	0.71	0.62
Percentage of children aged one who have received the measles vaccination prior to survey	0.36 (0.03)	0.41 (0.01)	0.29 (0.03)	0.33 (0.01)	0.11 (0.03)	0.16 (0.02)	0.29	0.41	0.12	0.09
Percentage of births in the past two years for which a skilled attendant was present	0.38 (0.02)	0.40 (0.01)	0.28 (0.03)	0.29 (0.01)	0.07 (0.02)	0.06 (0.01)	0.10	0.13	0.03	0.04
Percentage of men and women aged 15–49 who were tested for HIV in the past 12 months	0.07 (0.01)	0.07 (0.00)	0.06 (0.01)	0.05 (0.00)	0.03 (0.01)	0.02 (0.00)	0.00	0.03	0.03	0.04
Percentage of children under five who slept under an ITN bednet last night	0.01 (0.00)	0.05 (0.00)	0.01 (0.01)	0.05 (0.00)	0.02 (0.01)	0.03 (0.01)	0.01	0.37	0.36	0.35
Percentage of children under two who are two standard deviations below the median height for their age	0.35 (0.02)	0.36 (0.01)	0.39 (0.02)	0.39 (0.01)	0.56 (0.03)	0.46 (0.01)	0.78	0.43	−0.35	−0.29
Gross Primary Attendance Ratio	0.91 (0.02)	0.85 (0.01)	0.86 (0.02)	0.79 (0.02)	0.52 (0.04)	0.52 (0.03)	0.81	0.99	0.18	0.18
Percentage of households with at least one mobile phone	0.05 (0.01)	0.50 (0.01)	0.02 (0.01)	0.35 (0.01)	0.00 (0.00)	0.18 (0.02)	0.04	0.35	0.31	0.20

Note: Difference-in-difference is the difference between two values: the change between 2006 and 2009 in the MV; and the change between 2006 and 2009 in the rural portion of the MV region, where the value for 2006 is linearly interpolated between the 2003 and 2008 values, and the value for 2009 is linearly extrapolated from the 2003 and 2008 values. Standard errors in parentheses account for two-stage cluster sampling in DHS. Standard errors for Gross Primary Attendance Ratio (only) are bootstrapped with 1000 repetitions.

4.2.4. Births delivered by skilled personnel. The fraction of live births in the previous two years delivered by a doctor, nurse, or community health worker rose in Kenya and Ghana, while showing little change in Nigeria. At the MV intervention site in Kenya, skilled birth attendance fell while it was rising in the surrounding region. In Ghana, skilled birth attendance rose at the MV site more than in the surrounding region. In Nigeria, it increased only slightly at the MV intervention site relative to the surrounding area, despite the construction of a new, doctor-staffed, $174,000 clinic within Pampaida locality (Boyd *et al*. 2009).

4.2.5. HIV testing. HIV testing rates rose greatly across Kenya and Ghana during this period and fell in Nigeria. In Kenya and Ghana, HIV testing at the MV intervention sites rose more than in the surrounding areas. In Nigeria, HIV testing at the MV site rose while falling in the surrounding area. Kenya faces a major HIV epidemic while HIV is present but much less prevalent in Ghana and Nigeria.

4.2.6. Bednet usage. Insecticide-treated bednet usage by small children increased enormously across Kenya and Ghana and by a small amount across Nigeria during the period. At the MV intervention sites in Kenya and Ghana, bednet usage also rose, somewhat more than in the surrounding area in Kenya, and notably more than in the surrounding area in Ghana. In Nigeria, it rose much more than in the surrounding area.

4.3. Outputs: relative trends inside and outside Millennium Villages

4.3.1. Chronic malnutrition. In all three countries, stunting rates for small children fell in the rural areas of the regions in which the Millennium Villages are located. In Ghana, stunting declined at the MV intervention site at the same rate as in the surrounding region. In Kenya and Nigeria, stunting fell much more at the MV intervention site than in the surrounding region.

4.3.2. Gross primary school attendance. There were different trends in gross primary school attendance across the three countries. In the regions where the MVs are located, attendance increased greatly in Kenya and Ghana, while dropping in Nigeria. At the MV intervention site in Kenya, attendance rose somewhat more rapidly than in the surrounding region. At Ghana's MV site, attendance rose less than in the surrounding area. Finally, at Nigeria's MV site, attendance rose while attendance in the surrounding area was nearly flat.[11]

4.3.3. Mobile phone ownership. Mobile phone ownership has skyrocketed across all three countries. In Kenya and Ghana, mobile phone ownership rose about as much at the MV intervention sites as in the surrounding areas. In Nigeria, it rose somewhat more at the MV site than in the surrounding area.

4.3.4. Malaria prevalence. MVP (2010c) documents declines in malaria prevalence among all ages, from 50 to 8 per cent in Sauri, from 28 to 13 per cent in Pampaida, and from 15 to 6 per cent in Bonsaaso. Malaria tests are not administered as part of the DHS, so definitive figures on malaria trends are not available. Thus we cannot construct a figure for malaria like the other figures. However, there is suggestive evidence that malaria rates have

dropped in Kenya overall, where the MVP site decline is greatest. First, in the DHS data, the percentage of children under five reporting fever during the past two weeks – a very rough indicator of malaria prevalence[12] – dropped from 40.6 to 23.7 per cent nationally and from 48.5 to 24.8 per cent in rural areas of Nyanza, where Sauri is located. Additionally, studies based on surveillance data in other areas of Kenya find large drops in malaria rates. O'Meara *et al.* (2008) find in Kilifi, Kenya that hospital admissions for malaria decreased from 18.43 per 1000 children in 2003 to 3.42 in 2007. Okiro *et al*. (2007) find a similar decline across coastal Kenya.

4.3.5. Maize yields. The MV sites experienced striking increases in maize productivity, from 1.9 to 5.0 tons per hectare in Sauri, from 0.8 to 3.5 tons per hectare in Pampaida, and from 2.2 to 4.5 tons per hectare in Sauri. The data needed to track maize productivity trends for comparison areas is not generally available. However, the Tegemeo Institute in Kenya has an ongoing panel survey of rural agricultural households in Kenya that makes it possible to follow increases in productivity over time. Kibaara *et al.* (2008) report that overall maize yields in the Western Lowlands region (near Sauri) were about 1.25 tons per hectare in 2007,[13] which is much lower than the 2008 yield in Sauri. This suggests that Sauri started out, before the intervention, with abnormally high maize yields for its region. Kibaara *et al.* (2008) also show that maize yields more than doubled across the Western Lowlands region, outside the MV intervention site, between 2004 and 2007 – due to roughly a doubling in fertiliser usage, high-yielding variety usage, and the number of varieties planted, all of which occurred outside the MV intervention site during the same period.

4.3.6. Child mortality. Although the MVP protocol (MVP 2010c) stipulates that under-five child mortality is the primary indicator for the evaluation, child mortality rates are not included in the mid-term evaluation report. Because they calculate these rates for five-year periods preceding the data of the survey, it is not yet possible to calculate these rates for a pure period of MVP interventions. The DHS data, however, can tell us the child mortality trends in the MVP countries. The most recent DHS surveys show large declines over 2003–2008/09: from 114.6 to 73.6 per 1000 in Kenya, from 111.2 to 80.0 per 1000 in Ghana, and from 200.7 to 156.8 per 1000 in Nigeria.[14] Any future rigorous impact evaluation of the MV intervention will need to take account of these large changes occurring around the intervention sites.

4.4. Differences-in-differences estimates compared with simple differences

In addition to reporting the numbers and standard errors behind the figures, Tables 1–3 juxtapose two simple estimates of the MVP's effects. The second-to-last column of each table shows the difference between the indicator values at three years and the baseline values, using the information presented in MVP (2010c). This before-versus-after estimate of the MVP's effects is shown as the 'simple difference within Millennium Village'. The last column of each table shows 'differences-in-differences' estimates of the MVP's effects, based on the MVP data along with DHS data. These estimates show the change in each indicator between the MV evaluation years, minus the change in each indicator in rural areas of the surrounding region between the same years. The values for rural areas of the surrounding region for the MV starting year (2005 in Kenya, 2006 in Ghana and Nigeria)

are linearly interpolated using the two DHS values. For Ghana and Nigeria, the 2009 values are linearly extrapolated from 2003–2008 trends. Figure 4 summarises these results, graphically comparing the before-versus-after (or 'simple difference') estimates compared with the differences-in-differences estimates for all three countries and all nine indicators (with the exception of the two indicators not reported in MVP 2010c for Sauri, Kenya).

There are two clear patterns in this comparison. First, in many cases the intervention sites perform somewhat better than the surrounding area. Second, the differences-in-differences estimates – which take account of trends outside the intervention sites – are frequently about half as large as the simple before-and-after differences in Kenya and Ghana. In a few cases for those two countries the simple difference shows an important improvement in the indicator while the differences-in-difference shows no relative improvement or even a relative decline. This pattern differs in Nigeria, where while in some cases the differences-in-differences are smaller than the simple differences, the reduction due to removing trends in the surrounding area is less than one-half of the magnitude of the simple difference. An exception to this pattern in Nigeria is access to improved sanitation, where the simple difference shows a large increase but the differences-in-difference shows no relative increase. Overall, the differences-in-differences estimates show substantially less improvement than the before-versus-after estimates for six out of seven cases in Kenya, eight out of nine cases in Ghana, and four out of nine cases in Nigeria.[15]

We highlight this contrast to show how important it is to provide credible comparison groups. We do not claim that either set of estimates captures the true effects of the intervention, and the differences-in-differences estimates cannot be seen as rigorous measurements of the effects of the MV intervention. There are many reasons for this. First, it could be that the selection process of the MV sites made them more likely to experience different trends than the surrounding regions (we return to this issue in the next section). Second, the trends in the surrounding regions are not based on measurements in exactly the same years as the measurements at the MV sites, and we compare the two trends assuming they are both linear. Departures from linearity would change these estimates – although not greatly, because the measurements are taken in years that are not far apart, and because many of these indicators are unlikely to exhibit large departures from linearity such as volatile up-and-down jumps from year to year. (In many cases, departures from linearity such as exponential growth would further reduce the magnitude of the increase at the intervention sites relative to the surrounding region.) Third, the MV regions as shown do include the MV themselves, and it is possible that there is some spillover effect from a MV to nearby areas.[16] This could potentially 'contaminate' rural areas of the site's region/province as a comparison group. But it is highly unlikely that the MV intervention substantially affected the measurements of change for rural areas of the entire region in which they are located, because the population of the MVs is in all cases a tiny fraction of the rural population of each region.[17] Furthermore, the MV region trends in most cases are similar to rural trends that span each nation. Despite these weaknesses, the differences-in-differences estimates presented here are better measures of the true effects than simple differences.

5. The Millennium Villages Project's currently planned impact evaluation design

The evidence in the previous section illustrates that the mid-term MVP evaluation report cannot provide good estimates of the effects of the programme's effects. In this section we consider whether the future stages of the MVP evaluation will produce rigorous impact estimates. Any impact evaluation of the project must assess the degree to which it has achieved or is likely to achieve its stated goal. The goal of the MVP is that of:

> providing immediate evidence that, by empowering communities with basic necessities and adequate resources, people in the poorest regions of rural Africa can lift themselves out of extreme poverty in five years' time and meet the Millennium Development Goals by 2015. (MVP 2010a)

Five years after the first MV intervention began, the project released details of its plans for evaluation of the project's effects.[18] The study protocol (MVP 2009) calls for assessing child mortality as the primary indicator for the programme and also details a number of secondary indicators. The core of the evaluation is a programme of data collection in one comparison village cluster for each of 10 MV sites (the protocol excludes the Sauri, Kenya and Koraro, Ethiopia sites from the evaluation plan).[19]

The protocol (MVP 2009, p. 27) mentions only briefly the mid-term evaluation report (MVP 2010c) described in Section 4, stating that '[i]nterim analyses post year 3 collection will be conducted to assess changes from baseline at MVP sites, as well as to compare levels of primary and secondary outcomes between matched intervention and comparison villages'. The mid-term report includes a subset of indicators specified in the protocol, along with additional indicators, and does not present all indicators for all sites.[20]

Beyond the mid-term evaluation report, the protocol details the plans for a final evaluation report. Data are to be collected for both MV sites and comparison villages twice: three years and five years after the beginning of interventions at the MV sites. The protocol is based on the premise that differences in various indicators between the MVs and the comparison villages are a good indicator of the MVP's effects.

This impact evaluation protocol has a number of weaknesses. The protocol document acknowledges some of these but does not fully explore their implications.

A first weakness concerns the process of the selection of the MV intervention sites. As the MV protocol notes, the 'non-random selection of intervention communities has the potential to introduce bias'. Specifically, the fact that the MV sites were not selected at random generates the possibility that they differ, in terms of both observed and unobserved characteristics, from other possible sites. The protocol explains this decision as follows:

> In the MVP context, purposive selection was undertaken to ensure a diversity of agro-ecological zones and range of operational settings that characterize the sub-Saharan African context. Communities were often among the most disadvantaged – with high levels of baseline mortality and nutrition deficiencies felt to be desirable characteristics in the initial site selection. Issues of feasibility, political buy-in, community ownership and ethics also featured prominently in village selection for participation in large scale development programs such as MVP. Finally, as the efficacy of various components of the MVP package have already been demonstrated, our goal is to assess the adequacy of the MVP model in a diversity of settings, while generating the best possible evidence of plausible and consistent relationships between intervention exposure and a range of pre-specified primary and secondary outcomes. (MVP 2009, p. 34)

It is understandable that, in the context of launching a multi-country demonstration project, concerns other than facilitating a rigorous impact evaluation came into play in the selection of the MVP sites. But the explanation in the above paragraph heightens rather than alleviates concerns that the MVP sites differed systematically from other potential sites; in particular, in that they were apparently seen to have greater levels of political buy-in and community ownership, which might well be important factors for the intervention's success. There is no indication that the project insisted on equal levels of political buy-in and community ownership in the comparison villages. This makes differences in outcomes between the treated and comparison villages harder to attribute to the effects of the intervention.

It is furthermore troubling that the protocol describes the selected intervention sites as 'often the most disadvantaged'. This means that the treatment effects measured there could be substantially higher than the effects that could be expected at other sites as the project scales up. For example, a doctor placed in a community where skilled birth attendance starts out exceptionally low – relative to other communities – can cause a larger *change* in skilled birth attendance than might be possible in a more typical community. Vaccination rates in villages where they start out exceptionally low can be raised further there than they can be raised in more typical villages.

A second weakness of the protocol concerns the process of selecting the comparison villages. The comparison village selection process is described as follows:

> The choice of candidate villages for comparison was at random, and informed by both scientific and operational issues. Scientifically, a detailed matching process is undertaken to adequately characterize village attributes. DHS data are not sufficiently disaggregated to the degree of granularity such that potential local variations in levels child mortality between MV and comparison candidates are not known in advance. The matching process therefore involves collecting data on village-level parameters with the potential to influence child mortality and related MDG outcomes including: agro-ecological zone, accessibility (distance to tar roads and markets), principal livelihood strategies, crop types, levels of electrification, numbers of clinics and schools, and the presence of other development projects in the area. A full inventory of village matching criteria is detailed in Appendix 2. This inventory is collected both for the intervention and comparison villages and updated in parallel with each round of surveys. Operationally, comparison villages had to be logistically accessible to evaluation field teams, yet far enough away from the intervention site that 'spillover' from the MVP would be unlikely. The same field teams will conduct assessments in both intervention and comparison sites to ensure a standardize approach to the evaluation, as well as to minimize and evenly distribute potential bias. (MVP 2009, p. 20)

This paragraph describes a process of selecting candidate comparison villages that permits considerable subjectivity. The 'matching criteria' comprise a handful of village traits that will be considered during matching – such as size, distance to a paved road, and agro-ecological zone. But the protocol does not indicate how closely villages will need to match, what weights the different traits will have, or how it can be known that comparison villages were not chosen using additional criteria not that do not appear on the list. Other factors that could shape the effects of the intervention in the treatment villages include the level of community organisation and willingness and ability of community leadership to work well with outside agencies. Thus even if the comparison villages were perfectly matched on the listed criteria, their usefulness as a rigorous comparison group would remain unclear.

The protocol continues to describe how comparison villages are chosen from among the matched villages:

> Using these matching criteria, up to three comparison village candidates are assessed for potential inclusion in the study. Given the range of matching criteria being assessed, and the need for comparison villages to be contiguous and within the same agro-ecological zone and farming system as the intervention site, finding three potential matches was felt to be maximum logistically feasible. Among three comparison village candidates, one will be selected at random for inclusion. (MVP 2009, p. 20)

Unfortunately, random selection among candidate villages selected in this way does not rectify the problems inherent in the selection method for the candidates. A random choice among three or fewer candidates, each of whose choice allows the possibility of subjectivity, still amounts to an overall choice that allows subjectivity. Without a clearly

objective selection procedure, the credibility of the comparison villages will always be in doubt. Rigorous evaluation methods, including fully objective matching, policy discontinuities, and randomised treatment, among others – although often criticised as methods only of interest to academic purists – offer precisely this objectivity and clarity to practitioners.

A third weakness of the MVP protocol is the lack of baseline data on comparison villages.[21] Baseline data on comparison villages reduce the possibility that the comparison villages differed from the treated villages to begin with. For example, if a treated village has higher school enrolment than a comparison village in 2009, is this because the treated village advanced more rapidly than the comparison village between 2006 and 2009, or because the treated village already had higher school enrolment in 2006? This question is difficult to answer without data on both the treated village and the comparison village at baseline. For just a single indicator – child mortality – it will be possible to generate baseline data using the data from the three-year survey, since such data are inherently retrospective.

The MV protocol does acknowledge 'that the absence of pure baseline data from comparison villages is a limitation' (2009, p. 20). But this is insufficient treatment of a very important issue. The lack of broad baseline data on comparison villages is one more reason why differences between treated and comparison villages cannot be scientifically attributed to the intervention.

The protocol describes the problem and asserts that '[f]or many indicators, in the absence of new interventions it would be unlikely that substantive year on year changes would be witnessed, such as malaria prevalence, institutional delivery rates or child nutrition'. The implicit argument is that if we assume that little changes without interventions, data for the comparison villages from the three-year point can be interpreted as roughly equivalent to baseline data. The protocol also notes that the 'presence of external interventions in comparison sites will be monitored throughout the evaluation to better understand the nature of secular changes taking place during the project period'.

There are a number of problems with this argument. First, changes may occur for hard-to-identify reasons that have nothing to do with local interventions. Economic growth, weather variation, changes in remittance flows, broad improvements in public administration, and national information campaigns are examples of factors that could change conditions in the comparison villages but not be identified as 'external interventions' by the MVP monitoring teams. Second, while changes over a single year for many indicators may typically be negligible for many indicators, over the relevant period of three years changes are more likely to be substantial. Third, if the MVP monitoring did identify interventions, it would not be possible to credibly adjust for them in the evaluation. Instead, the existence of external interventions that may have changed conditions in the comparison villages would only make it clearer why it is essential to have baseline data at those sites.[22]

A fourth weakness is the small sample size in the evaluation. There is only one comparison village cluster for each of the 10 treated village clusters being evaluated. Although 300 households in each cluster are to be surveyed, households within the same village cluster are likely to be similar to each other in many observable and unobservable ways and do not constitute independent observations of outcomes under treatment or non-treatment. Consequently, as the MVP protocol makes clear, the current evaluation will not produce reliable estimates of the effects at the level of any single intervention site. At best the MVP protocol aims to produce an estimate of the mean impact across all 10 sites.[23] The protocol outlines power calculations based on detecting changes in child mortality, noting that the

planned sample size will only be able to reliably detect a minimum drop of 40 per cent at intervention sites.[24] Achieving such a large drop in five years requires the child mortality rate to fall at more than double the rate called for in the already-ambitious MDGs.[25] Any smaller changes could not reliably be detected in a way that would allow investigators to reliably distinguish them from zero. Similarly large drops would be needed in other indicators measured as proportion (such as school attendance and access to sanitation) for them to be detected by the evaluation. In the next section using more modest assumptions of the impact, we estimate that somewhere around 20 matched pairs would be required to confidently conduct statistical inference on the project's effects in the absence of baseline data for the comparison villages.

A fifth important weakness is the short time horizon of the evaluation. The MVP (2009) evaluation protocol states that, although the project is intended to include two five-year phases, the 'evaluation is for the first five year project period' only. This is inadequate, for three reasons. First, the most important goal of the project is long term: to lastingly break the villages free from poverty traps and 'achieve self-sustaining economic growth' (MVP 2010a). Only a long-term evaluation, conducted several years after the intervention has ended, can assess whether its effects are self-sustaining. Second, earlier village-level package interventions of this type have had impacts that faded just five years after the intensive intervention ended (World Bank 1990, Chen *et al.* 2009). Third, even if the stated plan for short-term evaluation is just the first phase of a longer-term evaluation to be described in future public statements, it is still problematic. The project has stated that it intends 'the model's expansion throughout Africa' and that this is already happening in several countries (MVP 2010f). The fact that scale-up is beginning at a time when the only stated evaluation horizon is five years means that it will not be possible to determine whether the intervention is capable of doing what it claims before it is scaled up.

A further, minor weakness in the evaluation plan is the choice of primary indicator. In contrast to the programme's stated goals of reducing extreme poverty and sparking 'sustained economic development', the protocol emphasises under-five child mortality, and is subtitled 'Assessing the Impact on Child Survival in Sub-Saharan Africa'. However, as the protocol specifies, the MVP child mortality calculations require five years of retrospective data. Consequently, it was not possible to calculate child mortality rates for a pure period of MVP interventions for the mid-term evaluation, which is based on data collected three years after the start of the MVP. It will not be possible to make any estimate of the project's effect on child mortality – the protocol's primary indicator – until more than five years after the initiation of the project at each site.

6. How the true effects of the Millennium Villages Project could be measured

The analysis presented in Section 4 does not constitute a rigorous impact evaluation of the effects of the MV intervention. Due to the design of the project's initial phase, a rigorous impact evaluation cannot be performed. The calculations above serve only to concretely illustrate how important the choice of evaluation method can be.

It is too late to apply rigorous impact evaluation methods to the first phase of the MVP because of how it was set up, as Esther Duflo, a leading proponent of careful evaluation in development interventions, has observed.[26] But the effects of any future expansion of the project – that is to say, the future MV intervention sites, of which many are planned – can be rigorously evaluated at relatively low cost.

6.1. One way to rigorously measure the impacts of the MVP

A rigorous impact evaluation of the MVs can be conducted in the future. Several methods of rigorous impact evaluation are available, including careful differences-in-differences, regression discontinuity designs, randomised assignment of treatment, and others summarised by Gertler *et al*. (2010).

One of many good ways to proceed would be to randomise treatment between members of several matched pairs of village clusters, among the new intervention sites to which the MVP is planning to expand.[27] Such a design would proceed as follows. First, several pairs of village clusters would be chosen using a systematic matching procedure to resemble each other on a number of observable traits – such as population, average education, and a wealth index. This could be based on existing climate and population data, without additional surveys.[28] Second, only one cluster per pair of clusters would be selected – at random – to receive the package intervention. This eliminates the possibility of subjectivity in the choice of untreated comparison villages. Third, surveys of various development indicators would be performed for all the treated and untreated clusters, at multiple time points such as zero, five, 10, and 15 years after the intervention began. Differences between the treated and untreated clusters would be scientifically attributable to the intervention. These differences could be measured for all of the indicators in Section 4 above, plus household income, consumption, and assets.

How many matched pairs of village clusters would be required in order to achieve a given level of statistical power? This depends on many factors, including the intra-cluster and inter-cluster variance in the outcomes of interest. Because data on the MVP current and planned intervention sites are not publicly available, this calculation cannot be performed directly for the MVP sites. Here we roughly estimate the statistical power achievable with different numbers of matched pairs using the publicly available panel data for Kagera, Tanzania from the Kagera Health and Demographic Surveys (KHDS) conducted between 1991 and 1994. We use the matching method of Bruhn and McKenzie (2009) and the statistical power formula of Duflo *et al.* (2008).[29]

Figure 5 summarises the results of this example power calculation for two outcomes in the KHDS data: primary school completion and access to improved water sources. The horizontal axis shows the number of matched pairs of villages, and the vertical axis shows the minimum detectable effect of an intervention in percentage points. For example, the figure shows that a 10 percentage point change in primary school enrolment could be detected

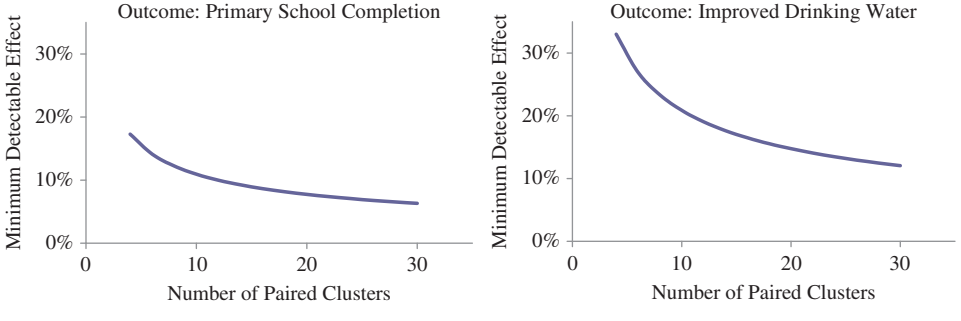

Figure 5. Sample power calculation using panel data from Kagera, Tanzania.
Note: Power = 0.90, significance = 0.05. Power calculation is for matched and paired clusters with baseline data; that is, power is calculated for changes rather than levels in the panel data of the KHDS 1991–1994. See text for details.

with 90 per cent probability, at the 5 per cent level of statistical significance, with at least 10 matched pairs of villages. This differs for different outcomes: the figure also shows that detecting a 15 percentage point change in access to improved drinking water, with the same reliability, would require more than 20 matched pairs of villages. Based on this very rough analysis we approximate that roughly 20 matched pairs of village clusters would probably be sufficient to yield reasonable statistical power for the experimental design we propose. But an exact calculation would require the MVP's internal, confidential data.

The design we propose would address the weaknesses of the current evaluation protocol we discussed above. It would allow measurement of short-run and medium-run indicators of various kinds, make the choice of treated villages clear and objective, make the choice of comparison villages clear and objective, and allow for complete baseline data on both groups. This would allow a rigorous measurement of the overall impact of the entire package, across all treatment sites, relative to no intervention.

There are several interesting questions that the design we propose cannot answer. Two of these follow, both of which also cannot be answered by the current MVP evaluation protocol, which likewise has one comparison village per treatment site. First, the design we propose would not allow evaluation of the relative effects of different components of the package. If a rigorous design using randomised treatment of village pairs revealed that the MVP reduced the child mortality rate five years after the intervention ended, for example, it would not be possible with this design to attribute the effect to any particular component of the package, such as improved sanitation. Second, the design we propose would not allow measurement of the impact of the intervention at any single site. There would only be a single pair of village clusters for each site, which would be insufficient to measure the impact at that site alone. Like the existing MVP protocol, our proposed design would aim to measure only the average effects across all of the intervention sites.

The design we propose could answer the most fundamental question about the project: whether a package intervention in a model village – across a range of intervention sites on average – truly is capable of sparking 'sustained economic development', the project's stated goal. While the protocol (MVP 2009) notes that 'the efficacy of various components of the MVP package [has] already been demonstrated', their efficacy in unleashing sustained economic development following an intensive combined package intervention has not. One reason we choose a 15-year time horizon for the evaluation is that it is a stated goal of the MVP (2007a) to provide 'immediate evidence' that 'people in the poorest regions of rural Africa can . . . meet the Millennium Development Goals'. Those goals are quantitative levels of indicators of human well-being to be achieved 15 years after the goals were set.

6.2. Cost-effectiveness of rigorous impact evaluation

A common objection to rigorous impact evaluation is that the additional information provided by rigorous methods does not justify their additional cost. This can be true in many evaluation settings. In others, however, scrutiny reveals that the direct cost of making impact evaluation rigorous is quite low. This is true in the case of the MVs for a number of reasons.

First, the costs of rigorous impact evaluation would have been similar to the costs of the evaluation the project is already doing. The changes needed to make the evaluation rigorous – per treated village – would have been mostly free. The project is already carrying out baseline surveys and multiple follow-up surveys in the treated villages, and a number

of surveys (although no baseline survey) in the non-random comparison villages. The only additional costs involved in making the impact evaluation design rigorous would be: the very low cost of identifying the matched pairs within which to randomise treatment, using rainfall and census data; and the additional cost of a baseline survey in the randomised comparison villages, which would only add one round of surveys, in one-half of the villages, to the multiple rounds of surveys the project is already planning. These additional costs would only represent a small percentage increase in the project's existing evaluation budget per treated village.

Second, the overall cost of any impact evaluation, even a rigorous one, would represent only a small percentage of the project's costs. This is important because it is not too late to rigorously evaluate the effects of future expansion of the MVs. For example, above we suggest that it would require only 20 matched pairs of village clusters to rigorously evaluate the impacts of the MVP at the new intervention sites to which it seeks to expand. Conservatively overestimating that it would cost US$100 per household per survey round to survey 300 households per cluster, four rounds of surveys in one pair of village clusters (zero, five, 10, and 15 years from start) would cost US$240,000. This is the cost per treated village. Even this conservatively high estimate is only 16 per cent of the $1.5 million total intervention cost for each treated village (MVP 2010e). This direct cost is small compared with the indirect cost of the large waste that would occur if the project was to be ineffective but was massively expanded without discovering the ineffectiveness early.

Third, researchers outside the MVP with access to the intervention sites might well be willing to conduct the impact evaluation rigorously from their own research budgets, providing the results free to the MVP. Interest in the MVP has been high among development researchers and policy-makers. Thus it is not obvious that any funds raised for the MVP intervention itself would need to be dedicated to a rigorous impact evaluation.

We recognise that fundraising is difficult and it can be easier to raise donations for interventions than for evaluation (Savedoff *et al.* 2006). But the MVP has already shown that it can raise funds for evaluation, and the additional costs of making evaluation rigorous are low in the MVP setting. It is worth making a strong case to donors for the critical importance of learning, objectivity, and careful impact evaluation.

6.3. Other common objections to rigorous evaluation

Project implementers frequently object to rigorous evaluation designs for reasons beyond cost. One common objection is that it is unethical to monitor people in untreated villages who are undergoing hardship. But the MVP does not have anywhere near the resources to intervene in every village in Africa. Thus it is literally impossible for there to be no untreated villages, and it cannot be unethical to observe what happens in untreated villages whose existence is obligatory. In the present case, at any rate, this objection is inapplicable; though the MVP has stated in the past that it is unethical to have comparison villages at all (Sanchez *et al*. 2007), its revised evaluation design already includes untreated comparison villages.

There is an extensive literature on process and impact evaluation in 'complex' interventions (for example, Campbell *et al.* 2007). Much of that research expresses legitimate concerns about rigorous designs like the one we propose: that they do not allow measurement of heterogeneous effects, focus attention on easily measurable outcomes, fail to capture long-run general equilibrium effects, and take too long to yield useful information to implementers.

But none of these objections bite in the MV setting. The project's stated goal is to unleash sustained economic development with an intensive package intervention. While it would be of interest to know exactly why some villages might respond more or less to such an intervention, it is also of paramount interest simply to know whether or not the package is effective on average over a broad range of villages, as it claims. It is true that the indicators reported capture only that which is easily measured and quantified, but it would be very difficult to argue that 'sustained economic development' is occurring in villages where none of the above quantitative indicators lastingly improve. It is true that the design we discuss cannot measure the average effect of the project in long-run equilibrium if every village in Africa was treated, but that point is a very long way off, and the marginal effect of treating additional villages for the foreseeable future is of great interest. Of course the design we discuss above could easily be used to evaluate medium-run and long-run effects as well, simply by surveying treated and untreated villages at later points in time. That said, the project has asserted from the beginning that it can break villages free of poverty traps in five years' time, so the short-run impacts are of interest also.

Finally, there is certainly time for proper impact evaluation before the MV intervention is scaled up. Many Africans have pressing needs, but they have needs for interventions that are effective, unlike so many of the wasteful and disappointing model village interventions that have recurred over the past half-century. Taking five to 15 years to acquire scientific understanding of a large intervention's effects before enormous scale-up is an appropriate investment.

7. Conclusion

Earlier we offered six criteria for the benefits of rigorous impact evaluation to exceed its costs. We have shown that evaluation methods make a big difference in how the effects of the MVP are assessed. In the case of the MVP, all six of our criteria for rigorous impact evaluation are satisfied: the cost of rigorous impact evaluation is relatively low; the policy decision of massive scale-up can and should wait until clear information about the project's effects is available; the consequences of scaling up an unproven and expensive intervention across an entire continent are high; resource constraints make it impossible for all but a few villages to receive the intervention at first; strong interests inside and outside the MVP raise the importance of clear and objective criteria to assess the project's real effects; and the evaluation setting is highly similar to the setting in which a scaled-up version of the MVP would occur. Despite this, the initial round of the MVs has been designed in such a way that truly rigorous impact evaluation cannot be done.

Without baseline data for the comparison villages, the MVP evaluation design implicitly relies on the assumption that no changes would have occurred without the MVP interventions. This assumption is clearly not compatible with the evidence we have presented. During the period of the MVP interventions, large changes were taking place in the countries where the MV sites are located – at the national level, in rural areas overall, and in rural areas of the province/region where the MV sites are located, outside the intervention sites. Kenya achieved marked gains in improved drinking water sources, improved sanitation facilities, measles vaccination, births delivered by skilled personnel, HIV testing, ITN (insecticide treated net) usage, mobile phone ownership, and child mortality rates. The incomplete evidence also suggests that across Kenya, maize yields are on a sharply upward trend and that rates of malaria have dropped. Ghana as a whole experienced improvement in drinking water and sanitation, measles vaccination, births delivered by skilled personnel,

ITN usage, child malnutrition, school enrolment, mobile phone ownership, and child mortality rates. Finally, across broad areas of Nigeria, there have been notable improvements in access to improved drinking water and sanitation facilities, rates of measles vaccination, mobile phone ownership, and child mortality.

These findings may be at odds with perceptions that all of Africa is permanently mired in poverty. The trends captured in our snapshots of trends in three countries match the broader findings of Radelet (2010), who demonstrates that the experiences of sub-Saharan African countries have been diverse and that many countries in the region have experienced steady economic growth, improved governance, and decreased poverty since the mid-1990s. Pinpointing the probable drivers of the broad changes in our limited set of indicators is beyond the scope of this paper, but we can hazard an informed guess that they are broadly the result of a combination of economic growth and improvements in public service delivery, unconnected to the MVP.[30]

Unfortunately, the information available currently and in the near future to evaluate the effects of the MVP will not allow even approximate measure of the project's impact with care. The before-and-after estimates in MVP (2010c), we have shown, can be extremely poor indicators of the impact of the project. And future evaluation work the MVP has planned, comparing selectively-chosen treated villages against opaquely-chosen comparison villages, is also unlikely to yield reliable information on the impact of the intervention.

One important reason for this is that the treated villages were chosen to be places with political buy-in and community ownership while the comparison villages were not. Schlesinger (2007) reports that Sauri, Kenya had 15 years of extensive involvement with international aid groups before the MVP arrived. And there may be other, unobserved reasons why the selection of treated villages differed from the selection of comparison villages. A second reason is that the process of choosing the comparison villages allowed for considerable subjectivity. A third reason is the lack of baseline data on the comparison villages. This means that it will be impossible to control for even the *observable* portion of differences that arise from selection of treated and comparison villages, rather than from the true effects of the project, to say nothing of other differences less easily observed and measured. A fourth reason is the small sample size of the currently-planned evaluation. Although it will measure outcomes in a large number of households, the fact that those households will be clustered together makes it fairly likely that statistically reliable statements about the overall effects of the project on various development indicators would be impossible even if none of the other weaknesses were present. A fifth reason is the short time horizon of the current evaluation protocol, which will make it impossible to assess crucial claims about the long-term effects of the intervention before the project's large planned scale-up across Africa.

Rigorous impact evaluation is essential in the MVP setting. The project costs large amounts of money – $1.5 million per village, in very poor areas, spent annually at a rate that exceeds the size of the entire local economy. Very large and lasting effects of such a sizeable intervention could be commensurate with this level of expense; more modest effects would call into question its cost-effectiveness. Unfortunately, questions of this kind will be difficult to answer for the foreseeable future.

But it is not at all too late for the design of the MVP to be changed to allow for clear and rigorous evaluation of its short-term and medium-term effects. So far there are only 14 intervention sites, and the publications of the project discuss the possibility of hundreds or even many thousands of intervention sites in the future. Effects of the intervention at the

next 20 or so sites could easily be rigorously evaluated – by the method of randomised treatment among matched pairs that we suggest above or by another equally rigorous method – at relatively low cost. The MVs offer an opportunity to learn what does and does not work in development interventions, an opportunity that will be lost if impact evaluation ceases after the first few intervention sites or simply recurs in its current form.

Notes

1. For an example of the research to which Binswanger-Mkhize refers, see Lele (1975).
2. There are numerous other model village efforts now underway around the world – including hundreds in India, such as the Kuthambakkam model village in Tamil Nadu and the Pattori model village in Bihar.
3. The project's stated objectives have focused on the five-year time horizon: 'In five years, not only will extreme poverty be wiped out, Sauri will be on a self-sustaining path to economic growth' and 'These investments, tailored to meet the needs of each community, are designed to achieve the Millennium Development Goals in 5 years' (Millennium Promise 2007); and 'The Millennium Villages project is an integrated development initiative providing immediate evidence that, by empowering communities with basic necessities and adequate resources, people in the poorest regions of rural Africa can lift themselves out of extreme poverty in five year's time . . ..' (MVP 2007a.) The calendar of MVP key activities (MVP 2010b) presents a five-year programme showing 'Outcomes' for years three, four, and five as 'Achievement of Millennium Development Goals for child mortality, education, environment, health, gender equality, maternal mortality and water'. The mid-term evaluation report (MVP 2010c, p. 2) explains that the project is conceived of as 'a ten-year initiative spanning two five-year phases', where the first phase 'focuses on achieving quick wins, especially in staple crop production and disease control, and on establishing basic systems for integrated rural development that help communities escape the poverty trap and achieve the MDGs'. The second phase will 'focus more intensively on commercializing the gains in agriculture and continuing to improve local service delivery systems in a manner that best supports local scale-up'.
4. It is worth noting that the counterfactual is not the absence of all interventions of any type, because the MVP evaluation cannot be and should not be an evaluation of all publicly-funded activity of any kind. Rather, the proper counterfactual for an impact evaluation of the MVP whatever interventions the MVP sites would have received in the absence of the MVP.
5. Fisman and Miguel (2008, pp. 202–206) raise concerns about the lack of rigour in the MVP impact evaluation design, and posit that broader national trends might be responsible for some of the changes observed in the MV intervention site at Sauri, Kenya.
6. The period consisting of the three years of the program varies by country: 2005–2008 for Sauri, and 2006–2009 for Bonsaaso and Pampaida. The DHS data are from 2003 and 2008/09 for Kenya, 2003 and 2008 for Ghana, and 2003 and 2008 for Nigeria.
7. Appendix 2 gives the definitions of all indicators used. The DHS are nationally-representative household surveys containing individual-level data on indicators of population, health, and nutrition, carried out by the Measure DHS Project in cooperation with local governments and non-governmental organisations in countries all over the world since 1984. Although they are often used to study maternal and child health, they are representative of all households – not just those with children. Comparable and standardised survey data are collected roughly every five years in many countries, and made publicly available online (http://www.measuredhs.com). The project is principally funded by the US Agency for International Development. In July 2010 the most recent publicly-available standard DHS microdata for Uganda covered the years 2000/01 and 2006, which do not overlap with the MV initial evaluation period of 2006–2009. The most recent data for Malawi covered the years 2000 and 2004, which also do not overlap with the MVP initial evaluation years. DHS surveys from Rwanda (2005 and 2007/08) overlap with the intervention period for the MV site in that country (2006–2009), but indicators for the Rwanda MV site were not published in MVP (2010c).
8. The indicators shown in the figures as 'MV region, rural' are for rural households in the region in which the MV is located. This is rural Ashanti in Ghana, rural Nyanza in Kenya, and the rural Northwest Region in Nigeria. The Nigeria DHS does not provide state-level data in 2003.

9. Standard errors are not reported in the interim MV report, so we are not able to report standard errors for the indicators for the MV sites, the trends at the MV sites, or the differences between the trends at MV sites and the surrounding areas.
10. For example, the Nigeria pane of Figure 3 shows a higher linear slope at the intervention site than in the rural area of the surrounding region, but all the points in both could hypothetically lie on roughly the same exponential growth curve.
11. The gross attendance ratio is difficult to interpret, and a high ratio is not unambiguously positive. Ratios above one (such as those for Kenya and Ghana) indicate the presence of underage and/or overage children – that is, primary school attendees outside the age range six to 11 – and an increase in grade repetition can increase the ratio.
12. For example, Ashraf *et al.* (2010) use changes in DHS data on fever occurrence as a rough proxy for changes in malaria morbidity in Zambia.
13. Kibaara *et al.* (2008, Table 8) shows a yield of 0.506 tons/acre in Western Lowlands in 2007 (that is, 1.25 tons/hectare). In 2004 it was 0.231 tons/acre.
14. The child mortality figures are for the five-year period preceding each survey.
15. We define 'substantially less improvement' as an absolute difference of more than five percentage points between the estimates by the two methods. Because MVP (2010c) does not include standard errors for the indicator estimates, we are unable to estimates confidence intervals for either the before-versus-after or the differences-in-differences estimates. As a result, we cannot determine which estimates are statistically significant. Also note that 'improvement' in the indicator means a positive change, except for stunting, for which a decline is improvement.
16. For example, Sanchez *et al*. (2007) note that a health clinic in Sauri receives some patients who are non-residents.
17. Sauri comprises 1.32 per cent of the rural population of Nyanza Province, Kenya. Bonsaaso comprises 1.59 per cent of the rural population of Ashanti Region, Ghana. Pampaida comprises 0.02 per cent of the rural population of the Northwest Region, Nigeria. This calculation assumes: 65,000 residents of Sauri, 35,000 in Bonsaaso, and 6,000 in Pampaida, as reported by the MVP (2010c); each national government's census-based regional population estimates (5,442,711 for Nyanza Province, Kenya in 2009; 4,589,377 for Ashanti Region, Ghana in 2008; and 35,786,944 for Northwest Region, Nigeria in 2006); and the percentage of people defined as living in 'rural' areas weighted by sampling weight in the most recent DHS survey data (90.2 per cent for Nyanza Province, Kenya in 2008; 48.0 per cent for Ashanti Region, Ghana in 2008; and 79.5 per cent for Northwest Region, Nigeria in 2008).
18. Millennium interventions bean in Sauri, Kenya in late 2004. The study protocol document was registered with the ClinicalTrials.gov registry site in May 2010.
19. What we describe here is the section of the protocol related to evaluation of programme effects. Separate sections of the protocol not discussed here describe a process evaluation and measurement of the programme's costs.
20. The protocol lists 10 sites for the evaluation and excludes the sites established in 2005: Sauri, Kenya and Koraro, Ethiopia. Sauri is, however, included in the mid-term evaluation report, along with four of the sites listed in the protocol. According to the calendar in both the protocol and the evaluation report, interim assessments for four other sites should have been completed by the first half of 2010. Specifically, the report includes the following indicators not mentioned in the protocol: malaria prevalence among all individuals, maize yields, percentage of children in primary school receiving school meals, primary gross attendance ratio, rates of HIV testing among men and women 15–49, and mobile phone ownership. The report excludes the following indicators specified in the protocol: wasted and underweight nutrition measures, duration of breastfeeding, age at introduction of complementary feeding, proportion of children under five with diarrhoea in past two weeks, proportion of children under five with diarrhoea in past two weeks who received oral rehydration solution, proportion of children under five treated for pneumonia, prevalence of malaria among children under five, proportion of children under five with fever in the past two weeks who receive appropriate anti-malarial treatment, proportion of pregnant women who received an HIV test, proportion of newborns receiving a postnatal check in the first week of life, survival rate to last grade of primary education, an asset-based wealth index, and the proportion of households reporting not enough food for one of the past 12 months. It is unclear why some data are not given in the report for some village sites; for example, why no data on access to improved sanitation are given for Sauri.

21. The protocol states that the overall project evaluation will adhere to Transparent Reporting of Evaluations with Nonrandomized Designs (TREND) guidelines, which are detailed in Des Jarlais *et al.* (2004). Among the information that should be reported according to the TREND guidelines are 'data on study group equivalence at baseline and statistical methods used to control for baseline differences'. But the MVP protocol does not provide for the collection of baseline data on comparison villages.
22. In describing the details of how outcomes at the MV sites will be compared with those from the comparison villages, the protocol says that '[f]or variables where no data on baseline status exist, rural sub-national data will be imputed'. It is not clear how this imputation will work, and imputed baseline figures cannot substitute for true baseline data.
23. Comparing 300 treated households at one site that are very similar to each other against 300 non-treated households at a comparison site that are very similar to each other is closer to being a study in which $n = 2$ than one in which $n = 600$. A measurement with very low n has a very large statistical confidence interval.
24. Here, 'reliably detectable' means that with a probability of 80 per cent, the difference will be detectable at the 5 per cent level of statistical significance.
25. MDG number five calls for a drop of two-thirds in child mortality between 1990 and 2015. This equals an annual decline of $1 - (1 - (2/3))^{1/25} = 4.3$ per cent per year. A drop of 40 per cent in five years, the minimum change reliably detectable by the sample size in the current MV evaluation protocol, equals an annual decline of $1 - (1 - (0.4))^{1/5} = 9.7$ per cent per year.
26. Esther Duflo is a professor of economics at the Massachusetts Institute of Technology. Parker (2010) paraphrases an email sent by Duflo to MVP creator and Columbia University professor Jeffrey Sachs in 2009 stating that, in her opinion, it was 'too late' to use rigorous evaluation methods on the existing programme, although 'the methods could be used in any later expansion'.
27. Pair-matched randomisation may not be the most efficient impact evaluation design in this setting for a fixed evaluation budget; more statistical power might be obtained from having matched triplets or matched quadruplets, of which only one receives treatment. We suggest pair-matched randomisation in this setting because it is efficient given a fixed number of treatment sites, a condition more relevant to this circumstance than a fixed evaluation budget.
28. This could be done using widely-available rainfall data plus rich information contained in the most recent census. Recent censuses are available for many countries, including Ethiopia (1994, 2007), Ghana (2000), Kenya (1999, 2009), Malawi (1998, 2008), Mali (1998, 2008), Nigeria (1991, 2006), Rwanda (2002), Senegal (2002), Tanzania (2002), and Uganda (2002). Access policies for the census microdata vary by country. For many countries that do not make their microdata publicly available, given the extremely high profile of the MVP, we believe that it is likely that access to the data sufficient to conduct the procedure we describe could be negotiated.
29. All data from the KHDS are publicly available online (http://www.edi-africa.com/research/khds/introduction.htm). We use baseline data from 1991 and follow-up data from 1994 on each household. This dataset consists of a total of 578 households across 51 clusters in Kagera. We then take the difference in outcome variable for each household from wave one to wave four. The household-level outcomes chosen are primary school completion rate among children in each household and whether the household had an improved source of drinking water, both of which are dummy variables. A household was given a value of one if any child in the household had completed primary and zero otherwise. A household received a value of zero if their source of drinking water was the lake and one otherwise. Outcomes were chosen based on: similarity to outcomes of interest in the MVP evaluation; a mean value across households that was not close to zero or one; and ease of calculation. We generate matched pairs of village clusters by the method of Bruhn and McKenzie (2009). We match on six cluster-level characteristics: number of households in each cluster, major economic activity, major religion and ethnicity, distance to nearest paved road and electrification of villages. These characteristics are chosen to be similar to the matching criteria sketched in an annex to the MVP evaluation protocol. Finally, the residuals from a regression of the change in outcome 1991–1994 on cluster-pair fixed effects are used to calculate the minimum detectable effect for a given sample size of paired clusters using the following formula (Duflo *et al.* 2008):

$$\beta_{MDE} = \frac{t_{1-\kappa} + t_{\alpha/2}}{\sqrt{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}} \sqrt{\sigma_u^2 + \sigma_e^2}$$

where J is the number of clusters, P is the proportion of clusters that are treated, α is the significance level, κ is the desired power, σ_u^2 is the variance across clusters, σ_e^2 is the variance across households in each cluster and $\rho \equiv \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2\right)$. We assume that one cluster in each pair is treated and the other is control ($P = 0.5$), there are 100 households per cluster ($n = 100$), the significance level is 5 per cent, and the probability that the minimum detectable effect can be detected at this significance level is 90 per cent.

30. Radelet (2010) argues that five fundamental changes have driven the broader turnaround in the 'emerging African countries' that he profiles: more democratic and accountable governments; more sensible economic policies; the end of the debt crisis and major changes in relationships with the international community; new technologies that are creating new opportunities for business and political accountability; and a new generation of policy-makers, activists, and business leaders.

References

Acemoğlu, D., 2010. Theory, general equilibrium, and political economy in development economics. *Journal of economic perspectives*, 24 (3), 17–32.

Angrist, J.A. and Pischke, J.-S., 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton: Princeton University Press.

Angrist, J.A. and Pischke, J.-S., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24 (2), 3–30.

Ashraf, N., Fink, G. and Weil, D.N., 2010. *Evaluating the effects of large scale health interventions in developing countries: the Zambian malaria initiative*. Cambridge, MA: National Bureau of Economic Research, Working Paper 16069. Available from: http://www.nber.org/papers/w16069 [Accessed 23 August 2011].

Barker, R. and Herdt, R.W., 1985. *The rice economy of Asia*. Washington, DC: Resources for the Future.

Binswanger-Mkhize, H., 2011. *On the wrong path: CGIAR strategy and results framework*. Blog post. Available from: http://hansvins.blogspot.com/2011/03/on-wrong-path-cgiar-strategy-and.html [Accessed 20 April 2011].

Boyd, G., Asiabuka, C.C., Medupin, A. and Osunsanya, A.B., 2009. *Mid-term assessment of the Millennium Villages Project in Nigeria at Ikaram/Ibaram in Ondo State and at Pampaida in Kaduna State*. Draft final report for the African Millennium Villages Initiative. Available from: http://erc.undp.org/evaluationadmin/downloaddocument.html?docid=3611 [Accessed 31 August 2010].

Bruhn, M. and McKenzie, D., 2009. In pursuit of balance: randomization in practice in development field experiments. *American economic journal: applied economics*, 1 (4), 200–232.

Cabral, L., Farrington, J. and Ludi, E., 2006. *The Millennium Villages Project – a new approach to ending rural poverty in Africa?* London: Overseas Development Institute, Natural Resource Perspectives 101, August.

Campbell, N.C., *et al.*, 2007. Designing and evaluating complex interventions to improve health care. *British medical journal*, 334 (7591), 455–459.

Carr, E.R., 2008. The millennium village project and African development: problems and potentials. *Progress in development studies*, 8 (4), 333–344.

Chen, S., Mu, R. and Ravallion, M., 2009. Are there lasting impacts of aid to poor areas? *Journal of public economics*, 93 (3–4), 512–528.

Darley, G., 2007. *Villages of vision: a study of strange utopias*. 2nd rev. ed. Nottingham: Five Leaves Publications.

Deaton, A., 2010. Instruments, randomization, and learning about development. *Journal of economic literature*, 48 (2), 424–455.

de Janvry, A., Murgai, R. and Sadoulet, E., 2002. Rural development and rural policy. *In*: *Handbook of agricultural economics*. Vol. 2, Part 1. Amsterdam: Elsevier, 1593–1658.

Des Jarlais, D.C., Lyles, C. and Crepaz, N., 2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *American journal of public health*, 94 (3), 361.

Diamond, N., 1983. Model villages and village realities. *Modern China*, 9 (2), 163–181.

Duflo, E., Glennerster, R. and Kremer, M., 2008. Using randomization in development economics research: a toolkit. *In*: T.P. Schultz and J. Strauss, eds. *Handbook of development economics*. Vol. 4. Amsterdam: Elsevier, 3895–3962.

Fisman, R., and Miguel, E., 2008. *Economic gangsters: corruption, violence, and the poverty of nations*. Princeton, NJ: Princeton University Press.

Gertler, P.J., *et al.*, 2010. *Impact evaluation in practice*. Washington, DC: World Bank.

Heilmann, S., 2008. From local experiments to national policy: the origins of China's distinctive policy process. *China journal*, 59: 1–30.

Imbens, G.W., 2010. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of economic literature*, 48 (2), 399–423.

Kibaara, B., *et al*., 2008. *Trends in Kenyan agricultural productivity: 1997–2007*. Tegemeo Institute of Agricultural Policy and Development. Nairobi: Egerton University, Working Paper 31/2008.

Lele, U., 1975. *The design of rural development: lessons from Africa*. Washington, DC: World Bank. Available from: http://go.worldbank.org/1UOEW9NCX0 [Accessed 23 August 2011].

Millennium Promise, 2007. *Millennium Villages: a revolution is possible*. Available from: www.un.org/esa/coordination/Alliance/MPBooklet.pdf [Accessed 31 August 2010].

MVP, 2007a. *Millennium Villages: overview*. Available from: http://www.millenniumvillages.org/docs/MVInfokit_rev16.pdf [Accessed 6 April 2010].

MVP, 2007b. *Millennium promise: cluster reports, organization highlights*. Quarter 3, 2007. New York: Millennium Promise.

MVP, 2008a. *The Millennium Villages Project annual report: January 1 to December 31, 2008*. New York: The Earth Institute at Columbia University.

MVP, 2008b. Response by John W. McArthur and Jeffrey D. Sachs on behalf of the Millennium Village Project to Buse, K., E. Ludi, M. Vigneri. (2008). *Beyond the village: the transition from rural investments to national plans to reach the MDGs. Sustaining and scaling up the Millennium Villages. Synthesis Report of a formative review of the Millennium Villages Project*. London: Overseas Development Institute. Available from: http://www.odi.org.uk/resources/download/2494.pdf [Accessed 4 October 2010].

MVP, 2009. *Study protocol, integrating the delivery of health and development interventions: assessing the impact on child survival in sub-Saharan Africa*. Available from: https://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP+Evaluation+Protocol.pdf [Accessed 15 August 2010].

MVP, 2010a. *Background and history. Millennium Village Project website*. Available from: http://www.millenniumvillages.org/aboutmv/index.htm [Accessed 18 March 2010].

MVP, 2010b. *Key activities*. Available from: http://www.millenniumvillages.org/aboutmv/keyactivities.htm [Accessed 30 September, 2010].

MVP, 2010c. *Harvests of development: the Millennium Villages after three years*. New York: The Earth Institute at Columbia University.

MVP, 2010d. *Achieving the goals, progress to date*. Available from: http://www.millenniumvillages.org/progress/index.html [Accessed 20 September 2010].

MVP, 2010e. *Sustainability and cost*. Available from: http://www.millenniumvillages.org/aboutmv/mv_3.htm [Accessed 31 August 2010].

MVP, 2010f. *Scaling up*. Available from: http://www.millenniumvillages.org/progress/index.htm [Accessed 26 September 2010].

Okiro, E.A., *et al.*, 2007. The decline in paediatric malaria admissions on the coast of Kenya. *Malaria journal*, 6, 151.

O'Meara, W.P., *et al.*, 2008. Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. *The lancet*, 372 (9649), 1555–1562.

Parker, I., 2010. The poverty lab: transforming development economics, one experiment at a time. *New Yorker*, 17 May, pp. 79–89.

Radelet, S., 2010. *Emerging Africa: how 17 countries are leading the way*. Washington, DC: Center for Global Development.

Sachs, J.D., 2005. *The end of poverty: economic possibilities for our time*. New York: The Penguin Press.

Sanchez, P., *et al.*, 2007. The African Millennium Villages. *Proceedings of the National Academy of Sciences*, 104 (43), 16775–16780.
Savedoff, W., Levine, R. and Birdsall, N., 2006. *When will we ever learn? Improving lives through impact evaluation*. Washington, DC: Center for Global Development.
Schlesinger, V., 2007. The continuation of poverty: rebranding foreign aid in Kenya. *Harper's magazine*, May, 58–66.
Scott, J.C., 1998. *Seeing like a state: how certain schemes to improve the human condition have failed*. New Haven: Yale University Press.
Sutton, K., 1984. Algeria's socialist villages: a reassessment. *Journal of modern African studies*, 22 (2), 223–248.
Unger, C.R., 2007. Modernization à la mode: West German and American development plans for the Third World. *GHI bulletin*, 40, 143–159.
Woolcock, M., 2009. Toward a plurality of methods in project evaluation: a contextualized approach to understanding impact trajectories and efficacy. *Journal of development effectiveness*, 1 (1), 1–14
World Bank, 1990. *Project completion report. Kenya. Second Integrated Agricultural Development Project*. Washington, DC: World Bank Africa Regional Office, Credit 959-KE/IFAD loan 25-KE, May 11.

Appendix 1. Size of the MVP intervention relative to the size of the local economy

The MV intervention costs roughly US$150 per resident of the MV clusters, in 2009 dollars (MVP 2008a, p. 57). This is roughly the annual average income per capita of residents of those clusters, which means that the MVP intervention is roughly as large as the entire local economy of the intervened sites. Here we show this with a rough calculation using one of the Millennium Villages for which publicly-available data allow the calculation: Mwandama, Malawi.

In Mwandama, income per capita is somewhere between US$100 and US$150 (exchange rate dollars), and probably closer to US$100. We estimate this with two different methods:

(1) About 90 per cent of people in the Mwandama cluster live in extreme poverty (http://www.millenniumvillages.org/aboutmv/mv_mwandama.htm). The World Bank's PovCalNet figures for Malawi in 2004 indicate that 74 per cent of the national population lives below PPP$38/month, the World Bank's criterion for 'extreme poverty' (http://go.worldbank.org/NT2A1XUWP0), where PPP indicates measurement in Purchasing Power Parity dollars – the dollars necessary to purchase the same standard of living at US prices. If the poverty line referenced by the MV website is similar to that used by the World Bank, then given that the Mwandama headcount rate is higher than the national rate, this suggests that PPP$38/month (PPP$456/year) is a reasonable upper bound on the typical income of someone living there. According to the price data collected by the International Comparison Program, the PPP conversion factor to market exchange rate ratio for Malawi in the most recent year available (2006) is 0.333. Therefore US$152, converted into Kwacha at market exchange rates, will give roughly the same living standard at Malawian prices as US$456 would give at US prices. So the typical Kwacha income per capita of a Mwandama resident could be purchased with less than US$152 in cash, perhaps substantially less because the income figure is an upper bound.
(2) Unskilled wages in Mwandama are around US$0.50/day; that is, US$156/year with a six-day workweek (MVP 2007b, p. 20), and of course per-capita income

would include many non-wage-earning dependents, suggesting that average income per capita is closer to US$100 or less. This realistically assumes that few in Mwandama are high-skill workers.

Thus the US$150 per year spent on the MVP per resident of the Mwandama cluster could purchase an amount of Kwacha on the order of 100 per cent of typical local income per capita, and possibly more. Should we exclude administrative costs from this figure? The MVP's 2008 *Annual Report* states that about one-sixth of the project cost is spent on 'logistical and operational costs associated with implementation, community training, and monitoring and evaluation'. So even if we exclude that amount, the big story is the same. And it is not clear that items like 'community training' should be excluded from the amount if the goal is to compare the size of the intervention with the size of the local economy.

Appendix 2. Data and definitions

This annex describes the precise definition of variables used in our analysis of the DHS and compares them with the definitions in MVP (2010c). For some variables, there are slight differences in the definitions. In each of those cases, it is likely that the variable described by the MV definition is highly correlated with the variable described by the DHS definition. Consequently, although there may be resulting differences in the levels of the two variables, it is likely that the trends – the principal focus of the analysis in this paper – are very similar. Because in the microdata analysis the definitions were designed to maximise comparability to the MV measures, in some cases the DHS definitions differ from those of variables for which figures are published in DHS reports. These cases are noted below.

- **Access to improved drinking water sources.** *MV definition*: The fraction of households using improved drinking water sources, such as borehole or protected spring, that protects water from outside contamination. *DHS definition*: The fraction of households using improved drinking water sources. In the 2008/09 surveys these are piped water into dwelling, piped water into yard/plot, public tap/standpipe, tube well or borehole, protected well, protected spring, and rainwater. In the 2003 surveys these are piped water into dwelling, piped water into compound/plot, public tap, covered well in compound/plot, covered public well, and rainwater. The 2003 survey data do not distinguish between protected spring (improved) and unprotected springs (unimproved). We assume that the increase in access to springs consisted entirely of unprotected springs. This definition is conservative in that it is likely to understate the increase in access to improved drinking water in the DHS figures.
- **Access to improved sanitation.** *MV definition*: The fraction of households who use improved sanitation facilities, such as pit latrine with concrete slab, that hygienically separates human excreta from human contact. *DHS definition*: The fraction of households who use improved sanitation facilities. In the 2008/09 surveys these are a flush toilet, ventilated improved latrine, and pit latrine with slab. In the 2003 surveys these are a flush toilet or ventilated improved latrine. The 2003 survey does not distinguish between pit latrines with and without slab. We assume that the increase in pit latrines use consisted entirely of pit latrines without slab, which are classified as 'unimproved'. This definition is conservative in that it is likely to understate the increase in access to improved sanitation in the DHS figures.

- **Measles immunisation.** *MV definition*: Fraction of children under one year of age who have received at least one dose of a measles vaccine. *DHS definition*: Fraction of children aged 12–23 months who received a dose of a measles vaccine at any time before the survey. The World Health Organization recommends measles vaccination at the age of nine months or later (http://www.emro.who.int/vpi/measles).
- **Births delivered by skilled personnel.** *MV definition*: The fraction of births among children under two years of age who were attended by doctors, nurses, or midwives. *DHS definition*: Identical. Note: Published figures in the DHS reports are for live births during the five years previous to the survey.
- **HIV testing.** *MV definition*: The fraction of women and men ages 15–49 who received a human immunodeficiency virus (HIV) test within the past year. *DHS definition*: Identical. Note: Published figures in the DHS reports are for fraction of women and men ages 15–49 who received *results* from a HIV test within the past year.
- **Bednet usage.** *MV definition*: Fraction of children aged zero to 59 months who slept under an insecticide treated mosquito net the night prior to the survey. *DHS definition*: Identical. Note: Estimates using the 2003 microdata differ slightly from those in the corresponding DHS reports.
- **Chronic malnutrition.** *MV definition*: Proportion of children under two years of age whose height for age is less than minus two standard deviations from the median for the international reference population ages zero to 23 months. *DHS definition*: Identical, but possibly different reference population. Note: Figures shown in DHS reports are for children under five years of age. It is unclear whether the MV definition employs the NCHS or WHO reference populations. We employ the more recent WHO standard.
- **Gross primary school attendance.** *MV definition*: The number of pupils attending primary school, regardless of age, expressed as a fraction of the population of official primary school age. *DHS definition*: Identical. Note: Figures published in DHS reports restrict the population counted in the numerator of this calculation to those age six to 24. Calculated figures using the 2003 microdata and for Nigeria differ slightly from those in the corresponding DHS reports.
- **Mobile phone ownership.** *MV definition*: The fraction of households who own a mobile phone. *DHS definition*: The fraction of households who own a mobile phone (2008/09) or the fraction of households who own a mobile or landline phone (2003). Note: The DHS 2003 definition was determined by the question asked in the surveys that year. The effect of this difference is that our DHS figures understate the true growth rate of mobile phone ownership.
- **Under-five child mortality.** *MV definition*: This is a synthetic measure calculated with five years of retrospective data, using a procedure detailed in MVP (2010c). *DHS definition*: Identical.
- **The Kagera Health and Demographic Surveys (KHDS).** All waves of the KHDS panel data are downloadable at http://www.edi-africa.com/research/khds/introduction.htm.