The Millennium Villages Project: A protocol for the final evaluation

Shira Mitchell, Andrew Gelman, Rebecca Ross, Uyen Kim Huynh, Lucy McClellan, Matthew Harris, Sehrish Bari, Joyce Chen, Seth Ohemeng-Dapaah, Patricia Namakula, Sonia Ehrlich Sachs, Cheryl Palm, Jeffrey D Sachs

Abstract

The Millennium Villages Project is a ten-year integrated rural development project implemented in ten sub-Saharan African sites. This protocol documents the final evaluation of the project, including its basic principles, site selection, and a five-part evaluation plan. The evaluation will include an adequacy assessment, impact evaluation, cost assessment, process evaluation, and description of systems design and tools. We describe data collection and analysis for each component, as well as our plan for transparency and study limitations. Taken together, this evaluation is designed to assess the Millennium Village Project's model for achieving the Millennium Development Goals in rural sub-Saharan Africa.

Contents

1 Background

In September 2000, world leaders came together at the UN Millennium Summit to adopt the Millennium Declaration, which committed their nations to a new global partnership to reduce extreme poverty and set targets with a deadline of 2015 that have become known as the Millennium Development Goals (MDGs).[1]

The UN Millennium Project, an independent advisory effort from 2002-2006 initiated by UN Secretary-General Kofi Annan and directed by Professor Jeffrey D. Sachs, identified steps designed to achieve the MDGs.[2] The project recommended investments in scientifically-driven interventions, in the context of open, well-governed, and market-based economies.

The Millennium Villages Project (MVP) was initiated in 2005 to implement the UN Millennium Project's recommended interventions across multiple sectors in rural Africa. The MVP included general design of interventions, systems to measure and track intervention activities and outcomes, and programs for scale up of effective delivery systems.[3] The MVP was piloted in Sauri, Kenya and Koraro, Ethiopia in 2005, and expanded in 2006 to include fourteen sites across ten countries covering roughly half-a-million inhabitants in total. The project's goal was to facilitate rural populations to achieve the MDGs and move communities toward self-sustaining economic growth.

2 Project Description

The MVP model for achieving the MDGs in rural, sub-Saharan Africa adheres to several core principles:

- The implementation of multi-sectoral and integrated interventions grounded in well-managed delivery systems;
- The implementation of scientifically-driven technologies and practices;
- The participation of local communities in the planning, execution, and monitoring of a set of interventions, localized to the conditions of each Millennium Village (MV) site;
- Co-planning and implementing the MV concept at the local and district level with government agencies;

- Cost-sharing with government, donors, and the local community; and
- Learning by doing (adaptive co-management) in the design and implementation of interventions and systems.

The MVP approach includes interventions in multiple sectors appropriate for extremely poor, rural areas, including food production, nutrition, education, health services, roads, energy, communications, water supply and sanitation, enterprise diversification, environmental management, and business development. The MVP delivers diverse, simultaneous interventions, both to address multiple MDG objectives and also to enable possible synergistic gains through positively interacting interventions, motivated by the idea that the whole may be greater than the sum of its parts.[4, 5, 6] The MVP uses technologies and techniques such as agroforestry, improved cookstoves, insecticide-treated bednets, antiretroviral drugs, community deworming, remote sensing, and geographic information systems.

The MVP is a ten-year project with two five-year phases. The first phase concentrated mostly on "quick-win" interventions, which include:

- Free mass distribution of insecticide-treated bednets and effective antimalarial medications;
- Elimination of user fees for primary schools and essential health services;
- Expansion of school meals programs;
- Construction of roads and other infrastructure; and
- Subsidized fertilizers, improved crop varieties and tree seeds/seedling to replenish soil nutrients to smallholder farmers on recuperated degraded lands.

At the end of this initial phase, the MVP evolved to also focus on systems design and long-term sustainability. This included infrastructure systems (such as microgrids), agribusiness (including farmer-based organizations), and health and education systems.

2.1 Millennium Village study site selection

The project grew organically at the start, beginning as a single site in Sauri, Kenya and by the end of 2006 had expanded to a total of fourteen sites in ten countries. Candidate countries were selected based on essential characteristics: reasonable peace and stability, good governance and accountability, and commitment of the government to the Millennium Development Goals (MDGs). The guidelines for selection of Millennium Village (MV) sites within countries were as follows:

- located in areas of severe chronic malnutrition,
- located in varied agroecological zones in sub-Saharan Africa,
- and recommended by expert committees, including government officials.

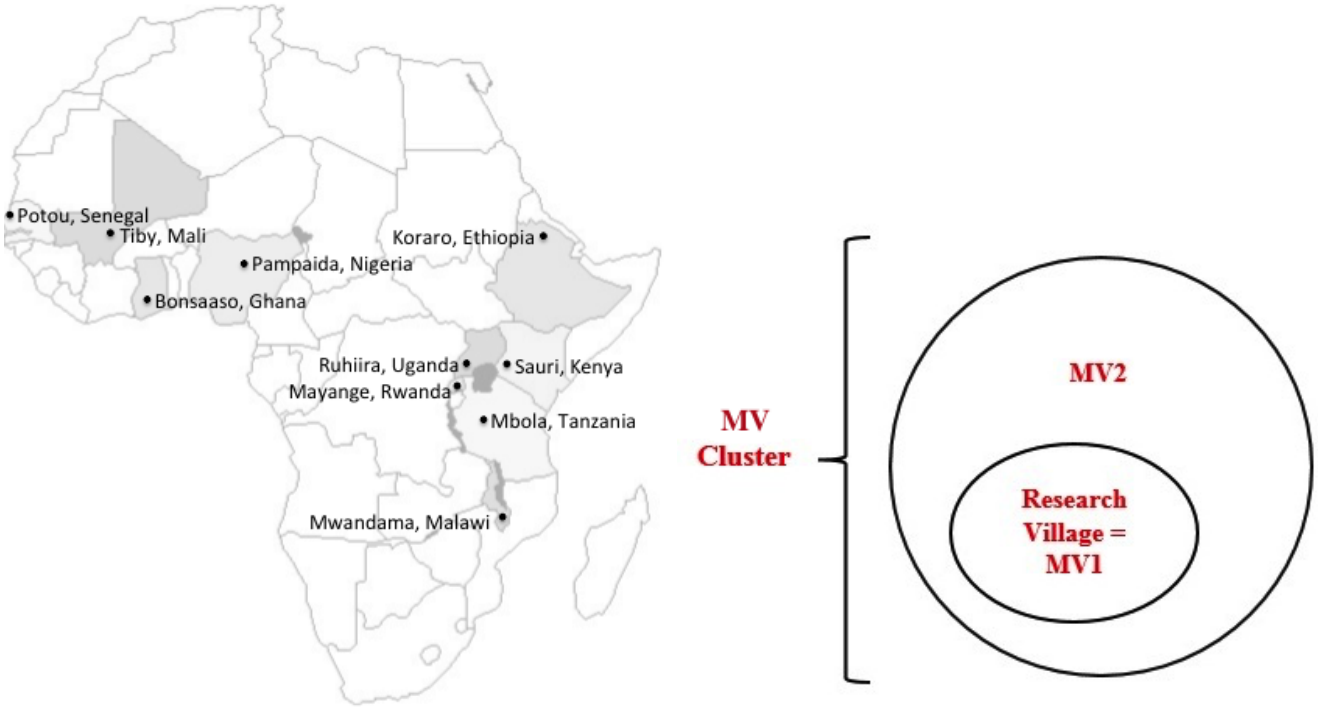
Of the fourteen MV sites, the project was able to fund ten of these to scale up to approximately 25,000 inhabitants for up to ten years, see Table 1. Scale up in these sites was intended to demonstrate to policymakers that the project could be implemented at a low cost per capita across large populations in rural sub-Saharan Africa.

In these ten sites, project resources were concentrated in a core area, referred to as the *MV1*. As additional resources became available, the MVP interventions extended to surrounding villages, referred to as the *MV2*. Taken together, the MV1 and the MV2 constitute the MV cluster, or equivalently the MV site (see Figure 1b). The ten clusters have an average population of approximately 45,000 inhabitants, ranging from 25,000 to 80,000. The MV1s have an average of roughly 6000 inhabitants (roughly 1000 households).

See Table 1 for a more detailed description of the ten clusters.

Four sites were not scaled up for various reasons and therefore did not attain the economies of scale needed to sustain full operations: Toya, Mali; Ikaram, Nigeria; Dertu, Kenya; and Gumulira, Malawi. Ikaram's United Nations Development Programme (UNDP) funding ended in 2010 and was therefore discontinued as an MV site. Gumulira began with funding from a donor grant and was taken over by another non-governmental organization (NGO) in 2011. Dertu and Toya remained small and were caught in regional conflict.

Due to disruptions in some sites, this evaluation is restricted to the ten scaled up sites, listed above. One can consider our estimates of program impact as applying to the population of village-clusters in rural, sub-Saharan Africa in which scale-up to 25,000 inhabitants would have been made by the MVP (after project start) and in which regional violence did not disrupt the project.



(a) Locations of the Millennium Village clusters.

(b) Project resources were concentrated in a core area: the MV1. As additional resources became available, the Millennium Villages Project interventions extended to surrounding villages: the MV2. Taken together, the MV1 and the MV2 constitute the Millennium Village cluster.

Figure 1: Millennium Village cluster locations and structure.

3 Evaluation Questions and Components

While specific interventions within the MVP package have been shown to be effective in environments similar to MVP,[7, 8, 9] the package of interventions and its implementation elicit questions of interest:

1) Are the Millennium Development Goals and Millennium Village Project targets met within each core intervention area (MV1)?

Millennium Village Clusters in geographic order (West to East and North to South)	Number of villages in MV1	Agroecological Zone	Start date	Number of Households in MV1 (population)	Number of Households in MV1+MV2 (population)
Potou, Senegal	14	Agro-silvopastoral	Q1 2006	717 (7227)	3137 (32,823)
Tiby, Mali	8	Agro-silvopastoral	Q1 2006	986 (14,290)	5529 (80,131)
Bonsaaso, Ghana	11	Tree Crop	Q3 2006	1201 (6049)	5555 (25,257)
Pampaida, Nigeria	28	Cereal-Root (Sudan savanna)	Q2 2006	924 (6244)	4152 (28,057)
Koraro, Ethiopia	3	Highland Mixed	Q1 2005	1171 (5914)	16,620 (67,711)
Sauri, Kenya	11	Maize Mixed (bimodal)	Q1 2005	996 (5112)	13,685 (67,315)
Ruhiira, Uganda	9	Highland Perennial	Q1 2006	1159 (5663)	9948 (46,570)
Mayange, Rwanda	9	Highland Perennial	Q3 2006	726 (3343)	5724 (25,710)
Mbola, Tanzania	3	Maize Mixed (unimodal)	Q2 2006	1041 (6952)	5972 (37,024)
Mwandama, Malawi	5	Cereal-Root (Southern miombo)	Q3 2006	889 (3598)	9038 (37,153)

Table 1: Description of the ten Millennium Village clusters (MVs) covered in this evaluation. The remaining four sites were not scaled up for various reasons and therefore did not attain the economies of scale needed to sustain full operations: Toya, Mali; Ikaram, Nigeria; Dertu, Kenya; and Gumulira, Malawi. The population counts in this table were collected in 2010-2012 via a detailed demography census in the MV1, and an abbreviated census listing (household counts) in the MV2 (2010-2013).

2) What are the program effects on each of the outcomes of interest within the core intervention area (MV1)? In other words, what progress towards the targets is attributable to the program?

3) Does the MVP stay within the target of $120 annual per capita cost?

4) Which factors have most hindered or best facilitated the implementation of intervention packages? What are the biggest lessons learned?

5) What new tools, including software, hardware, and systems design, have been developed by the MVP that may be adapted at low cost in similar settings?

Our evaluation aims to answer these questions using the components outlined below.

1) Adequacy assessment: to assess the adequacy of reaching the targets in the MV1s.

2) Impact evaluation: to attempt to isolate the effect of the program in the MV1s (in other words, to answer the question of causality).

3) Cost assessment: to compute all annual on-site costs of carrying out the MVP interventions and activities in each of the sites - by sector, year, stakeholder, and MV1 versus MV2 - relative to the project's $120 annual per capita cost-sharing model.

4) Process evaluation: to document and assess the factors at each site that have contributed to the MVP's relative "successes" and "failures" implementing a ten-year, multi-sectoral project with communities and local governments.

5) Description of systems design and tools: a detailed description of systems design and tools in situ, with a focus on scale up and replication.

The first two evaluation components (adequacy and impact) are restricted to the core intervention areas (the MV1s), but the remaining three components (cost, process, and systems design) study the entire Millennium Village clusters (both MV1s and MV2s). The survey work is restricted to the MV1s due to budget constraints, so we focus on estimating the program effects in the core intervention areas (the MV1s).

4 Outcomes of Interest

The primary outcomes of interest, for both the adequacy assessment and impact evaluation, are a subset of Millennium Development Goals (MDG) indicators and proxies.[10] This list includes indicators of poverty alleviation, agriculture, education, gender equality, health, environmental sustainability, and infrastructure. What we refer to as "outcomes" are a mixture of output, outcome, and impact indicators. These outcomes are listed in Appendix A.1. MDG indicators that have been excluded are listed in Appendix B. In addition to the MDG and MDG proxies, we will analyze the project on a number of indicators that are relevant to systems delivery (see Appendix A.2). We refer to these throughout as 'MVP indicators.'

5 Adequacy Assessment

Adequacy assessments require no control groups and only depend on comparison with internally established criteria.[11, 12] From its inception, the project has focused on demonstration of service delivery at real scale and reasonable cost, rather than assessment of causal impact of interventions. A demonstration can only show that reaching the MDGs was possible under these interventions, but would not show the counterfactuals, i.e. if other interventions (including no interventions) would have led to similar outcomes. In each of the MVP villages we will measure progress towards targets established by the project. Targets were set based on the following: (1) Official UN MDG targets, (2) International standards, and, where no official UN target or international standard exists, (3) goals set by the MVP sector leaders. Measurement of these indicators will take place in 2015, towards the end of the project, following approximately ten years of intervention exposure.

For each MV1, point estimates and 95% uncertainty intervals will be computed and reported for each indicator, using data collected in 2015. These intervals will be compared to the targets defined in Table 3 using a plot similar to that in Figure 2. Such plots will also be produced for the comparison villages, introduced below.

See Appendices A.1, A.2, and C for the explicit numerical targets and by whom they were established.

6 Impact Evaluation

In this section we describe the impact evaluation for the Millennium Villages Project (MVP). By *impact evaluation*, we mean a measurement of the program's effect with great attention to determining causal relationships. An impact evaluation attempts to assess whether the stated "effect," "result," "impact," or "achievement" of a program represents the difference between what happened with the program and what would have happened without that program.[6]

The MVP was not designed as a controlled experiment. This decision was justified on the basis of focus on an adequacy assessment, in addition to logistical, financial, and ethical complexities. The project is a village-cluster-level intervention, focused on learning about community-based delivery systems. Since it is not feasible to randomize people into villages, individual-level randomization was not an option. Compared to trials that are randomized to individuals or households, this creates a sample size issue.[13, 14] Based on a rough analysis, Clemens and Demombynes conclude that 20 matched pairs of village clusters would probably be sufficient to yield reasonable statistical power.[6] Their analysis assumes a model without parameters

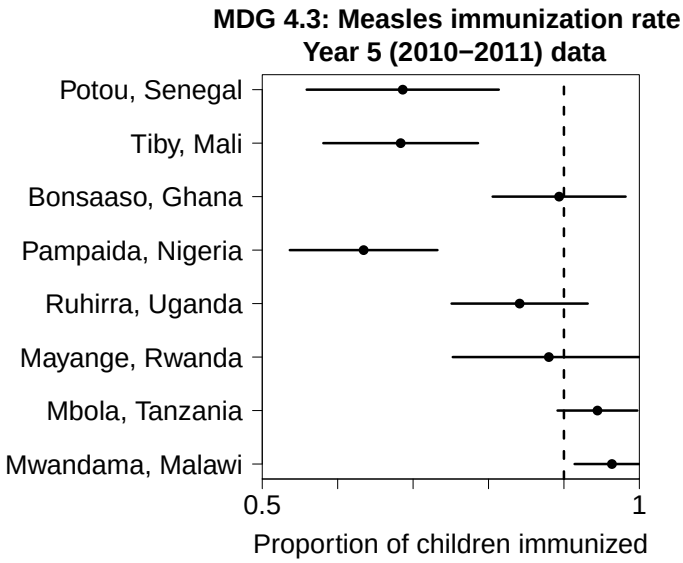


Figure 2: The above plot exemplifies the planned presentation of the adequacy assessment results. For demonstration, we show the 2010-2011 estimates of the measles immunization rate indicator within eight Millennium Village clusters, in geographic order from West to East and North to South (data are missing from two sites). We draw the target of 90% in a dashed line. The adequacy assessment will produce several plots of this kind for outcomes listed in Appendices A.1 and A.2.

that vary by country or agroecological zone. Taking this variation into account (which increases the number of parameters) would increase sample size requirements. The project's expansion from one village cluster was uncertain at the start, as it was unclear how much funding would be available. Given the project's eventual size, a design aimed to identify a causal effect could have been: at baseline, select groups of areas that match closely on geographic and poverty characteristics, and randomly assign treatment to a few areas. This would have freed us from some untestable assumptions, and would have enabled both measurement of outcomes and tracking of interventions in non-MVP areas. In this design, one ethical complexity includes how to present the study to communities not receiving the interventions.

There has been a continuing call for an impact evaluation of the MVP.[6, 15, 16, 17, 18, 19, 20] To address this gap in evaluation, in 2011 the United Kingdom Department for International Development (DFID) awarded a five-year grant to implement a new Millennium Village in northern Ghana, which is being evaluated independently by Information, Training And Development (ITAD), an international development advisory agency. In the initial design document for the impact evaluation of the Northern Ghana MV site,[21] ITAD raised many important issues, which were taken into account in the design of this evaluation.

There are many practical challenges in designing and implementing a rigorous and useful impact evaluation for the MVP. Given the sample size considerations mentioned above (see the simulation study in Mitchell et al.)[22] and the limited availability of baseline data (see below), we cannot guarantee "statistical significance." Furthermore, our results will be consistent estimates of the causal effects of the MVP only under assumptions discussed below. In the case of the MVP, consideration of these assumptions is particularly complicated and important.

Without "statistical significance," we may be able to see suggestive patterns in the results (e.g. all indicators related to a particular sector showing more impact than others). Without the assumptions

necessary for causal inference, our results will be comparisons between the Millennium Villages (MVs) and villages chosen to be similar to the MVs in 2005. Finally, we emphasize that our pursuit of quantitative estimates of causal effects represents only one component of the evaluation. A comprehensive analysis of the MVP requires understanding of local contexts and mechanisms of action. In carrying out an impact evaluation for the MVP we will be transparent about challenges, uncertainty, and assumptions, which may inform the design of future projects and evaluations.

In this section we discuss the challenges of such an evaluation, the design we have chosen, and the range of findings we anticipate. We first discuss mid-term reports, then we outline the design, and in subsequent sections go into the data sources, the matching procedure, candidate causal models, design analysis, and identification of externalities. Mitchell et al. go into the technical details of the proposed impact evaluation.[22]

6.1 Mid-term reports

Previous articles about the project's evaluation leave unanswered questions and critiques.[23, 6, 24, 25].

The project released its first public report in June 2010.[23] The report computed before-after comparisons of selected MDG indicators in the MVs. Estimates were descriptive and not an attempt at impact evaluation, but the report mistakenly used the word "impact," which understandably caused confusion. The question of interest, what impact has the project had, is answered only under the assumption that the trend in outcomes in the absence of the intervention would have been flat. Though the report did stress that the results were preliminary, it did not state this assumption as a strong caveat.

Clemens and Demombynes contrasted the reported effects with estimates from a difference-in-differences analysis.[6] They looked at three MVs, using as a comparison group the rural households in the region where the MV is located, with before and after data obtained from the Demographic and Health Surveys (DHS). Clemens and Demombynes did not provide intervals of uncertainty for the difference-in-differences estimates, and the analysis did not adjust for covariates.[6] The crucial assumption of *additivity* is needed with this strategy: in the absence of the MVP intervention, the differences in outcome over time would be the same across the MVs and comparison group.[26] This assumption can be made more believable by adjusting for covariates through matching and regression.[22]

Pronyk et al. also used difference-in-differences methods (for outcomes for which retrospective questions could provide baseline data), with adjustment for covariates via both matching and regression.[24] However, there were concerns about the usefulness of the comparison villages, due to possible differences in political buy-in between comparison villages and the MVs and the unclear selection procedures of the comparison villages.[6, 27] This evaluation will involve the selection of new comparison villages the the methodology is clearly described in detail, see below.

Wanjala and Muradian used a method related to our proposed method (see Section 6.2),[25] combining matching methods with regression estimation to look at the treatment effect in Sauri, Kenya MV. They appear to adjust for variables that may be affected by treatment, which may be a source of bias.[28] Their analysis assumes no village effects, attributing differences between the MV and comparison villages only to the treatment.[22]

The final evaluation will take into account issues and inadequacies of previous reports to provide a more rigorous and transparent analysis. Additionally, the final evaluation will be the first to consider the project in its entire ten-year context.

6.2 Design

For the impact evaluation, we consider only the MV1s, the core areas in each Millennium Village that receive the full set of interventions. As described above, these ten MV1s are located in ten distinct countries and each contains roughly 1000 households.

With unconfoundedness and noninterference assumptions (defined below), comparisons are key to estimating the causal effect of the MVP: the difference between what happened (in terms of measured outcomes) in the treatment villages with the MVP and what would have happened in them without the MVP. At end-line, in 2015, funding is available for surveying comparison villages outside of the Millennium Village clusters. Our design for the impact evaluation will be matching to select comparison villages, followed by collecting outcome data in both the MV1 treatment villages and matched comparison villages, followed by comparing outcomes using regression models. Our outcomes are the indicators defined in Appendices A.1 and A.2. For the impact evaluation we use undichotomized data.[29, 30]

The causal effect is defined in terms of "potential outcomes," - i.e. outcomes that would have happened with the MVP or without. Even with comparison data, establishing causal claims about the impact of the MVP relies on untestable assumptions, whose justifications rely on context-specific knowledge.

Our first assumption is the *stable unit treatment value assumption*, which requires that units do not interfere with one another, and for each unit there is only one 'version' of the treatment.[22, 31] This assumption can be satisfied by considering only two levels of treatment: either a unit (an individual, household, or village) is within a Millennium Village (at the scale at which the project operated, exposed to the site-specific set of interventions implemented); or a unit is far enough away from areas where the project operated that it cannot be affected by it. We aim to minimize interference (i.e. externalities, or "spillovers") by requiring our candidate comparison villages to be at least ten kilometers away from the MV cluster, outside a "buffer zone" of very likely interference, see Section 6.6. The MVs are far enough apart from each other to avoid interference among them.

A second assumption used to exploit comparisons for estimating causal effects is *unconfoundedness*, also known as no unmeasured confounders, (strong) ignorability, selection on observables, or, more broadly, a regular assignment mechanism.[32, 33, 34, 31, 26, 35, 36, 37] It states that the distribution of potential outcomes should be the same for the MVs and comparison villages, once we control for the observed confounding variables. This assumption comes "for free" in randomized experiments, but the MVP treatment was not randomized among eligible sites (see Section 2.1). Therefore, we work to make unconfoundedness as plausible as possible by controlling for many variables that are not affected by treatment.[28]

These variables are sometimes referred to as "pre-treatment" variables, but they need not be temporally before treatment, as long as we can be sure that the project could not have affected them (e.g. temperature). Adhering to the advice in the literature, we follow *matching* with *regression*.[38, 39, 26, 40, 41, 42] The combination of matching and regression is more robust than each alone.[43, 44, 36, 31] Matching serves to make the treatment and comparison groups more similar, with more overlap in covariates. This avoids using the regression to extrapolate to areas of poor overlap, which would rely heavily on the correctness of the model.

To estimate the effect of treatment on the treated villages, we want to select comparison villages that match the Millennium Villages as closely as possible on geographic, socioeconomic, education, and health variables at baseline. This strategy seeks to figure out the effect of the MVP on the Millennium Villages and

does not necessarily capture what effects the MVP would have on other villages. To inform our selection of comparison villages, and for regression adjustment, we need measures of pre-treatment variables in candidate comparison areas. In Section 6.3 we discuss available data sources. We gathered documents and correspondences from the site-selection process in order to understand the treatment assignment mechanism. These resources guided our search for relevant pre-treatment data.

In Section 6.4 we propose a matching procedure to select comparisons within the same country for each MV. After the matching procedure, we will have groups of treatment areas (the MV1s) and comparison villages. The matching algorithms search for five matches per MV1 (i.e. per country), with this sample size chosen due to a combination of logistics and design analysis. After outcome data are obtained, we will fit the models suggested in Section 6.5.

If the stable unit treatment value assumption holds, and we include enough variables to satisfy unconfoundedness, a combination of matching and regression should approximate a randomized experiment.[45, 46, 47] As previously stated, our results will be interpretable as causal effects only to the extent that the stable unit treatment value and unconfoundedness assumptions are believed. Without these assumptions, our results would be comparisons between the MV1s and similar areas (to the extent that our pre-treatment variables allow), without a causal interpretation. Each variable that we match on represents a refinement over comparing the treatment areas to country, region, or region-rural trends.[6] The comparisons become more and more relevant, perhaps approaching unconfoundedness, with each additional match on a pre-treatment variable. We use machinery from causal inference literature (e.g. the concepts of unconfoundedness and pre-treatment variables, and matching algorithms) in order to get as close as possible to relevant comparisons. For most of our causal models, our estimands are treatment effects averaged across all ten Millennium Villages, adjusted for pre-treatment variables. We also propose to extend these models to allow treatment effects to vary by MV cluster. These models can identify MV-specific treatment effects only if pre-treatment variables account for most of the within-country village variability.[22]

Mitchell et al. describe a design analysis to demonstrate our anticipated uncertainties given the design.[22] As emphasized above, we cannot guarantee statistical significance, and the impact evaluation will only be one summary of the MVP's contributions.

6.3 Data sources in candidate comparison areas

We require pre-treatment variables in the ten countries, measured at a fine enough geographic scale to identify good comparison villages and for regression adjustment in our causal models. Below we describe the data sources that were considered.

Geographic data

We collected geographic data from geographic information system (GIS) databases, including agroecological zone, travel time to nearest city of more than 100,000 population, soil composition, vegetation index, temperature, elevation, and population density.[48, 49, 50, 51, 52, 53, 54, 22]

Given the need to match the MV1s to comparison areas of similar geographic size, and the scale of the geographic variables, the geographic data sources were processed at a country level using fishnets with square grid cells approximately equal in area to each country's MV1, ranging from 2km $\times$ 2km to 12km $\times$ 12km. Processing outputs were merged into a single table for each country. These grid cells cover

each country, making them a favorable choice for matching units. In each of the ten countries, we define treatment units as the set of grid cells that overlap the MV1, and have either at least 40% area in MV1 and MV2 combined or have at least 20% area in MV1. Though this includes space outside of the treatment area, the level of precision in our data sources requires that we assume that nearby areas have similar variable values (this is often referred to as "spatial smoothness"). Furthermore, it is plausible that variable values in neighboring areas are correlated with outcomes of interest (e.g. the soil quality in a nearby area may influence a location's health outcomes). Within each country, these treatment grid cells are contiguous. The number of them varies from two to four, depending on the country. Our set of candidate comparisons excludes these treatment grid cells, as well as any cells overlapping the MV2 or overlapping the ten kilometer buffer zone enveloping the MV cluster.

Census data

With our site teams we are reaching out to national statistical authorities to acquire georeferenced census data. Census data for our ten countries is not readily accessible at a fine enough (i.e. roughly MV cluster-size) geographic scale with which to identify potential comparisons. Furthermore, boundary data is required in order to map administrative units. For selection of comparisons we can only use census data prior to the project baseline (2005), for which boundary data at a fine enough geographic scale is often unavailable. We have been working to resolve this issue. Due to time and resource constraints, census data will not be acquired and processed in time for selection of comparisons.

Demographic and Health Surveys

The Demographic and Health Surveys (DHS) measures many of our outcomes of interest using survey tools comparable to ours.[55, 56] Census enumeration areas serve as the primary sampling units (i.e. clusters) for the DHS two-stage sampling. At the second stage, roughly 20-30 households are sampled from each cluster.[57] To protect anonymity, DHS displaces GPS locations of clusters by up to five kilometers in rural areas.[57, 58] Unlike the geographic data, the DHS data do not correspond to grid cells, but instead to *DHS buffers*, circles around DHS cluster points with radius equal to the maximum possible anonymity displacement of five kilometers. Effectively, we approximate the enumeration area boundary with the DHS buffer. This is a reasonable approximation if neighboring areas tend to be similar on characteristics measured by the DHS (health, wealth, and education indicators). Unfortunately, DHS data is geographically sparse, with approximately 350-900 out of 8000-600,000 enumeration areas sampled per country, and approximately 20-30 households sampled within each enumeration area. See Figure 3 for an artificial example, with comparison grid cells chosen for the purpose of demonstration in this paper.

Other surveys

In addition to the DHS data, other survey data sources include UNICEF's Multiple Indicator Clusters Surveys (MICS) and World Bank Living Standards Measurement Study (LSMS) surveys. However, the MICS and LSMS are only georeferenced at the district level. This resolution is too coarse and therefore cannot be used to determine comparisons. Instead, during further stages in our analyses, we propose to include district-level aggregates from these data sources as covariates in the small area models discussed below.

Combining data sources - small area estimation

In the future (after selection of comparison areas), we plan to combine DHS data with census and geographic data to obtain better estimates of cluster-level DHS variables.[59, 60, 61, 62, 63] These cluster-level estimates (often known as "small area estimates") are associated with DHS buffers due to the anonymity displacement. It is expected that our most accurate estimates of DHS variables will be obtainable within DHS buffers.

Before selection of comparison villages, our small area models will only use geographic variables to predict DHS variables, due to delays in acquiring and processing finely georeferenced census data. See Mitchell et al. for the small area models we use to estimate DHS variables using geographic data.[22] However, we will continue to work with site teams to procure census data, and we will eventually fit small area models using these data. These results can be used in the causal models described in Section 6.5.

Processing of geospatial data

All processing was done in the WGS-84 geographic coordinate system.[64]. DHS buffers were generated using ArcGIS Buffer tool which supports geodesic buffers, i.e., those that account for the actual shape of the earth in the calculation of the buffers.[65] At our latitudes and resolutions, the differences that could result when the data are projected are likely to be insignificant for both the grid cell fishnet and buffers.

6.4 Selecting comparison villages

As mentioned above, our matching units are grid cells of size equal to the MV1 treatment areas. Corresponding to these grid cells are geographic variables, and corresponding to DHS buffers are wealth, education, and health variables. In Mitchell et al. we detail the conversion between the grid cells and DHS buffers.[22] We wish to select the best subset of five grid cells (with "best" described below), to serve as our comparisons.

For all countries except Tanzania, Nigeria, and Ethiopia, at least one DHS buffer overlaps the treatment grid cells. For these seven countries we restrict the set of candidate matches to grid cells overlapping DHS buffers. This ensures that we have pre-treatment data on many outcomes of interest in our selected comparison grid cells. Although the MVP collected baseline data in 2005 and 2006, we have doubts about its comparability to DHS data. Therefore, we do not restrict our matches for Tanzania, Nigeria, and Ethiopia to come from areas with DHS data. We consider DHS data from DHS buffers overlapping any of the treatment grid cells as relevant to all (two to four) treatment grid cells. This is justifiable because we believe that wealth, education, and health indicators vary (somewhat) smoothly across rural areas, with neighboring areas having similar values.

For each MV1, our matching procedure begins with a restriction to grid cells within the same country and agroecological zone, given the project's emphasis on farming systems. Each MV cluster is contained within a district (or local equivalent of district). We limit matched comparisons to come from this district or any districts that border the MV cluster. This is primarily for logistical reasons related to the survey fieldwork, but also because we suspect that areas closer to the MV are likely to be closer matches on many relevant variables than more distant areas.

As described above, we aim to minimize spillovers by choosing comparison villages at least ten kilometers away from the MV cluster, outside a buffer zone of very likely spillover effects. In several countries, the district containing the MV is small, leaving only a few grid cells within the district that are outside of the MV cluster and buffer zone of likely spillovers. Therefore, following Stuart and Rubin,[66] we choose

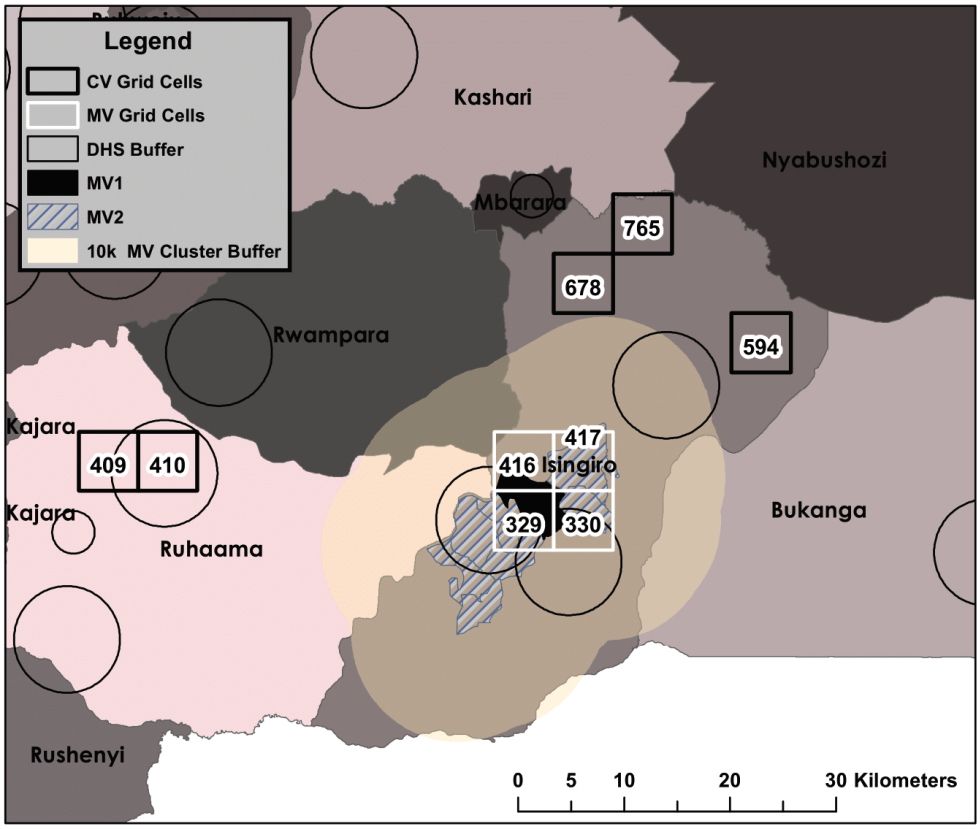


Figure 3: Here is a map of Uganda in the region surrounding the Millennium Village cluster. The core area in each Millennium Village that receives the full set of interventions, the MV1, is colored in black. The areas that received a subset of interventions, the MV2, are striped. A ten kilometer buffer is shaded in blonde. The DHS buffers, circles around DHS cluster points with a five (for rural) or two (for urban) kilometer radius, are drawn as circles. Treatment and comparison grid cells are in white and black, respectively, with comparison grid cells displayed for demonstration only. District boundaries, in different shades of gray, are drawn using boundaries from GADMv2.[52] The Millennium Village cluster is located in Isingiro district.

matches both inside and outside the MV district, matching on variables described below. There is tension between a desire for within-district matches (government programs are often implemented at the district level) and wanting close matches on other variables. The literature does not offer a simple solution to this tradeoff, so we consulted with subject-matter experts who recommended requiring that at least two of the five matched grid cells be within the district containing the Millennium Village.

We want to match on many variables to make the unconfoundedness assumption credible. If treatment assignment is unconfounded given covariates, then assignment is unconfounded given the propensity score (the average probability of treatment for those with common values of the covariates).[67] For selecting matched controls, it is often simpler to use a single value (the propensity score) rather than a collection of variables. However, in the MVP case, there are few treatment units (i.e., grid cells clustered in the ten MV clusters), and therefore, it is difficult to fit propensity score models with many covariates. Instead, we consolidate groups of relevant variables into indices to reduce the number of dimensions on which to match.

First, we select variables most closely related to our outcomes of interest available from the Demographic and Health Surveys (DHS). Next, we combine related variables into wealth, education, and health indices. As a measure of wealth, we use the DHS asset index of each household, which is the first principal component

from a principal components analysis.[68, 69] Our chosen geographic and DHS variables, and our procedure to create indices are detailed in Mitchell et al.[22] There we also describe our handling of missing data in both the geographic and DHS variables. For geographic variables we match on the missingness pattern (this is justified in Mitchell et al.[22]), and we create DHS indices using only available cases, leaving more sophisticated methods of handling missingness to future work.

After we restrict to neighboring districts and the MV's agroecological zone (and for all but three countries, to grid cells overlapping DHS buffers), we can search through all matched sets within a reasonable time (somewhat arbitrarily set at under 48 hours). We specify that at least two matched controls must be within-district (see above). Each iteration computes the match's "badness score," a measure of covariate imbalance described in Mitchell et al.[22] The measure combines two standard balance measures: the standardized difference in means between treatment and comparison groups,[70] and the ratio of standard deviations between treatment and comparison groups.[31] Because the DHS indices summarize baseline values of the outcome variables, the badness score gives them greater influence on the choice of matches.

As discussed previously, in Tanzania, Nigeria, and Ethiopia we do not restrict matches to grid cells overlapping DHS buffers because these countries' treatment cells do not overlap any DHS buffers. Treatment cells in Kenya and Uganda do overlap DHS buffers, but in Kenya only one grid cell within the district and agroecological zone overlaps DHS buffers, and in Uganda there are none. Therefore, in Kenya and Uganda we will select two or three within-district matches using geographic data only, but restrict the remaining two or three matches to come from areas with DHS data, from the neighboring districts (thus providing enough DHS information to estimate the imbalance measures mentioned above).[22]

In addition to the statistical and algorithmic procedures, selection of matched comparisons will involve subject-matter experts. We will present our matching results to development economists, public health practitioners, geographers, and agricultural scientists. We do not present these experts with maps as in Figure 3, since this might allow experts to use post-treatment information about the displayed matched grid cells. Instead, our format for presentation is shown in Figure 4. This visual allows the experts to see differences between treatment and comparison grid cells in the pre-treatment variables. If a particular variable appears to be poorly matched, we can rerun our procedures with an adjusted badness score that gives more weight to the problematic variable. This interaction between statistics and subject-matter knowledge improves the relevance of the selected comparisons.

After the selection of comparison grid cells, we will provide our field teams with the grid cell coordinates. For each grid cell, they will list all villages for which a majority of households fall within the cell boundary. We will then randomly select one village per grid cell in which to survey, see Section 10.

A detailed description of the matching procedure and resulting selected comparison villages will be released in subsequent papers. See Mitchell et al. for technical details.[22]

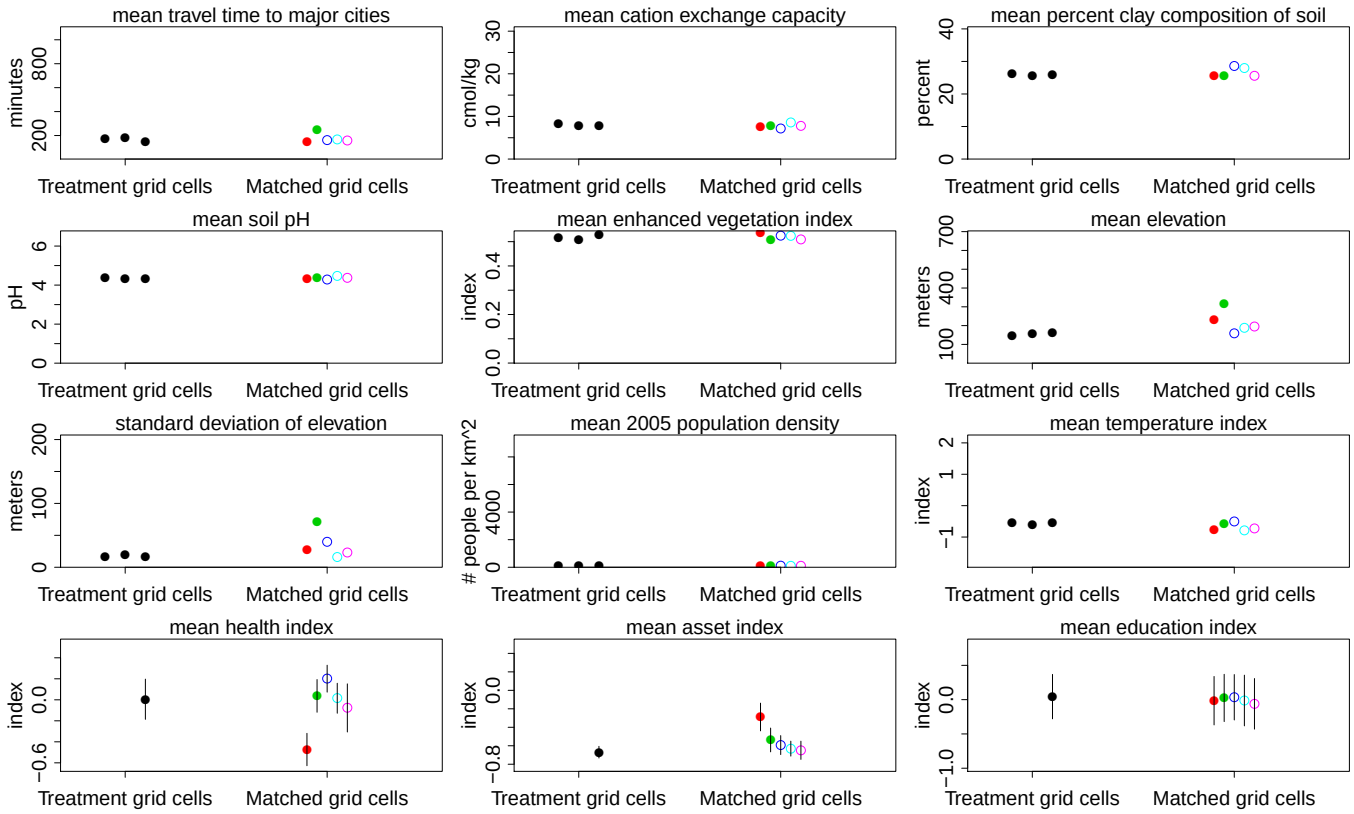


Figure 4: Values for the matching variables in both the treatment grid cells and matched comparison grid cells in Ghana. Each circle represents a grid cell. The treatment grid cells are black circles, while the comparison cells are colorful circles. These colors are used to identify each comparison cell, allowing the experts to compare across variables (e.g. one color/comparison may do well on one variable, and poorly on another). Within-district grid cells are represented by filled-in circles represent, while out-of-district grid cells are represented by empty circles. For the DHS indices (education, assets, and health), we also present 95% uncertainty intervals from our small area estimation procedure. Since only a subset of the treatment grid cells overlap DHS buffers, there are fewer black circles for these variables. The axes for the DHS variables are scaled such that they begin at the minimum value in Ghana and end at the maximum value in Ghana. This provides context for the unit-less indices.

6.5 Candidate models for causal inference

In Mitchell et al. we detail the multilevel Bayesian causal models that we will fit to the 2015 survey data.[22] These models range from simple to complex, building up from a very basic model to one that accounts for various study aspects (e.g. adjusting for covariates, allowing treatment effects to vary across countries, adding village-level variation). Our models only adjust for pre-treatment variables.[28] Though we have panel data in the MVs (see Section 10), we do not have corresponding data in candidate comparison areas. Therefore, we are limited to adjusting for aggregate baselines.[22] The final evaluation will report and compare results from all models and all outcomes (reporting posterior intervals of uncertainty), reducing the scope for fishing, i.e. reporting a model based on the realization of the conclusion.[71, 72]

First we will fit regression models to each outcome individually. Additionally, we will construct a multi-outcome model that will allow similar indicators to inform each other, as recommended in Gelman et al.[73] We define groups of similar indicators as follows:

- poverty indicators: composed of our MDG 1 indicators, our MVP agriculture indicators (a.1 to a.4), and MDG indicator 8.15 (access to mobile phones);
- education indicators: composed of our MDG 2 and 3 indicators, and MVP education indicators (b.1 to b.3);
- child health indicators: composed of our MDG 4 and 7 indicators, and MVP health indicator c.1;
- maternal health indicators: composed of our MDG 5 indicators; and
- HIV-malaria indicators: composed of our MDG 6 indicators.

For each of the above groups of indicators, we will have a summary measure, either from the multi-outcome model or constructed from the individual-level models. Inevitably, with many separate analyses, there will be some that reach the "statistical significance" threshold and some that do not. To offset this issue, we will consider two overall summary measures: one of all the indicators and one limited to only the Millennium Development Goal indicators and proxies. See Mitchell et al. for further details.[22]

6.6 Externalities

There are three possible types of interferences between the treatment areas and candidate comparison areas, sometimes referred to as "externalities", which may interfere with the SUTVA, discussed above.[21] The first is a spread of services to nearby areas, reduction of infection risk, and use of services within the MV cluster by temporary or sustained migrants. Second, government spending within the district containing the MV may shift from the MV to other areas in the district. Third, areas outside the cluster may imitate the MVP interventions or adopt policies such as bednet and fertilizer distribution.

Our choice of comparison villages will be limited to areas at least ten kilometers away from the MV areas, outside a "buffer zone" of very likely spillover effects. This mostly works towards minimizing the first type of externality. Moreover, our causal models will try to estimate treatment effects that vary by distance to the MV1, though we may not have the precision to estimate these treatment effects without strong regularization via prior distributions.

We have not tracked government spending in the districts where MVs are located, and therefore cannot estimate the second type of externality. Since the start of the project, some of the MVP interventions were adopted outside of the MV area. However, we cannot know the extent to which the outcomes in these areas would have happened without the MVP. Our results can be interpreted as comparisons between the

MVP interventions and the set of interventions implemented in the comparison villages. See Section 7 for a related discussion about cost effectiveness ratios.

6.7 Software

For fitting multilevel models we use Stan in R.[74, 75] For geographic data processing we use ArcGIS.[65]

7 Cost Assessment

A fundamental hypothesis of the project is that the MVP package of interventions can be delivered at a modest cost. The needs assessment conducted by the UN Millennium Project estimated that achieving the Millennium Development Goals (MDGs) would require an annual investment ceiling of $120 per person per year (in 2005 USD) toward local service delivery and community-based investments during the ten-year period from 2005 to 2015.[76, 77]

The $120 is not an increment above a baseline level of spending. Rather, it is the estimated total cost of the package of interventions, some part of which is in place without the MVP. The MVP is therefore providing additional financial support, with the aim of adhering to a total investment ceiling of $120 per person per year. The incremental "cost" of the MVP is therefore not the full $120, but only the additional MVP money spent in the area, along with any additional money spent by the government and external donors, above what they would have spent in the absence of the MVP. We cannot know precisely how much incremental spending was a result of the MVP unless we assess costs in the comparison villages (which will not be possible due to budget constraints). We do know that the project has contributed an approximate annual maximum of $60 per person towards the $120 ceiling. The project budgeted $60 per person per year for the first half of the project, with a planned decline in project funding from 2011-2015. The decline in funding differs among each site, depending on each site's particular financial landscape. The project's exact investment figures for each year will be reported in the final evaluation, along with the investments from each external stakeholder.

The $120 annual per capita costing model does not reflect the entire cost of the MVP, but approximates the "on the ground" costs of delivering the services and interventions co-implemented by the MVP. It includes the costs of service delivery, implementation, and on-site management, including estimated values of in-kind donations. Off-site costs, comprising salaries and overhead for all scientific and support staff at the Earth Institute and Millennium Promise staff based in New York and at the regional MDG Centers in Dakar and Nairobi, are excluded from this cost assessment. These excluded off-site staff are primarily involved in project design, implementation research, monitoring and evaluation, logistics, and fundraising. They are not involved in direct operations, so their costs should be considered a one-time cost to design and operate the project, rather than an ongoing cost of running an MVP-style project in a scale-up context.

7.1 Methodology

The goal of the costing evaluation is to measure the cost of carrying out specific services, such as malaria control or irrigation development, and not the cost of specific outcomes. This type of detailed costing information is not typically collected and constitutes one of the substantial contributions of this project's evaluation.

Costs are collected for the entire project area (MV1 and MV2), but a distinction between MV1 and MV2 costs will be drawn, allowing for the costs in the MV1, where investments have been more heavily concentrated, to be distinguished and analyzed separately. Due to the varying degree of detail in external stakeholders' expenditure records, as well as the spillover effect of certain investments and the difficulty of isolating beneficiary groups, it will not be possible to distinguish perfectly between MV1 costs and MV2 costs for every intervention. In cases where estimations are necessary, all assumptions made will be recorded and clearly outlined in the final evaluation.

Of the estimated $120 per capita annual cost ceiling, the project (Millennium Promise) supplied, on average, approximately $60 per capita per year during the first phase, and has supplied a reduced amount for each remaining year. National and local governments, external donors (including NGOs, multilateral organizations, and private donors), and the local community (whose contributions are mainly in-kind as labor and material inputs) supply the rest (Figure 5).[78] Understanding these inputs is critical to evaluating the success of the project in relation to the $120 per capita annual project ceiling, and to assess scalability of project interventions.[3] The nature and intensity of inputs is likely to differ substantially between clusters due to community needs, local disease profile, and local economic base.

Ideally, we would estimate two ratios to understand cost-effectiveness of the MVP. In order to describe them, we consider one outcome of interest, Y, and define the following variables (some of which cannot be measured):

$Y(1)$ is the outcome in the Millennium Village if the MVP package of interventions is implemented at the site (also including interventions implemented by government, community, and external donors).

$Y(0)$ is the outcome in the Millennium Village without the MVP package of interventions, including only interventions that would have been implemented by government, community, and external donors had MVP not existed.

$Y(-1)$ is the outcome in the Millennium Village without any interventions implemented by anyone, including the government, community, MVP, or other partners.

$\$N$ is the total amount spent in the rest of the district per capita by the government, community, and external donors.

$\$MVP$ is the total amount spent in the Millennium Village per capita (including government, community, external donors, and project funding).

Assume $Y(1) > Y(0) > Y(-1)$, in other words, that with more interventions, the value of the outcome increases, and that higher values are better outcomes. We also assume that $\$MVP > \$N$, i.e. that the project spends more money per capita than the government, community, and external donors. One ratio that could be of interest considers the incremental cost versus effect of the MVP relative to interventions only by the government, community, and external donors,

$$\frac{\$MVP - \$N}{Y(1) - Y(0)}. \tag{1}$$

A second ratio considers the total cost versus effect of the MVP relative to no interventions,

$$\frac{\$MVP}{Y(1) - Y(-1)}. \tag{2}$$

Neither ratio 1 nor 2 can be estimated by this evaluation. As previously mentioned, no costing information is going to be collected from the comparison villages. Thus, the numerator of ratio 1 is not estimable. We cannot measure the denominator of ratio 2, since the counterfactual outcome for areas with no interventions is not estimable. (The impact evaluation uses comparison villages to estimate $Y(0)$, outcomes in areas that do not receive the full package of services provided by the MVP. See Section 6.)

The only estimable ratio, combining the impact evaluation and cost assessment, is

$$\frac{\$MVP}{Y(1) - Y(0)}, \tag{3}$$

which is likely higher than both ratios 1 and 2, overstating the cost per outcome. Our evaluation, however, will not estimate this ratio and instead estimate the numerator and denominator separately.[79, 80, 81]

The goal of the cost assessment is to measure the cost of carrying out specific services, such as malaria control or irrigation development, and not the cost of specific outcomes. This type of detailed costing information is not typically collected and constitutes one of the substantial contributions of this project's evaluation.

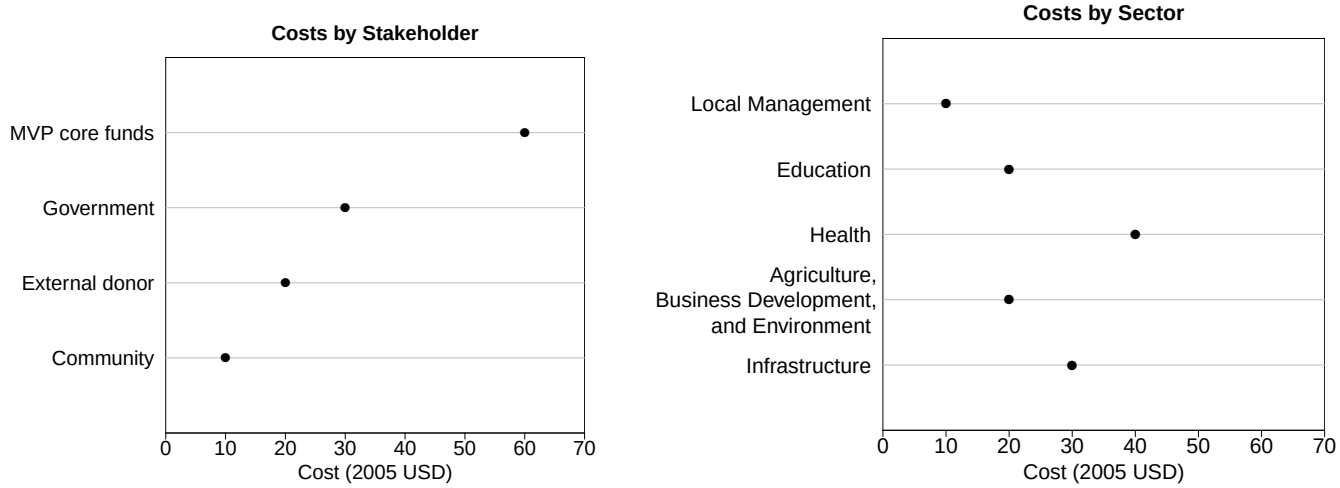


Figure 5: Annual per capita costing model, by stakeholder and sector.

A full economic cost assessment, in line with established methods of social and health policy interventions is underway in each project cluster.[82, 83, 84, 85, 86] The aim of the assessment is to document the annual on-site costs of the project by site, stakeholder, sector, year, and within the MV1 only as well as the entire cluster.

Core project investments and external stakeholder investments are tracked via two complementary mechanisms. Core project expenditures (those made with funds that flow through the Millennium Promise bank account) are tracked and reported quarterly via the project's internal cost-tracking system. Expenditures and in-kind donations provided by external stakeholders (government, community, and other donors) within

the sectors of education; health; agriculture, animal husbandry, business development, and environment; and infrastructure, are collected or estimated periodically by local site team members.

A series of data collection templates have been created for each stakeholder operating within each project cluster (including both MV1 and MV2). There are approximately 20 government, donor, and community stakeholders per project cluster. All costs within the defined project sectors (see above) are collected. The costing data collected via these individual stakeholder templates are amalgamated with the core project costing data collected via the internal tracking system to form a single comprehensive costing database for each project cluster.

For contributions made in kind, all prices are documented using the standard cost imputation method recommended for multi-center interventions.[87, 88, 89] This method involves establishing local unit costs for each in-kind contribution (e.g. daily wage rate in the case of labor contributions). These unit costs are then used along with qualitative data collected during key informant interviews to calculate a total cost for each contribution (e.g. daily wage rate $\times$ number of laborers $\times$ number of days worked).[87, 88, 89]

7.2 Data management and analysis

After the costing data from all stakeholders have been collected and aggregated, the data will be archived and available for analysis. For the purposes of cross-site comparison, compatibility with core project expenditures, and measurement against the MVP costing model, all expenditure amounts will be converted to 2005 US dollars using average annual exchange rates for each project year. Annual per capita expenditures will be calculated for each project cluster, MV1, sector, and stakeholder, using the total cluster population (MV1 and MV2) as well as the MV1 only. Analysis of the data will be focused around questions of sustainability, replicability, and scalability.

8 Process Evaluation

Since the MVP was designed as a demonstration project to achieve the MDGs, its multi-sectoral model of integrated rural development has been implemented across a wide range of physical and social environments. The MV site selection process originally emphasized agroecological diversity (see Section 2.1), but diversity across sites is further reflected in their many geographical features, political, economic, cultural, and social structures. The sites have also been subject to varying degrees of government involvement since the project's inception, as well as different forms of community engagement. In order to capture some of these complexities, a qualitative process evaluation (PE) component will be included in our evaluation design. Our 'process evaluation' is a synthesis of ten case studies (one per site) using data from stakeholder interviews and focus group discussions. For the purposes of this evaluation, the PE will also provide an entry point for other qualitative research related to the state of the MV sites by 2015 (see Section 8.1). Individual investigation of the ten MVP cases should further clarify how the MVP model evolved within each site's context. Key development processes, factors, and outcomes will then be compared across sites, largely based on community, MVP field staff, and government interpretation of the project's implementation.

8.1 Process evaluation objectives

Qualitative data will be collected to serve the following four objectives:

1) To describe the development and implementation of the distinct packages of multi-sectoral interventions within each MV site's context (e.g. summary timeline of interventions, design of service delivery systems, core sector strategies, cross-sector cooperation, sustainability management, etc.);

2) To provide insights into the causal mechanisms which contributed to the observed outcomes, outlining the major barriers to and enabling factors for the implementation of the MVP core interventions;

3) To elicit stakeholder input on the project's biggest successes and failures and its long-term sustainability; and,

4) To disseminate its greatest lessons learned for policymakers and development practitioners.

Researchers will explore various themes raised during interviews and focus group discussions. Although we cannot determine the causal mechanisms generating each site's treatment outcome, we will use PE data to complement and contextualize the quantitative results from household surveys. Similar types of qualitative assessments have been recommended in mixed methods approaches to evaluating complex interventions.[90, 91, 92]

8.2 Changes from previous PE research

Earlier in the project, the focus of PE research was to assist project field staff in improving the course of interventions. Accordingly, previous rounds of qualitative research focused on topics such as determining the most appropriate sequence of implementing activities, assessing unintended consequences, and documenting the step-by-step approaches involving all stakeholders. The needs of a final project evaluation require a much broader approach. In addition to implementation research, the PE will now study perception of the MV sites in 2015, reflections on the entire MVP trajectory, and opinions on post-2015 sustainability (see Section 8.1).

8.3 Methodology

The PE will employ two approaches to collect evidence: 1) key informant interviews and 2) focus group discussions. For both, we will target three categories of MVP stakeholders: project field staff, government officials, and community members across the entire cluster. Approximately ten individual key informant interviews and thirteen focus group discussions will be conducted at each MV site. Semi-structured questionnaires will allow flexibility to adapt the discussions as needed. Additional data will come from archived reports or documents found at each MV site's field office. Various forms of participant-observation, consented photographic evidence, and field notes may also be used.

Project field staff (the team leader, sector coordinators, and operational personnel) offer technical expertise and intimate knowledge of their site's implementation history. First, the team leader and sector coordinators will be interviewed to explore: overall perception of the MVP; sector processes and highlights; sector impact insight; government and community relations; and miscellaneous feedback. Second, one focus group discussion will be conducted with sector coordinators to investigate: multi-sectoral implementation and outcomes; project impact insight; sustainability, replication, and scale-up; and other cross-cutting themes. Other senior staff members at the site may also be included in the focus group discussion at the

discretion of each team leader. Third, MVP staff members will be issued an open-ended, written-response template, which they may use to provide additional feedback at their discretion.

Local, district, regional, and national government officials will offer insight on policy externalities, spread of services to nearby areas, and government investments in the project. We will conduct three to five key informant interviews with local and district-government officials who have worked closely with the project field team, including those who have been seconded to the MVP as field facilitators or planning officers. We will also attempt to interview government officials at the regional and national levels to provide their perspectives on the MVP's successes and failures, influence on policy changes, and post-2015 sustainability. Although a causal analysis of policy changes will not be possible in this evaluation, we will attempt to qualitatively document the roles that government actors have played in the MVP, as well as any notable policy changes that may have taken place as a result.

Community members (e.g. adult men and women, opinion leaders, teachers, local business owners, health workers, political leaders) will explain their experiences with the MVP and their perception of its impact. There will be approximately two focus group discussions on the following topics: nutrition; maternal and child health; HIV/AIDS, malaria, and TB; education; agriculture and business development; water, sanitation, and infrastructure. Additional topics might include gender relations, institutions, private sector activities, cultural practices, mobility, technology, or other MVP-pertinent topics that emerge from the data collection process.

8.4 Recruitment of informants

For the MVP field staff informants, team leaders, sector coordinators, and operational personnel will be invited to participate. For the government informants, officials seconded to work with the MVP, technical experts, local political leaders, and other government partners of interest will be invited to participate. For the community informants, MVP field staff and government extension agents will identify voluntary research participants via word of mouth and community boards. The ages of participants will range from 16-59 years old. Their informed verbal consent must be acquired before they become eligible to participate. Focus group discussions will be carried out at either the cluster level or in targeted localities of both the MV1 and MV2. Due to limited resources, we will not be able to conduct any qualitative research in the comparison villages.

8.5 Qualitative research epistemology

PE research borrows more heavily from postpositivist and interpretivist traditions in the social sciences than from the positivist sciences.[93, 94] Whereas the adequacy assessment and impact evaluation collect data from a random group with the aim of drawing inference to a larger population, PE collects data from a nonrandom group with the aim of learning from different stakeholder vantages. For some aspects of the PE, this means managing informant bias to verify the objectivity of statements made. For other aspects, it means embracing the subjective quality of informants' responses and putting various interpretations of stakeholder knowledge, attitudes, historical memory, and common practices into conversation with each other. Interview and focus group discussion data will be analyzed with these subjectivities and interpretations in mind.

Many factors also shape PE responses during the data collection process. These may include practical factors (e.g. time of day, location of discussion, expectation of remuneration, research fatigue) as

well as social ones (e.g. gender, age, religion, social status, parties present, political position, etc.). In order to optimize technical consistency across sites without compromising the valuable input of in-country knowledge, the PE data will be collected through the combined work of: 1) external (i.e. often Western) researchers who will provide technical expertise and conduct qualitative research training, 2) in-country qualitative consultants who will contribute to sampling and data collection, and 3) regionally-based research assistants/interpreters. This socially mediated process of data collection will be a cross-cultural and cross-linguistic one. We acknowledge that identity politics may affect the PE data we collect, and in the absence of long-term participant observation, underscore the fragmented quality of the data that will be collected within the allotted timeframe. Despite some of the challenges qualitative field researchers confront,[95, 96, 97, 98] much can still be learned across the etic/emic divide of "outsider" and "insider" knowledge, wherein the combined contribution of multiple research vantages should strengthen the overall quality of the data collected.

8.6 Data management and analysis

Key informant interviews may be conducted in either English, French, or other local language, depending on the linguistic comfort of the informant. Content from qualitative interviews and focus group discussions will be digitally voice recorded in the field, then transcribed and archived. Researchers will also draft field notes. We will use qualitative analysis software, Nvivo,[99] to code and memo data by thematic content. Ten country case studies (one per site) will then be generated from the organized data sets analyzing the objectives described above. In addition, thematic reports and 'lessons learned' documents will describe other qualitative and policy findings from the PE research. The intended audience includes development practitioners, government actors, policymakers, academics, and other members of the international community.

8.7 Mixed methods interpretations

Mixed methods approaches to causal inference and program evaluation further suggest that qualitative and quantitative data components should complement each other.[100] Coupled with the household survey results, for example, qualitative data can help explain certain outcomes. Understanding the mechanisms and processes behind how multi-sectoral interventions were designed and carried out will provide social context for interpretation of household survey and impact evaluation results (e.g. the roles of Koranic and formal schools in the educational landscape of Senegal). Once all data have been collected, we will combine findings from various evaluation components to optimize insight.

9 Description of Systems Design and Tools

An additional component of this final evaluation will be documentation of the various Millennium Villages Project (MVP) systems designs, tools, instruments, and other technical features of the project. These features will include the SharedSolar microgrid technology,[101] the Formhub data collection platform,[102] and the Community Health Worker system and mobile health application on the CommCare platform.[103] Implementation guides will also be made available for the MVP health, education, and agribusiness system components. We expect that this output will be used, refined, and adjusted by development practitioners, leading to further innovations.

10 Survey Data Collection

Throughout the implementation of the MVP, data have been collected in four ways: population-based surveys (in years 2005-2006, 2008-2009, 2010-2011), routine monitoring systems, economic cost data, and qualitative interviews. Survey data were collected from the MV1s (the core areas of each Millennium Village). As in previous years, prior to the survey administration, a census will be conducted in each MV1 to record basic demographic data of all household members.

We define household members to be those who have lived in the household for at least 3 of the past 12 months, and who 'normally eat from the same pot.' Additionally, the following persons are always considered to be household members: the main provider for the household, and infants who are less than 3 months old.

This section describes the plan for final survey data collection in the MV1s and matched comparison grid cells, modified from previous data collection plans. These data are inputs to the adequacy assessment and impact evaluation. Additionally, we describe the routine monitoring systems and their use as supporting data sources.

At the start of project implementation (2005-2006), 300 households were randomly sampled from each MV1. Those remaining from the original 300 households were included in subsequent survey rounds (in 2008-2009 and 2010-2011). At end-line we will survey those of this original sample of households who were measured in all subsequent survey rounds and still reside in the MV1. We refer to these households as "panel households" and their data as "panel data." Within-household sampling for panel households is described below. (The panel data do not serve an explicit purpose in the design of this evaluation, as they have no equivalent in the control villages. The impact evaluation and adequacy assessment use cross-sectional data only. However, other researchers have requested these data because ten-year-long panel data on households in rural sub-Saharan Africa is rare and potentially useful. We describe collection of panel data for completeness.)

Additionally, cross-sectional survey data will be collected within each MV1 and comparison village using a two-stage cluster design.[104, 105] The first stage will be a simple random sample (SRS) of households each selected with equal probability. We refer to these households as "cross-sectional households" or the "cross-sectional sample," and we refer to their data as the "cross-sectional data." We will sample 300 households out of the roughly 800 to 1200 households in each MV1, and 300 households across the comparison villages for each MV1. In the Millennium Villages, with probability almost one, some households selected for the cross-sectional sample will by random chance also be panel households. The second stage sampling differs from module to module as is described below, with technical details, including design analyses used to guide sample size recommendations, in subsequent papers.[22] The target sample sizes for each module and age-sex group were determined based on a combination of budget, logistics, and relative importance of different vulnerable populations and intervention beneficiaries.

As described above, we will randomly select one village per matched grid cell, giving five comparison villages for each MV1. Sampling in each comparison village will be identical to sampling in the MV1 cross-sectional sample, with sample sizes divided by the number of comparison villages. In comparison villages we will not have demographic census data prior to survey data collection. Rather, prior to sampling households within each comparison village, we will create a household list: a list of all non-abandoned households (determined by outside appearance, without consultation of household members), with GPS

coordinates identifying the location. From this list, we will choose a simple random sample of households. Next, we will administer a demographic census in those sampled households, which will serve as the sampling frame for the within-household sampling. Enumerators will be provided with pre-populated survey tools identifying the sampled individuals prior to administration of each survey module. If a sampled individual is not found at home when the enumerator visits, any present household members will be asked if the sampled person resides in the household and why they are absent. If the person resides in the household, the enumerators will be instructed to make up to six visits to the household to attempt to reach them. Generally, they will consult with other household members to determine a time when the person is most likely to be available. If the person cannot be reached after six attempts, this will be recorded as missing data.

The MVP survey tools draw from the Demographic and Health Surveys (DHS), UNICEF's Multiple Indicator Clusters Surveys (MICS) and the World Bank Living Standards Measurement Study (LSMS) surveys.

10.1 Household surveys

The household survey will be administered to all household heads (or other knowledgeable household members) within the cross-sectional households in the MV1s and their comparison villages, and within panel households in the MV1s. This module will capture information on household demography, education, malaria bed net usage, agricultural and non-agricultural sources of income, assets, expenditure, consumption and access to basic services including water and sanitation, health care, energy, transport and communication.

In- and out- migration

In the models for the impact evaluation, we hope to include a variable for time exposed to the MVP interventions. However, during the demographic censuses completed before each survey round (in the years 2005-2006, 2008-2009, 2010-2011), no information was collected on whether migration was from or to villages outside the MV1 cluster of villages. Questions have been added to the final household survey about the household head/primary provider's duration of residence in the MV1s. Given the difficulty in tracking all individuals' migration histories over the past ten years, the household head/primary provider's migratory experience will serve as a proxy for individual-level in-migration. This individual will be asked when they moved to their current village and from where they moved. The data will be used to generate a variable for household head time exposed to MVP interventions.

Since the migration module described above will only be administered to current (2015) household heads, permanent out-migration (households that have left the MV1 and do not return by the 2015 data collection period) will not be logistically feasible for our team to measure.

10.2 Adult surveys

A sex-specific adult survey will be administered to men and women of reproductive age (15 to 49 years) within the cross-sectional households in the MV1s and their comparison villages, and within panel households in the MV1s.

In each MV1 the sampling will be done as follows: If the total number of men (respectively, women) age 15 to 49 years in the cross-sectional households is at most 500, then all are sampled. Otherwise, a systematic sample of 500 men (women) will be taken from these cross-sectional households.[22, 105] We will continue to either sample all men (women), or to systematically sample, in the rest of the panel households that are not in the cross-sectional sample. Sampling in each comparison village will be similar, without panel households, with the sample divided amongst the comparison villages within a country.[22]

The male and female surveys include questions regarding marital status, sexual and reproductive health, treatment seeking behavior, HIV knowledge, and mental health. In addition, the female survey also examines birth histories, contraceptive use, child health and immunization history, and infant and young child feeding practices. We will administer the birth history module to more women than those sampled for the adult female survey, in order to get better estimates of under-5 and infant mortality rates. We will take a simple random sample of extra households outside of our cross-sectional sample, and administer the birth history module to women age 15 to 49 years in those extra households, matching the adult female survey sampling method (all women or a systematic sample of women). The number of extra households will be chosen so that the total number of women sampled for the birth history module is expected to reach 1000 women in each MV1, and a total of 1000 women across each MV1's comparison villages.[22]

10.3 Nutrition surveys

We will administer a food frequency questionnaire to men and women age 15 to 49 years in sampled households. The food frequency questionnaire was adapted and piloted in each site in previous rounds to reflect the local diet and to measure food intake and dietary diversity over a one-month recall period. Due to (presumably) high intra-house correlation in food consumption habits, we will only administer the food frequency questionnaire to one adult per household.

In each comparison village, we will randomly divide the sampled households into two equal-sized groups (or as close as possible if there is an odd number of households). One of these groups will be designated to contribute men, and the other to contribute women. In each MV1, we will consider three strata: 1) households only in the cross-sectional sample and not in the panel households, 2) households in both the cross-sectional sample and panel households, i.e. the overlapping households, and 3) households only in the panel and not the cross-sectional sample. We will randomly divide each stratum into two equal-sized groups (or as close as possible if there is an odd number of households in a stratum). One of each of these three pairs of groups will be designated to contribute men, and the other three to contribute women.

For each household designated to contribute men (respectively, women), we will select a random man (woman) age 15 to 49 years, if available. For households in this group that do not have any men (women) age 15 to 49 years, no food frequency questionnaire will be administered.

10.4 Biological and anthropometric data

Biological testing

We will conduct sampling for malaria and anemia in four age-sex groups in sampled (cross-sectional) households: children age 6 to 59 months, school-aged children (5-14 years old), men age 15 to 49 years, women age 15 to 49 years. Malaria and anemia will be tested on the same sample of people. No malaria and anemia testing will be conducted in panel households that are not in the cross-sectional sample. In each MV1 the

sampling will be done as follows: If the total number of children aged 6 to 59 months in the cross-sectional households is at most 300, then all are sampled. Otherwise, a systematic sample of 300 children will be taken from these cross-sectional households.[22, 105] If the total number of school-aged children (5-14 years old) in the cross-sectional households is at most 100, then all are sampled. Otherwise, a systematic sample of 100 children will be taken from these cross-sectional households.[22, 105] We repeat this for the men and women age 15 to 49 years.

Anemia will be tested using HemoCue HB 301 point-of-care device.[106] Those found to be anemic will be referred to the nearest health center. Malaria parasitemia will be tested using Rapid Diagnostic Tests (RDTs) from Access Bio Inc.[107] CareStart™ Malaria HRP2/pLDH (Pf/Pv) Combo test G0161 will be used in Ethiopia and Rwanda, where *Plasmodium Vivax* is prevalent.[108] In all other sites, CareStart™ Malaria HRP2/pLDH(Pf) test G0181 will be employed. Both G0161 and G0181 were tested in the WHO-Foundation for Innovative New Diagnostics (FIND) Malaria RDT Evaluation Program and meet the WHO recommended selection criteria.[109, 110] Those who test positive for malaria will be treated with Artemisinin-based Combination Therapies (ACTs).

Anthropometric data

We will assess weight, height, length, and mid-upper-arm circumference (MUAC) among children age 6 to 59 months in sampled (cross-sectional) households. No anthropometric measurements will be taken in panel households that are not in the cross-sectional sample. In each MV1 the sampling will be done as follows: If the total number of children aged 6 to 59 months in the cross-sectional households is at most 400, then all are sampled. Otherwise, a systematic sample of 300 children will be taken from these cross-sectional households.[22, 105]

Weight will be measured to the nearest 0.1kg using Seca 874 weighing scales in all sites except Senegal, where HealthOMeter 498K scales will be used. Length and height will be measured to the nearest 0.1cm using portable baby/child length and height measuring boards commonly used in the field.[111] Recumbent length will be measured for children 6-23 months of age, standing height will be measured for children 24-59 months of age. If bilateral oedema is present in a child's feet or if the MUAC measurement for a child (6 months and older) is less than 125 mm, then the child will be referred to the nearest health facility.

10.5 Quantitative data collection and management

Enumerators will be hired and trained prior to survey rounds, per Institutional Review Board (IRB) requirements. Trainings will be co-facilitated by Regional M&E Supervisors and NY-based M&E specialists. Enumerators will specialize in various modules of the survey tools. Surveys will be administered verbally in the local language after an informed consent process. Random household visits will be undertaken by field supervisors to ensure quality control. All questionnaires will be quality checked after enumeration and re-enumerated as needed.

Data entry clerks will use a template developed in CSPro containing a series of pre-programmed range, skip, and logic checks to minimize errors in data capture.[112] Double data entry will be employed for key variables to reduce errors in data capture. Data cleaning will be conducted concurrent to data entry using CSPro's batch edit functionality to perform an additional series of data checks. Indicators will be tabulated and data analysis completed in CSPro, Stata and R statistical packages.[113, 75]

10.6 Use of supporting data sources

The MVP also collects extensive data in real-time and at routine intervals for operational purposes. Members of the local site team collect these data during their interactions with households and facilities (e.g. clinics, schools). The tools used for collection are primarily questionnaires on mobile phones. This operational data will be used to calculate indicators that are not generated from our survey data.

Cross-checking operational data with survey data is a difficult task for several reasons. Community data is collected at the household level by extension workers during visits. These visits do not necessarily cover all households, but rather, a nonrandom subset based on personnel availability and efficiency. The questionnaires used to collect data differ from the survey tools. Therefore, data collected by extension workers can only generate proxies for the survey-based indicators.

Electronic facility data is only available at the aggregate level. Facilities generally report data to regional centers without disaggregating by residency, i.e. within or outside of the MV1. Electronic facility-based data is not available for the entire ten-year period, only the past three years. Additionally, many of the survey indicators do not directly translate to facility indicators.

11 Transparency

We will register our selection of comparison villages with RIDIE.[114] These will be time-stamped before we begin surveys, and locked until release until 2016, once surveys are complete. We wait until 2016 for public release because there is a risk that public knowledge of comparison villages before surveys are complete may compromise the survey work. However, the time-stamped nature of the registration will avoid all possibility of altering comparison village choice when outcome data are available.

The development of this research protocol involved extensive informal collaboration with many technical experts in the fields of causal inference, program evaluation, economic development, agriculture, and public health. We hope that with its publication, the broader communities will scrutinize this protocol, and provide feedback and suggestions. Though we cannot, at this stage, change our data collection procedure (due to timelines and budget constraints), we will synthesize feedback on the analysis plan and come to a decision through an interactive process of external critique.[20]

To enable replicability, we will document all variables (and their data sources) used in the matching process to select comparison villages. We will publicly release this documentation, along with matching algorithms code and geographic data. Our other data source for selection of comparison villages, the Demographic and Health Surveys (DHS), is publicly available. By 2017, final survey data, costing datasets, and process evaluation summaries will be made available in a secure online data enclave, which can be accessed by researchers after submission of a research proposal and application. Approval will involve a thorough review of the proposal by a designated research committee, signing of an end-user license agreement, and completion of human subjects research training per Institutional Review Board (IRB) requirements. Analysis code for the impact evaluation and the adequacy assessment will also be made public.

The Independent Expert Group (IEG), led by Professor Robert Black of Johns Hopkins University, graciously agreed to partake in discussions on ways to evaluate this complex and multi-dimensional project. We would like to take this opportunity to thank them for their feedback and recommendations throughout the planning phase. All errors are the responsibility of the authors alone.

In addition, we intend to engage the African Population and Health Research Center (APHRC), headquartered in Nairobi, to help support quality assurance checks of survey data collection. The APHRC will also review and improve survey data cleaning and processing systems.

12 Evaluation Timeline

We note that due to the project's time and budget constraints, impending staff departures related to the end of the ten-year project, and the ongoing transition of project responsibilities to local governments, one MV site commenced survey data collection of the household survey module a few days prior to submission of this protocol. Costing and process evaluations in the MVs also commenced prior to submission of this protocol for the same reasons. The remaining survey modules in the first site and all survey modules at the nine other sites will begin after submission of this protocol, and we plan for their data collection to be concurrent within-country between MV and comparison villages.

The final 2015 MVP evaluation will consist of five major components with staggered releases of its findings. In 2016, the adequacy assessment will be made public. The findings from the remaining four evaluation components: (1) impact evaluation, (2) cost assessment, (3) process evaluation, and (4) description of systems design and tools will be released within a year, following the 2016 adequacy assessment. These findings will be disseminated in high impact, peer-reviewed publications, project reports, implementation reviews, and presentations.

In addition to these components, the MVP will release in 2016 a package of outputs of lessons learned from the project, including: books, articles, policy briefs, MV tools and the MV Field Guide. These outputs will also describe examples of how the MVP model and its policies have been adopted, replicated, and scaled up by governments and programs outside of the core MV sites. Data from the cost assessment and process evaluation will be used to inform these subsequent outputs.

13 Ethical Issues

A number of important ethical issues have been addressed for the purposes of the study protocol:

1) **IRB approval:** Survey modules, questions and procedures employed as part of this assessment will undergo review and approval at Columbia University's Institutional Review Board (IRB). Approval will also be obtained from an accredited IRB or ethics review committee in each country.

2) **Community-level Informed consent:** In the MV1s, the MVP ground team (MVP team leader or other high level MVP staff member) will consult village leadership prior to conducting assessments in all communities in order to gain acceptance to operate within the village. In comparison villages, MVP ground team must document written attestation that village leaders allow MVP to survey their population.

3) **Individual-level informed consent:** Informed verbal consent will be obtained from all survey subjects and documented by the enumerator. For minors under 18 years, verbal consent will be obtained from their parent or guardian. For biological specimen collection (anemia and malaria), written informed consent through a signature or thumbprint will be obtained. For the process evaluation, verbal consent will be obtained from participants prior to interviews.

4) **Minors:** As per the MVP protocol, adults will consent on behalf of survey respondents under 18 years old. Adults will give signed consent for biological testing among under-5s and school-age children.

5) **Non-coerced:** Explicit mention is made on the informed consent documents that participation is completely voluntary and that refusal to participate will not affect any individual's or household's ability to receive interventions or services at the time of the survey or in the future.

6) **Confidentiality:** Confidentiality will be ensured through a number of mechanisms as per the existing quality assurance and data storage plans including: all source documents will be kept in locked cabinets at the MV site and field offices; all data sent from the villages will be encrypted when transferred or stored; data will only be stored on limited access password protected computers; all database managers and investigators will have undergone IRB-approved training; all enumerators will be trained on confidentiality and privacy protections; all data will be anonymized prior to dissemination.

7) **Referral of the seriously ill:** All under-5s with fever, who are malnourished (MUAC of < 125 mm or presence of bilateral pitting oedema) or who have moderate to severe anemia (Hemoglobin $< 110 \mathrm{g/l}$) will be immediately referred to the nearest health center for assessment. All adults and children with positive malaria RDT results will also be referred to the nearest health facility.

14 Study Protocol Limitations and Future Areas of Research

Our design and analysis have several unavoidable limitations. These include:

1) **Study design.** As discussed in Section 6.2, the impact evaluation is severely weakened by the nonrandom design and lack of comparison village data at baseline. We believe we have outlined an approach that makes the best use of available data, adjusting for observable differences between treatment and comparison groups. The impact evaluation will be unavoidably subject to errors that will not be entirely quantifiable, but with all assumptions made clear, we hope that they can be discussed transparently. Power is also a concern, because treatment is assigned at the cluster-level and lack of panel data in comparison villages prohibits adjustment for individual-level baselines.

2) **Both the intervention recipients and evaluation team are un-blinded to the intervention.** This has the potential to introduce interviewer or reporting bias and has been cited as a common challenge to community intervention trials.[115, 116] The use of standardized training of study personnel with clear standard operating procedures for field and data management systems is intended to minimize errors in survey enumeration, data capture, cleaning and analysis. During the informed consent procedure, it is made clear to respondents that participation in the evaluation has no bearing on the delivery of interventions at the household level.

3) **Population migration.** It is likely that those who leave or enter a community may differ from those who do not. High income earners may move to urban areas, which might lead to an underestimation of program impact.[21] Similarly, an influx of people to the MV1 may lessen the interventions' impact due to new residents having shorter exposure to the project or less services being delivered per person than originally planned (interventions were budgeted as lump sum, not per capital basis). We will attempt to measure population movement, as explained in Section 10.1.

4) **There are no systems in place to monitor a number of important outcomes.** These include HIV infection and TB incidence. In addition, given the evaluation design, sample sizes are insufficient to detect cluster-level (i.e. MV-level) changes in other important indicators such as maternal mortality or adolescent fertility.

5) **Self-reporting.** Many indicators are based on self-reported data either at the individual level (survey data) or the facility level (operational data). Other indicators are tested for or measured by survey enumerators (malaria, anemia, anthropometry) and not susceptible to self-reporting bias.

6) **Potential for recall bias.** Some indicators, such as child mortality rates, are themselves susceptible to recall bias. The further back in history one measures, the greater the potential for error. In addition, non-surviving births are thought to be more frequently omitted than surviving births.[117] While this would cause mortality decline to be masked or underestimated, provided these errors are randomly distributed between intervention and comparison villages, the overall effect of these errors on final risk ratios should be limited.

7) **Respondent fatigue.** Respondents may get tired of answering survey questions, causing a deterioration in data quality.

8) **Definitive statements regarding mechanism of action will be difficult to make.** The process evaluation described in Section 8 will try to qualitatively reveal some of the mechanisms of action by studying differences between faulty intervention concepts and poor delivery (implementation failure). We believe that measuring these mechanisms (i.e. mediation analysis) is likely to be very difficult with this design. The regression framework proposed by Baron and Kenny relies on many strong identifying assumptions.[118, 119] Additionally, two very interesting questions remain largely unanswerable by this evaluation:

   - **We are unable to estimate synergistic effects.** Estimating synergistic effects would answer one interesting and relevant question underlying the MVP:[5, 4] is the whole integrated package better than the sum of its parts?[120, 13, 14] However, in the absence of an experimental design including arms with all possible treatment combinations, it is very difficult to establish synergistic treatment interactions. Such studies require large sample sizes. Meeting these sample size requirements is easier with individual-level randomization, for programs such as the Ultra Poor Graduation program.[121] Qualitative assessments of synergies are difficult and insufficient. Perceived benefits due to synergies are likely to be very unreliable.
   - **We are unable to establish which component interventions are most effective.** However, it may be possible to find variables that are likely to be affected by one intervention and not the other, teasing apart which interventions work best.[122] This work will not be included in the this final evaluation, but may be in subsequent analyses.
   - **Governance and leadership.** We are unable to quantitatively estimate how much of the MVP treatment effect is due to intervention technologies versus political actors, quality of leadership, or other features of the local, district, regional, and national governance. The process evaluation described in Section 8 will attempt to qualitatively assess the extra government attention towards the project areas. The cost assessment described in Section 7 will highlight how government

financial investments have trended over the project's duration, i.e. whether local governments have increased MVP-related spending as the project approaches its ending.

9) **External validity.** Extrapolating program effects beyond the villages operating thus far is problematic for a number of reasons. First, it is difficult to get estimates of the impact of the MVP at increasing scales. It seems likely that increasing the size of the population receiving health care would have a more positive effect on each individual in the society. Conversely, if a primary reason for success of the MVs is related to institutional accountability, scaling up may prove difficult, because the attention that enforces accountability is limited. In addition, extrapolating to time periods beyond 2005-2015 is difficult both because of global changes, and because lessons were learned in 2005-2015 that would likely be used in the next implementation of the MVP model.

10) **Externalities are difficult to assess.** In Section 6.6 we discuss the difficulty of measuring some key types of externalities. We hope that the process evaluation described in Section 8 will be able to trace some of the externalities from the MVP, but we are not able to establish statistically the spillover effects from the MVP through imitation and policy changes.

11) **Sustainability is difficult to assess.** Without measurements in years following the MVP intervention, it will be difficult to assess the sustainability of its impacts. ITAD proposes assessment of sustainability via measurements five years after the end of the intervention.[21] They also mention examining variables that have lasting impacts, such as child stunting. A definitive claim about sustainability of the impact in the years following implementation cannot be made with the data available in 2015-2016. The costing and process evaluations in Sections 7 and 8, respectively, will assess the ability of the sites to maintain the MVP's delivery systems and interventions after 2015.

12) **Comparison to other projects with the same budget will be difficult to make.** An interesting, but largely unanswerable question given available data, is whether the MVP model is the best use of development money. For example, it is unknown if giving everyone in a village a cash transfer of equivalent value to the per capita cost of implementing the MVP, including interventions and management, would result in similar or better outcomes.[123, 124] To try to answer this question, one should find comparable areas and randomly assign the MVP intervention to some, and the others receive a cash transfer. This would require a budget roughly double that of the MVP, to look at, for example, 20 villages, with ten randomly assigned to the MVP and ten to cash transfers of equal value.

Contributors

Shira Mitchell - study design, data analysis, writing, figures, literature search, serves as the corresponding author, had final responsibility for the decision to submit for publication; Andrew Gelman - study design, data analysis, writing, figures; Rebecca Ross - study design, data analysis, writing, figures; Lucy McClellan - study design, data collection, data analysis, writing, figures, literature search; Uyen Kim Huynh - study design, writing, literature search; Matthew Harris - study design, data collection, data analysis, writing, literature search; Sehrish Bari - data collection, data analysis, writing, figures; Seth Ohemeng-Dapaah - data collection; Patricia Namakula - data collection; Joyce Chen - writing, figures; Sonia Ehrlich Sachs - study design, data interpretation, writing; Cheryl Palm - study design, data interpretation; Jeffrey D Sachs - study design, data interpretation, writing.

All authors contributed to writing the final report and approved the version to be published.

Declaration of interests

We declare that we have no conflicts of interest.

Acknowledgements

The authors would especially like to thank the following people for very valuable feedback and ideas: Michael Clemens, Macartan Humphreys, Elizabeth A. Stuart, Avi Feller, and Alan M. Zaslavsky.

We thank the following members (past and present) of the MVP team: Maria Muniz, Nathalie Mumaw, Xiaoyi An, Eva Quintana, Saira Qureshi, Madeline Woo, Kyle DeRosa, Ryan Marriott, May Hui, Paul Veldman, Paola Kim-Blanco, Paul Musingila, Susan Karuti, Elizabeth Katwan, Meir Brooks, Yanis Ben Amor, Jilian Sacks, Awash Teklehaimanot, Andrew Thorne-Lyman, Roseline Remans, Radhika Iyengar, Alia Karim, Vijay Modi, Belay Begashaw, Amadou Niang, Rafael Flor, Sara Lizzo, Caroline Fox, James Ossman, and MVP site teams and M&E coordinators.

We thank the following members of the Center for International Earth Science Information Network (CIESIN): Marc Levy, Linda Pistolesi, Olena Borkovska, Erin Doxsey-Whitfield, Susana B. Adamo.

We thank the following statistics and economics consultants: Susanna Makela, Abhishek Chakrabortty, Natalie Exner Dean, Keli Liu, Peng Ding, Rachael Meager, Natalie Bau, Marcia Castro, Joseph K. Blitzstein, Qixuan Chen, Jennifer Hill, Johannes Haushofer, Alberto Abadie, Christopher Blattman, and Dean Karlan.

This work will be carried out with the generous support of the Foundation to Promote Open Society, and the Governments of Japan, Korea, Mali, Senegal, and Uganda. The Earth Institute also gratefully acknowledges the support of the Bill and Melinda Gates Foundation for ongoing MDG-related activities of the Earth Institute. These funders had no direct role in the drafting of this protocol.

All mistakes are our own.

Appendices

A Outcomes of Interest

Outcomes for both the adequacy assessment and impact evaluation consist of three types: Millennium Development Goal (MDG) indicators;[10] MDG proxies; and Millennium Village Project (MVP) indicators that are relevant to systems delivery. We list these outcomes below. "I" denotes that the indicator will be used in the impact evaluation. These include all indicators computable from the survey tools (as opposed to project operational data, which is only available in treatment areas). "A" denotes that the indicator will be used in the adequacy assessment (i.e. it has a target).

Targets are defined by the UNDP,[125] unless otherwise indicated: $^{(u)}$ indicates a target defined by UNESCO;[126] $^{(w)}$ indicates a target defined by WHO;[127] and $^{(m)}$ denotes a target defined by the MVP,[128] for indicators without specific externally defined targets.

Targets are either absolute or defined relative to a 1990 national rural average or project baseline data. When 1990 national rural data are not available, we use data temporally closest to 1990, see Table 3 in C. Reference data were compiled from a variety of sources, including the World Bank, World Health Organization, the Demographic Health Surveys (DHS), and United Nations Statistics Division databases, see Table 3. See Table 4 for the village-specific targets.

A * next to the indicator number labels those indicators measured by DHS. All data come from surveys unless explicitly noted that the indicator is derived from operational data. Proxy indicators are labeled as approximations to official MDG indicators with a "≈" sign, with subscripts to distinguish different proxies for the same MDG indicator.

A.1 Millennium Development Goal Indicators and Proxies

MDG 1: Eradicate extreme poverty and hunger

1.1 **Proportion of population below 1.25 USD (PPP 2005) per day** (A, I)

Definition: proportion of all people that live below 1.25 USD (PPP 2005) per day

Target: reduce to 50% of the level in 1990

≈ 1.1 **Asset index** (I)

Definition: an indicator of household wealth that combines both asset ownership and housing characteristics, reducing to one dimension using principal component analysis (PCA). For a formal definition see equation 3 in Filmer and Pritchett.[68, 69].

Target: no target

1.2 **Poverty gap ratio** (A, I)

Definition: the Foster-Greer-Thorbecke metric $\mathsf{FGT}_1^{(f)} = \frac{1}{n}\sum_{i=1}^{q}\frac{z-y_i}{z}$ summing over all q people below the poverty line, $z = 1.25$ USD (PPP 2005), where y_i is income of person i, and n is the number of people sampled. (Indicator 1.1 is FGT_0)

Target: reduce to 50% of the level in 1990

1.8* **Prevalence of underweight children under-5 years of age** (A, I)

Definition: proportion of children under-5 with Weight-for-Age (WFA) z-score of < 2 of the WHO standard

Target: reduce to 50% of the level in 1990

$\approx_s$ 1.8 **Proportion of children under-5 years of age who are moderately or severely stunted** (A, I)

Definition: proportion of children under-5 with Length-for-Age (LFA) or Height-for-Age (HFA) z-score of < 2 of the WHO standard

Target: reduce to 50% of the level in 1990

$\approx_w$ 1.8 **Proportion of children under-5 years of age who are moderately or severely wasted** (A, I)

Defintion: proportion of children under-5 with Weight-for-Length (WFL) or Weight-for-Height (WFH) z-score < 2 of the WHO standard

Target: reduce to 50% of the level in 1990

$\approx$ 1.9 **Proportion of population below minimum level of dietary energy consumption** (I)

Definition: proportion of the population below a daily diet diversity score of five out of ten food groups[129]

Target: no target

MDG 2: Achieve universal primary education

$\approx_n$ 2.1* **Adjusted net attendance ratio in primary education** (A, I)

Definition: proportion of children of official primary school age who attend primary or higher education

Target: $\geqslant 90\%^{(m)}$

$\approx_g$ 2.1 **Gross attendance ratio for primary education** (A, I)

Definition: total attendants in primary school, regardless of age, expressed as a proportion (which can exceed 100%) of the population in the official age group corresponding to primary school

Target: $\geqslant 90\%^{(m)}$

2.2 **Proportion of pupils starting grade 1 who reach last grade of primary education** (A, I)

Definition: estimated probability of a student in grade 1 advancing to the end of primary school, subject to retention rates in the year of the survey, estimated by the *reconstructed cohort method*[10]

Target: $\geqslant 90\%^{(m)}$

MDG 3: Promote gender equality and empower women

3.1* **Gender parity in primary education** (A, I)

Definition: proportion of girl gross attendance ratio to boy gross attendance ratio [gross attendance ratio = (# of people in primary school)/(# of children of primary school age), note that this can be greater than 1]

Target: $0.97 - 1.03^{(u)}$

MDG 4: Reduce child mortality

4.1* **Under-5 mortality rate** (A, I)

Definition: estimated probability of a child dying before age five years (usually reported as deaths per 1000 live births), subject to survival rates in the five years preceding the survey[56]

Target: reduce to 33% of the level in 1990

$\approx$ 4.1 **Under-5 mortality rate, from *operational data*** (A)

Definition: estimated probability of a child dying before age five years (usually reported as deaths per 1000 live births), subject to survival rates in the 1 year preceding the survey[56]

Target: reduce to 33% of the level in 1990

4.2* **Infant mortality rate** (A, I)

Definition: estimated probability of a child dying before age 1 year (usually reported as deaths per 1000 live births), subject to survival rates in the 1 year preceding the survey[56]

Target: reduce to 33% of the level in 1990

$\approx$ 4.2 **Infant mortality rate, from *operational data*** (A)

Definition: estimated probability of a child dying before age 1 year (usually reported as deaths per 1000 live births), subject to survival rates in the 1 year preceding the survey[56]

Target: reduce to 33% of the level in 1990

4.3* **Measles immunization rate of 1 year-old children** (A, I)

Definition: proportion of children aged 12-23 months who received measles vaccine before their first birthday

Target: $\geqslant 90\%^{(w)}$

MDG 5: Improve maternal health

5.2* **Skilled birth attendance** (A, I)

Definition: proportion of women age 15-49 years with a live birth in the last 2 years who were attended by a skilled health personnel during their most recent live birth

Target: reduce proportion of unattended births by 75% of level in 1990

5.3^*_{A} **Contraception prevalence rate, ANY method** (A, I)

Definition: proportion of women age 15-49 years who are currently married or in a union where she or her partner is using ANY contraceptive method

Target: 25% nominal increase from level in $1990^{(m)}$

5.3^*_{M} **Contraception prevalence rate, MODERN method** (A, I)

Definition: proportion of women age 15-49 years who are currently married or in a union where she or her partner is using a MODERN contraceptive method

Target: 25% nominal increase from level in $1990^{(m)}$

5.5(1) **Antenatal care coverage- at least one visit with a skilled health personnel** (A, I)

Definition: proportion of women age 15-49 years with a live birth in the last two years who receivesd antenatal care at least once during their last pregnancy (with a skilled health personnel)

Target: $\geqslant 80\%^{(m)}$

5.5(4)* **Antenatal care coverage - at least four (4) visits with any provider** (A, I)

Definition: proportion of women age 15-49 years with a live birth in the last two years who received antenatal care at least four (4) times during their last pregnancy (with any provider)

Target: $\geqslant 80\%^{(m)}$

MDG 6: Combat HIV/AIDS, malaria and other diseases

$\approx_{\mathrm{p}}$ 6.1 **Proportion of pregnant women tested for HIV during their pregnancy** (A, I)

Definition: proportion of number pregnant women tested during their pregnancy to estimated number of pregnant women

Target: $\geqslant 90\%^{(m)}$

$\approx_{\mathrm{m}}$ 6.1 **Mother To Child Transmission Rate, from *operational data*** (A)

Definition: proportion of infants born to HIV-positive mothers who are shown to be HIV-positive (at either 6 weeks or 18 months)

Target: $< 5\%$

6.3* **Proportion of population aged 15-49 years with comprehensive correct knowledge of HIV/AIDS** (A, I)

Definition: proportion of population aged 15-49 years who correctly identify the two major ways of

preventing the sexual transmission of HIV (using condoms and limiting sex to one faithful, uninfected partner), who reject the two most common local misconceptions about HIV transmission and who know that a healthy looking person can transmit HIV. This indicator is usually presented for women and men separately

Target: $\geqslant 90\%^{(m)}$

6.7* **Proportion of children under-5 sleeping under insecticide-bed nets the night before** (A, I)

Definition: proportion of children under-5 who slept under an insecticide treated mosquito net the night prior to the survey

Target: $\geqslant 80\%^{(w)}$

$\approx_p$ 6.7 **Proportion of pregnant women sleeping under insecticide-bed net the night before** (A, I)

Definition: proportion of pregnant women sleeping under bed net the night prior to the survey

Target: $\geqslant 80\%^{(w)}$

$\approx_h$ 6.7 **Proportion of households with at least one insecticide-bed net** (A, I)

Definition: proportion of households with at least one insecticide-bed net

Target: $\geqslant 90\%^{(w)}$

$\approx_n$ 6.7 **Proportion of households with at least one insecticide-treated bed net per two persons sleeping in the house the night before** (A, I)

Definition: proportion of households with at least one insecticide-treated bed net per two persons sleeping in the house the night prior to the survey

Target: $\geqslant 90\%^{(w)}$

$\approx_m$ 6.8 **Among children under-5 with fever, proportion who were tested for malaria** (A, I)

Definition: among children under-5 with fever, proportion who had blood taken from a finger or heel

Target: $\geqslant 90\%^{(w)}$

$\approx_p$ 6.8 **Among children under-5 with fever tested for malaria, the proportion who received a positive test result for malaria** (I)

Definition: among children under-5 with fever tested for malaria, the proportion who received a positive test result for malaria

Target: no target

$\approx_a$ 6.8 **Among children under-5 with positive result for malaria, the proportion who took any artemisinin-based combination therapy (ACT)** (A, I)

Definition: among children under-5 with positive result for malaria, the proportion who took any artemisinin-based combination therapy (ACT)

Target: $\geqslant 90\%^{(w)}$

$\approx$ 6.9 **Death rate associated with TB, from *operational data*** (A)

Definition: proportion of TB patients who died who were diagnosed with smear positive TB in the previous 12 months

Target: 0%

$\approx$ 6.10 **Proportion of TB cases successfully treated under DOTS (new smear positive), from *operational data*** (A)

Definition: proportion of patients diagnosed with new smear positive TB in the previous 12 months

who were successfully treated as per WHO guidelines (either confirmed "cured" or who completed their treatment)

Target: $\geqslant 85\%^{(\mathfrak{m})}$

MDG 7: Ensure environmental sustainability

7.8* **Access to improved drinking water** (A, I)

Definition: proportion of all persons who use an improved source of drinking water

Target: reduce proportion without access to 50% of the level without access in 1990

7.9* **Access to improved sanitation** (A, I)

Definition: proportion of all persons who use improved sanitation facilities

Target: reduce proportion without access to 50% of the level without access in 1990

MDG 8: Develop a global partnership for development

$\approx$ 8.15 **Household access to mobile phones** (A, I)

Definition: proportion of households that own at least one mobile phone

Target: $\geqslant 80\%^{(\mathfrak{m})}$

A.2 Millennium Village Project Indicators

a. Agriculture

a.1 **Agriculture incomes** (A, I)

Definition: annualized household agriculture income (USD 2005 PPP)[130, 131, 132, 133]

Target: 100% increase of baseline levels$^{(m)}$

a.2 **Proportion of farming households using mineral fertilizer** (A, I)

Definition: proportion of farming households (all households that use any land for farming or crop production as determined by the household survey) who reported using any mineral fertilizer on farms over the past one year

Target: $\geqslant 80\%^{(m)}$

a.3 **Average amount of nitrogen (N) used by farming households** (A, I)

Definition: average amount of nitrogen (N) (in kilograms) used per farming household over the past one year

Target: $\geqslant 50$ kg of nitrogen $(\mathsf{N})^{(m)}$

a.4 **Proportion of farming households using improved seed** (A, I)

Definition: proportion of farming households who report using improved seeds on farm fields over the past one year

Target: $\geqslant 80\%^{(m)}$

b. Education

b.1 **Net attendance ratio for preschool** (A, I)

Definition: proportion of children in the age group that officially corresponds to preschooling who attend preschool

Target: $\geqslant 90\%^{(m)}$

b.2 **Net attendance ratio in grades 1-3** (A, I)

Definition: proportion of children in the age group that officially corresponds to [primary school grade] who attend primary school

Target: $\geqslant 90\%^{(m)}$

b.3 **Net intake rate for the first grade of primary school** (A, I)

Definition: new entrants in the first grade of primary education who are of the official primary school-entrance age, expressed as a proportion of the population of the same age

Target: $\geqslant 90\%^{(m)}$

c. Health

c.1 **Proportion of children under 6 months who are being exclusively breastfed** (A, I)

Definition: proportion of children under 6 months who are being exclusively breastfed, i.e. who receive only breast milk - no water or other liquids and no soft, semi-solid, or solid foods, with exception of oral rehydration solution and drops/syrups of medicines, vitamins, or minerals

Target: $\geqslant 50\%^{(m)}$

d. Infrastructure

d.1 **Proportion of MV1 primary schools that have a functional improved primary water source, from *operational data*** (A)

Definition: proportion of MV1 primary schools which have at least one functional water source pro-

tected from outside contamination (piped water, protected wells, protected springs, rainwater collection)

Target: $\geqslant 80\%^{(m)}$

d.2 **Proportion of MV1 primary schools with handwashing stations, from *operational data*** (A)

Definition: proportion of MV2 primary schools that have at least one handwashing station

Target: $\geqslant 80\%^{(m)}$

d.3 **Proportion of MV1 primary schools functional improved sex-separate toilet facilities, from *operational data*** (A)

Definition: proportion of MV1 primary schools that have at least one female and one male functional and improved toilet facility (including flush or pour-flush pit latrine, Ventilated Improved Pit latrine, pit latrine with slab, composting toilet)

Target: $\geqslant 80\%^{(m)}$

d.4 **Proportion of MV1 health facilities that have a functional improved primary water source, from *operational data*** (A)

Definition: proportion of MV1 health facilities that have at least one functional water source protected from outside contamination (piped water, protected wells, protected springs, rainwater collection)

Target: $100\%^{(m)}$

d.5 **Proportion of MV1 health facilities with handwashing stations, from *operational data*** (A)

Definition: proportion of MV1 health facilities that have at least one handwashing station

Target: $100\%^{(m)}$

d.6 **Proportion of MV1 health facilities with four (4) or more improved toilet facilities for use by anyone in the health facility, from *operational data*** (A)

Definition: proportion of MV1 health facilities that have at least one improved toilet facility (including flush or pour-flush pit latrine, Ventilated Improved Pit latrine, pit latrine with slab, composting toilet)

Target: $100\%^{(m)}$

d.7 **Proportion of MV1 health facilities where there was adequate electricity for at least 3 days in a week, from *operational data*** (A)

Definition: proportion of MV1 health facilities where there was enough electricity for lighting and use of equipment for at least 3 days in the week prior to surveying

Target: $100\%^{(m)}$

B Excluded Millennium Development Goal Indicators

The list includes indicators that are not applicable in the context of the villages (e.g. proportion of seats held by women in national parliament, proportion of urban population living in slums or the official development assistance and global market access indicators); indicators that are too difficult or costly to measure (e.g. CO2 emissions, consumption of ozone-depleting substances) or too sensitive to measure (e.g. HIV prevalence among population aged 15-24 years); indicators that are not part of the core MVP interventions (e.g. literacy rates of 15-25, since MVP education related interventions focus on primary aged children); and indicators for which there are insufficient sample sizes through population surveys (e.g. maternal mortality).

Table 2: Excluded MDG Indicators

#	Indicator
	MDG 1: Eradicate extreme poverty and hunger
1.3	Share of poorest quintile in national consumption
1.4	Growth rate of GDP per person employed
1.5	Employment-to-population ratio
1.6	Proportion of employed people living below $1 (PPP) per day
1.7	Proportion of own-account and contributing family workers in total employment
	MDG 2: Achieve universal primary education
2.3	Literacy rate of 15-24 year olds, women and men
	MDG 3: Promote gender equality and empower women
3.2	Share of women in wage employment in the non-agricultural sector
3.3	Proportion of seats held by women in national parliament
	MDG 5: Improve maternal health
5.1	Maternal mortality ratio
5.4	Adolescent birth rate
5.6	Unmet need for family planning
	MDG 6: Combat HIV/AIDS, malaria and other diseases
6.1	HIV prevalence among population aged 15-24 years
6.2	Condom use at last high risk sex
6.4	Ratio of school attendance of orphans to school attendance of non-orphans aged 10-14 years
6.5	Proportion of population with advanced HIV infection with access to antiretroviral drugs
6.6	Incidence and death rates associated with malaria
	MDG 7: Ensure environmental sustainability
7.1	Proportion of land area covered by forest
7.2	C02 emissions, total, per capita and per $1 GDP (PPP)
7.3	Consumption of ozone depleting substances
7.4	Proportion of fish stocks withing safe biological limits
7.5	Proportion of total water resources used
7.6	Proportion of terrestrial and marine areas protected

Continued on next page

Table 2 – continued from previous page

#	Indicator
7.7	Proportion of species threatened with extinction
7.10	Proportion of urban population living in slums
	MDG 8: To develop a global partnership for development
8.1	Net ODA, total and to the least developed countries, as percentage of OECD/DAC donors' gross national income
8.2	Proportion of total bilateral, sector-allocable ODA of OECD/DAC donors to basic social services (basic education, primary health care, nutrition, safe water and sanitation)
8.3	Proportion of bilateral official development assistance of OECD/DAC donors that is united
8.4	ODA received in landlocked developing countries as a proportion of their gross national incomes
8.5	ODA received in small island developing States as a proportion of their gross national incomes
8.6	Proportion of total developed country imports (by value and excluding arms) from developing countries and least developed countries, admitted free of duty
8.7	Average tariffs imposed by developed countries on agricultural products and textiles and clothing from developing countries
8.8	Agricultural support estimate for OECD countries as a percentage of their gross domestic product
8.9	Proportion of ODA provided to help build trade capacity
8.10	Total number of countries that have reached their HIPC decision points and number that have reached their HIPC completion points (cumulative)
8.11	Debt relief committed under HIPC and MDRI Initiatives
8.12	Debt service as a percentage of exports of goods and services
8.13	Proportion of population with access to affordable essential drugs on a sustainable basis
8.14	Fixed telephone lines per 100 inhabitants
8.15	Mobile cellular subscriptions per 100 inhabitants
8.16	Internet users per 100 inhabitants

C Targets per MVP village

National-rural and baseline reference data used to set 2015 relative targets

#	Senegal	Mali	Ghana	Nigeria	Ethiopia	Kenya	Uganda	Rwanda	Tanzania	Malawi
1.1 (WB)	71.0% ('94)	64.8% (2001)	63.6% ('92)	36.4% ('92)	47.5% ('95)	47.9% ('92)	60.3% ('92)	66.1% (2000)	40.8% ('92)	46.8% ('94)
1.2 (WB)	23.6 ('94)	21.2 (2000)	18.5 ('92)	14.2 ('92)	12.9 ('95)	14.9 ('94)	20.9 ('92)	24.0 (2005)	11.8 ('92)	23.4 ('98)
1.8 (WN)	27.3% ('93)	33.9% (2001)	22.6% ('99)	38.2% ('90)	43.5% (2000)	21.2% ('93)	20.0% (2001)	24.7% ('92)	26.0% ('92)	25.6% ('92)
$\approx_s$ 1.8 (WN)	40.1% ('93)	47.0% (2001)	35.1% ('99)	53.1% ('90)	58.6% (2000)	41.7% ('93)	46.1% (2001)	57.5% ('92)	50.8% ('92)	57.4% ('92)
$\approx_w$ 1.8 (WN)	10.8% ('93)	13.7% (2001)	10.9% ('99)	12.6% ('90)	12.9% (2000)	7.3% ('93)	5.2% (2001)	5.1% ('92)	8.1% ('92)	7.0% ('92)
4.1 (D)	184 ('92)	273 ('95)	149 ('93)	208 ('90)	192 (2000)	96 ('93)	159 ('95)	163 ('92)	152 ('91)	244 ('92)
4.2 (D)	87 ('92)	145 ('95)	82 ('93)	96 ('90)	115 (2000)	65 ('93)	88 ('95)	90 ('92)	97 ('91)	138 ('92)
5.2 (M,D)	28.6% ('93)	26.3% ('95)	29.5% ('93)	23.2% ('90)	2.3% (2000)	40.3% ('93)	32.3% ('95)	23.7% ('92)	34.2% ('92)	50.8% ('92)
5.3A (D)	3.3% ('93)	3.3% ('96)	15.4% ('93)	3.6% ('90)	4.3% (2000)	30.9% ('93)	12.2% ('95)	20.8% ('92)	8.4% ('92)	11.7% ('92)
5.3M (D)	1.4% ('92)	1.9% ('95)	7.4% ('93)	1.9% ('90)	3.3% (2000)	25.4% ('93)	5.1% ('95)	12.6 % ('92)	4.5% ('92)	6.0% ('92)
7.8 (U)	43% ('90)	22% ('90)	37% ('90)	30% ('90)	8% ('90)	32% ('90)	39% ('90)	66% ('90)	44% ('90)	33% ('90)
7.9 (U)	22 % ('90)	23% ('90)	4% ('90)	36% ('90)	1% ('90)	27% ('90)	40% ('90)	22% ('90)	23% ('90)	41% ('90)
a.1	$334.8	$172.7	$249.3	$165.0	$180.93	$29.2	$56.0	$32.3	$134.2	$52.9

Table 3: For a subset of indicators in Appendices A.1 and A.2, setting targets for the adequacy assessment requires either baseline project data or country-specific national-rural data before project start dates. This table contains these values and corresponding years of source for relevant indicators, used to set 2015 targets shown in Table 4. For all indicators except the agricultural indicators, A.1 and A.5, national-rural reference data is used to set 2015 targets. Data sources include: (WB) World Bank PovCal: `http://iresearch.worldbank.org/PovcalNet/index.htm?3`, (WN) WHO NLIS: `http://apps.who.int/nutrition/landscape/search.aspx?dm=52&countries=`, (W) WHO: `http://apps.who.int/ghodata/?theme=country`, (D) DHS: `http://www.measuredhs.com/data/available-datasets.cfm` or `http://www.statcompiler.com/`, (U) UNSTATS: `http://mdgs.un.org/unsd/mdg/Default.aspx`, (M) MICS: `http://www.unicef.org/statistics/index_24302.html`

2015 targets, by Millennium Village, for indicators with relative targets

#	Senegal	Mali	Ghana	Nigeria	Ethiopia	Kenya	Uganda	Rwanda	Tanzania	Malawi
1.1	35.5%	32.4%	31.8%	18.2%	23.8%	24.0%	30.2%	33.1%	20.4%	23.4%
1.2	11.8	10.6	9.3	7.1	6.5	7.5	10.5	12.0	5.9	11.7
1.8	13.7%	17.0%	11.3%	19.1%	21.8%	10.6%	10.0%	12.4%	13.0%	12.8%
$\approx_s$ 1.8	20.1%	23.5%	17.6%	26.6%	29.3%	20.9%	24.4%	28.8%	25.4%	28.7%
$\approx_w$ 1.8	5.4%	6.9%	5.5%	6.3%	6.5%	3.7%	2.6%	2.6%	4.1%	3.5%
4.1	61	91	50	69	64	32	53	54	51	81
4.2	29	48	27	32	38	22	29	30	32	46
5.2	17.85%	18.43%	17.63%	19.2%	24.43%	14.93%	16.93%	19.08%	16.45%	12.3%
5.3A	28.3%	28.3%	40.4%	28.6%	29.3%	55.9%	37.2%	45.8%	33.4%	36.7%
5.3M	26.4%	26.9%	32.4%	26.9%	28.3%	50.4%	30.1%	37.6%	29.5%	31.0%
7.8	71.5%	61%	68.5%	65%	54%	66%	69.5%	83%	73%	66.5%
7.9	61%	61.5%	52%	68%	50.5%	63.5%	70%	61%	61.5%	70.5%
a.1	$669.7	$345.3	$498.6	$329.9	$361.86	$58.3	$112.0	$64.5	$268.3	$105.9

Table 4: A subset of indicators in Appendices A.1 and A.2 have targets based on baseline project data or country-specific national-rural data before project start dates, whose values are in Table 3. This table contains these 2015 targets.

D Adequacy Assessment - Sample Size Considerations

In Figures 6, 7, and 8 we plot confidence interval widths for eight MDG indicators. For binary indicators (all except indicator 2.2), we form Agresti-Coull confidence interval widths.[134] For MDG indicator 2.2, we use Greenwood's variance estimation with reconstructed cohort data.[10, 135]

We consider three sample sizes, 100 (blue), 300 (red), or 600 (green), with differing units for each indicator (e.g. children aged 12-23 months, for the measles immunization rate variable). In Figures 6 and 7 we make the simplifying assumption that there is no intra-house correlation. In Figure 8, we relax this assumption by simulating data with two children per household, using an intra-house correlation estimated from 2013 project data. We do not adjust for intra-village correlation. The x-axes represent true values for the indicators, and the curves the width of the confidence interval for each possible indicator value. For each indicator, the targets are marked in a black dashed line, while estimates from the fifth year of the project, averaged across all sites, are marked in a purple dotted line. Though the confidence interval width curves are identical for all indicators (except 2.2), the separate plots are used to show different plausible true values for each indicator (using the targets and fifth year project data).

The following indicators are included: 1.1, Proportion of population below 1.25 USD (PPP 2005) per day, averaged target = 33.23%, average 2010 estimate = 68.2%.; MDG 1.8, Prevalence of underweight children under-five years of age, averaged target = 14.17%, average 2010 estimate = 15.8%; MDG 2.1 Net attendance ratio in primary education, target $\geqslant$ 90%, average 2010 estimate = 74.2%; MDG 2.2 Proportion of pupils starting grade 1 who reach last grade of primary education, target $\geqslant$ 90%, average 2010 estimate = 75.4%; MDG 4.3 Measles immunization rate of 1 year-old children, target $\geqslant$ 90%, average 2010 estimate = 86.6%; MDG 5.2 Skilled birth attendance, target $\geqslant$ 70%, average 2010 estimate = 64.9%; MDG 6.7 Children under 5 sleeping under insecticide-bed nets, target $\geqslant$ 80%, average 2010 estimate = 57.3%; MDG 7.8 Access to improved drinking water, averaged target = 67.8%, average 2010 estimate = 0.789.

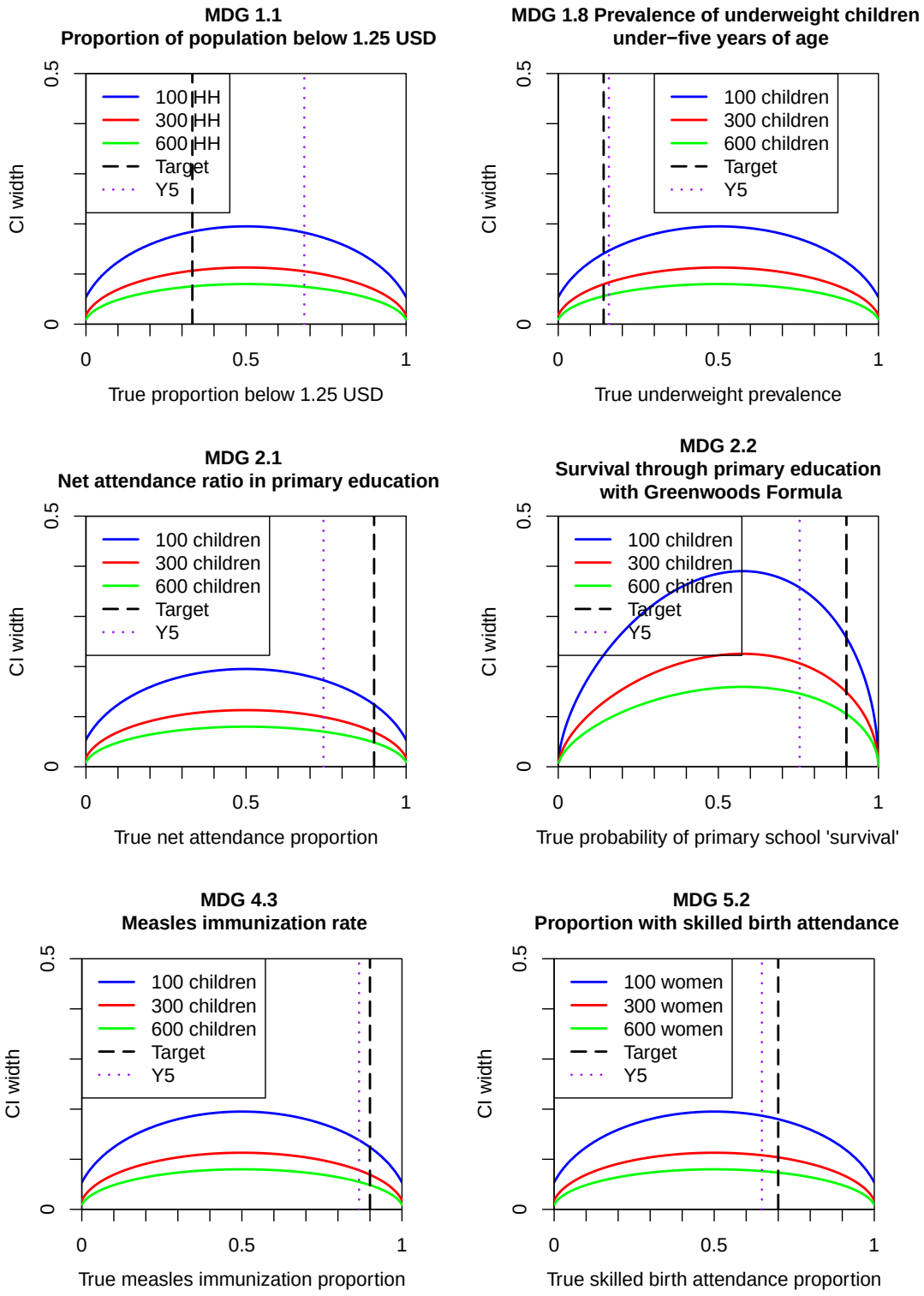


Figure 6: Confidence interval widths for MDG indicators 1.1, 1.8, 2.1, 2.2, 4.3, and 5.2. For each indicator, the targets are marked in a black dashed line, while estimates from the fifth year of the project, averaged across all sites, are marked in a purple dotted line.

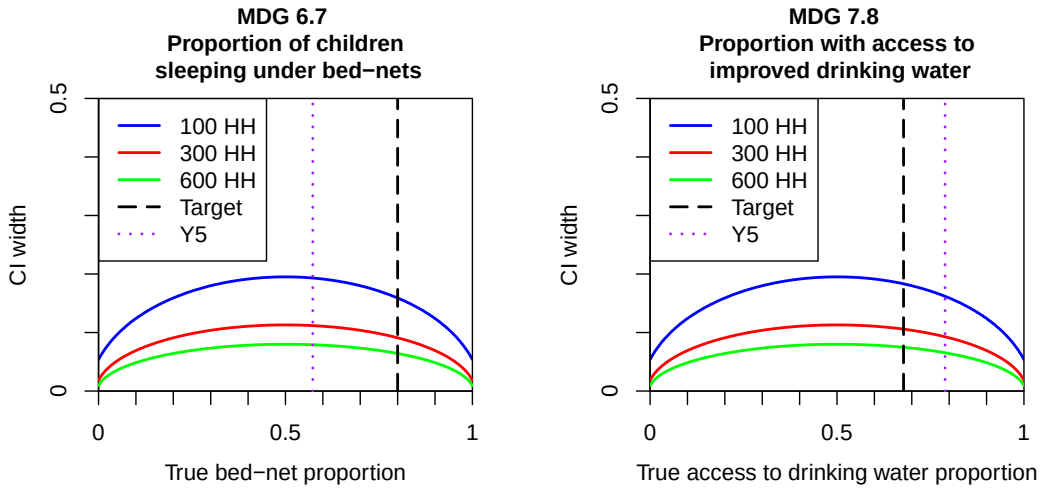


Figure 7: Confidence interval widths for MDG indicators 6.7 and 7.8. For each indicator, the targets are marked in a black dashed line, while estimates from the fifth year of the project, averaged across all sites, are marked in a purple dotted line.

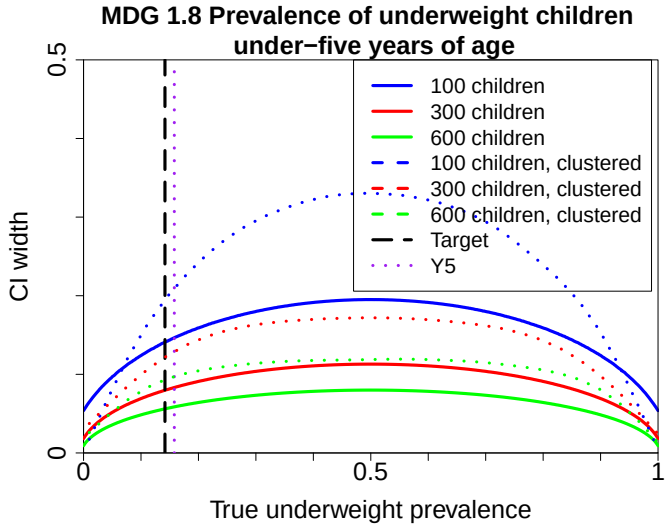


Figure 8: Confidence interval widths for MDG indicator 1.8, including the curves from Figure 6 and curves formed using simulated data with two children per household and an intra-house correlation estimated from 2013 project data. The target is marked in a black dashed line, while estimates from the fifth year of the project, averaged across all sites, are marked in a purple dotted line.

References

[1] United Nations GA. 55/2. United Nations Millennium Declaration; 2000. available from http://www.un.org/millennium/declaration/ares552e.htm.

[2] Sachs JD, McArthur JW. The Millennium Project: a plan for meeting the Millennium Development Goals. The Lancet. 2005 January;365(9456):347–353.

[3] Sanchez P, Palm C, Sachs JD, Denning G, Flor R, Harawa R, et al. The African Millennium Villages. Proceedings of the National Academy of Sciences. 2007;104(43):6775–80.

[4] Sachs J. Rapid Victories against Extreme Poverty. Scientific American. 2007 April;296(4):34.

[5] Sachs JD, McArthur JW, Schidt-Traub G, Kruk M, Bahadur C, McCord G. Ending Africa's Poverty Trap. Brookings Papers on Economic Activity. 2004;1:117–240. Available from: http://www.unmillenniumproject.org/documents/BPEAEndingAfricasPovertyTrapFINAL.pdf.

[6] Clemens MA, Demombynes G. When does rigorous impact evaluation make a difference? The case of the Millennium Villages. Journal of Development Effectiveness. 2011 September;3(3):305–339.

[7] Kremer M, Holla A. Improving Education in the Developing World: What Have We Learned from Randomized Evaluations? Annual Review of Economics;(One):513–542. Available from: http://scholar.harvard.edu/files/kremer/files/annual_review_kremer_holla_2009.pdf.

[8] Cohen J, Dupas P. Free Distribution or Cost-sharing? Evidence from a randomized malaria prevention experiment. The Quarterly Journal of Economics. 2010 February;125(1):1–45. Available from: http://www.stanford.edu/~pdupas/CohenDupas.pdf.

[9] Schofield H. The Economic Costs of Low Caloric Intake: Evidence from India; 2014. Working paper. Available from: http://scholar.harvard.edu/files/hschofield/files/schofield_calories_and_productivity_2014.01.27.pdf.

[10] UN Millennium Project; 2014. Available from: http://mdgs.un.org/unsd/mdg/Metadata.aspx.

[11] Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. International Epidemiological Association. 1999;28:10–18.

[12] McArthur JW, Pronyk PM, Sachs JD. Designing, Implementing and Evaluating Complex, Goal-Oriented Adaptive Interventions in the Millennium Villages; 2011. Available from: http://www.csae.ox.ac.uk/conferences/2011-EdiA/plenaries/csae-conf2011-panel2-McArthur.pdf.

[13] Blattman C. Am I actually sticking up for the Millennium Villages?; 2009. Blog post. Available from: http://chrisblattman.com/2009/10/15/am-i-actually-sticking-up-for-the-millennium-villages/.

[14] Blattman C. Evaluating the Millennium Villages: The saga continues; 2010. Blog post. Available from: http://chrisblattman.com/2010/10/28/evaluating-the-millennium-villages-the-saga-continues/.

[15] Clemens MA, Demombynes G, Kenny C, Minard S, Naudet J, Peccoud R. The Collision of Development Goals and Impact Evaluation; 2012. Working paper. Available from: http://www.afd.fr/webdav/shared/PORTAILS/PUBLICATIONS/EUDN/EUDN2012/interventions/Article-Michael-CLEMENS.pdf.

[16] Butler D. Poverty project opens to scrutiny. Nature. 2012;486:165–166.

[17] Nature editorial. With transparency comes trust. Nature. 2012 May;485(147). Available from: http://www.nature.com/nature/journal/v485/n7397/full/485147a.html.

[18] The Economist. Millennium Bugs: Jeffrey Sachs and the Millennium Villages; 2012. Available from: http://www.economist.com/blogs/feastandfamine/2012/05/jeffrey-sachs-and-millennium-villages [cited March 2014].

[19] Starobin P. Does it take a village?; 2013. Available from: http://www.foreignpolicy.com/articles/2013/06/24/does_it_take_a_village.

[20] Clemens MA, Demombynes G. The New Transparency in Development Economics: Lessons from the Millennium Villages Controversy. Washington DC: Center for Global Development; 2013. 342.

[21] ITAD evaluation for Northern Ghana. Impact Evaluation of a New Millennium Village in Northern Ghana: Initial Design Document. UK Department for International Development; 2013.

[22] Mitchell S, Ross R, Makela S, Stuart EA, Feller A, Zaslavsky AM, et al. Causal inference with small samples and incomplete baseline for the Millennium Villages Project; 2015. Working Paper.

[23] MVP. Harvests of development: the Millennium Villages after three years. New York, NY: The Earth Institute at Columbia University; 2010.

[24] Pronyk PM, Muniz M, Nemser B, Somers MA, McClellan L, Palm CA, et al. The effect of an integrated multisector model for achieving the Millennium Development Goals and improving child survival in rural sub-Saharan Africa: a non-randomised controlled assessment. The Lancet. 2012 June;379(9832):2179–2188.

[25] Wanjala BM, Muradian R. Can Big Push Interventions Take Small-Scale Farmers out of Poverty? Insights from the Sauri Millennium Village in Kenya. World Development. 2013;45:147–160.

[26] Gelman A, Hill JL. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, NY: Cambridge University Press; 2007.

[27] Bump JB, Clemens MA, Demombynes G, Haddad L. Concerns about the Millennium Villages project report. The Lancet. 2012;379(9830):1945.

[28] Rosenbaum PR. The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. Journal of the Royal Statistical Society A. 1984;147(5):656–666.

[29] Royston P, Altman D, Sauerbrei W. Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea. Statistics in Medicine. 2006;25:127–141.

[30] Gelman A, Park DK. Splitting a Predictor at the Upper Quarter or Third and the Lower Quarter or Third. The American Statistician. 2008;62(4):1–8.

[31] Imbens GW, Rubin DB. Causal Inference in Statistics and Social Sciences: An Introduction. New York, NY: Cambridge University Press; 2015.

[32] Rubin DB. Inference and missing data. Biometrika. 1976;63:581–592.

[33] Rubin DB. Bayesian inference for causal effects: The role of randomization. Annals of Statistics. 1978;6:34–58.

[34] Rubin D. For Objective Causal Inference, Design Trumps Analysis. The Annals of Applied Statistics. 2008;2(3):808–840.

[35] Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. Statistical Science. 1999 February;14(1):29–46.

[36] Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics. 2005;61:962–972.

[37] Angrist JD, Pischke JS. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press; 2009.

[38] Rubin DB. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. Biometrics. 1973 March;29(1):185–203.

[39] Rubin DB, Thomas N. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. Journal of the American Statistical Association. 2000;95(450):573–585.

[40] Ho DE, Imai K, King G. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis. 2007;15:199–236.

[41] Kreif N, Grieve R, Radice R, Sekhon JS. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation; 2011. Paper presented at the Causal Inference Group Meeting at the Harvard School of Public Health. Available from: `http://www.lshtm.ac.uk/php/hsrp/reducing-selection-bias/output/regression_adjusted_matching_and_double_robust_methods.pdf`.

[42] Abadie A, Imbens GW. Bias-Corrected Matching Estimators for Average Treatment Effects. Journal of Business and Economic Statistics. 2011;29(1):1–11.

[43] Robins JM, Rotnitzky A, der Laan MJV. Comment on the Murphy and Van der Vaart article, "On profile likelihood.". Journal of the American Statistical Association. 2000;95:431–435.

[44] Robins JM, Rotnitzky A. Comment on the Bickel and Kwon article, "On double robustness.". Statistica Sinica. 2001;11:920–936.

[45] Dehejia RH, Wahba S. Causal effects in Nonexperimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association. 1999 December;94(448):1053–1062.

[46] Dehejia RH. Practical propensity score matching: a reply to Smith and Todd. Journal of Econometrics. 2005;125:355–364.

[47] Shadish WR, Clark MH, Steiner PM. Can nonrandomised experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. Journal of the American Statistical Association. 2008 December;103(484):1334–1343.

[48] Dixon J, Gulliver A, Gibbon D. Farming Systems and Poverty: Improving Farmers' Livelihoods in a Changing World. Rome and Washington DC: FAO and the World Bank; 2001.

[49] Joint Research Centre: Land Resource Management Unit. Travel time to major citis: A global map of Accessibility;. Available from: http://bioval.jrc.ec.europa.eu/products/gam/sources.htm [cited October 2014].

[50] ISRIC: World Soil Information. Soil property maps of Africa at 1 km;. Available from: http://www.isric.org/data/soil-property-maps-africa-1-km [cited October 2014].

[51] GPWv3. Socioeconomic Data and Applications Center (SEDAC): Gridded Population of the World (GPW), v3;. Available from: http://sedac.ciesin.columbia.edu/data/collection/gpw-v3 [cited October 2014].

[52] GADMv2. Global Administrative Areas Databse (GADMv2); 2012. Available from: http://www.gadm.org [cited 2015].

[53] IRI/LDEO. IRI/LDEO Climate Data Library;. Available from: http://iridl.ldeo.columbia.edu/ [cited October 2014].

[54] The CGIAR Consortium for Spatial Information (CGIAR-CSI). SRTM 90m Digital Elevation Data;. Available from: http://srtm.csi.cgiar.org/ [cited October 2014].

[55] MVP. Survey Enumeration Manual: Guidelines for enumerators, field supervisors, and data managers. New York, NY: Millennium Villages Project; 2011. Available from: https://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP_Y5_Enumeration_Manual.pdf.

[56] Rutstein SO, Rojas G. Guide to DHS Statistics. Demographic and Health Surveys, Calverton, Maryland: Demographic and Health Surveys; 2006. Available from: http://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_29Oct2012_DHSG1.pdf.

[57] Measure DHS/ICF International. Sampling and Household Listing Manual: Demographic and Health Surveys Methodology. Measure DHS; 2012. Available from: http://www.measuredhs.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

[58] ; 2014. Available from: http://www.measuredhs.com/faq.cfm.

[59] Ghosh M, Rao JNK. Small Area Estimation: An Appraisal. Statistical Science. 1994;9(1):55–76.

[60] Ghosh M, Natarajan K. Small Area Estimation: A Bayesian Perspective. In: Ghosh S, Dekker M, editors. Multivariate Analysis, Design of Experiments and Survey Sampling. New York, NY: Wiley; 1999. p. 69–92.

[61] Nadram B. Bayesian Generalized Linear Models for Inference About Small Areas. In: Rey D, Ghosh SK, Mallick BK, editors. Generalized Linear Models. Boca Raton: CRC Press; 2000. p. 89–109.

[62] Rao JNK. Small Area Estimation. Hoboken, New Jersey: John Wiley and Sons; 2003.

[63] Jiang J, Lahiri P. Mixed Model Prediction and Small Area Estimation. Test. 2006;15(1):1–96.

[64] National Geospatial-Intelligence Agency. World Geodetic System 1984;. Available from: `http://www.unoosa.org/pdf/icg/2012/template/WGS_84.pdf`.

[65] ERSI. ArcGIS Desktop: Release 10.2. Redlands, CA: Environmental Systems Research Institute; 2013.

[66] Stuart EA, Rubin DB. Matching With Multiple Control Groups With Adjustment for Group Differences. Journal of Educational and Behavioral Statistics. 2008 September;33(3):279–306.

[67] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

[68] Filmer D, Pritchett LH. Estimating Wealth Effects Without Expenditure Data - Or Tears: An Application to Educational Enrollments in States of India. Demography. 2001 February;38(1):115–132.

[69] Michelson H, Muniz M, DeRosa K. Measuring Socio-economic Status in the Millennium Villages: The Role of Asset Index Choice. The Journal of Development Studies. 2013;49(7):917–935.

[70] Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science. 2010;25(1):1–21.

[71] Humphreys M, de la Sierra RS, van der Windt P. Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration. Political Analysis. 2013;21:1–20.

[72] Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time; 2013. Available from: `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

[73] Gelman A, Hill JL, Yajima M. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness. 2012;5:189–211.

[74] Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 1.3; 2013. Available from: `http://mc-stan.org/`.

[75] R Development Core Team. The R project for statistical computing; 2014. Available from: `http://www.r-project.org/`.

[76] Sanchez P, Swaminathan MS, Dobie P, Yuksel N. Halving hunger: it can be done. UN Millennium Project Task Force on Hunger; 2005.

[77] World Health Organization. A Summary of the Findings of the Commission on Macroeconomics and Health. World Health Organization CMH support unit; 2003.

[78] UN Millennium Project. Investing in development: A practical plan to achieve the milliennium development goals. New York: UN Millennium Project; 2005. Available from: `http://www.unmillenniumproject.org/reports/fullreport.htm` [cited 20 September 2012].

[79] Heitjan DF, Moskowitz AJ, Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. Health Economics. 1999;8:191–201.

[80] Heitjan DF, Moskowitz AJ, Whang W. Problems with interval estimation of the incremental cost-effectiveness ratio. Medical Decision Making. 1999;19:9–15.

[81] Gelman A. Fundamental difficulty of inference for a ratio when the denominator could be positive or negative; 2011. Available from: `http://andrewgelman.com/2011/06/21/inference_for_a/` [cited March 2015].

[82] Catterall JS. Economic evaluation of public programs. New directions for program evaluation. 1985;1985(26):99–103.

[83] Ahren P. Economic evaluation methods in community planning. Swedish Council for Building Research; 1976.

[84] Pushpangadan P. Conservation and economic evaluation of biodiversity. Pushpangadan P, Ravi K, Santhosh V, editors. India: Science Publishers Inc; 1997.

[85] Rahman ML, Alam MF. An economic evaluation of some credit programmes designed for the small farmers and landless poor in Bangladesh. Dhaka. Bangladesh Agricultural University: Bureau of Socioeconomic Research and Training; 1987.

[86] Hutchinson BG. The economic evaluation of urban transportation investments. London Centre for Environmental Studies; 1969.

[87] Grieve R, Cairns J, Thompson SG. Improving Costing Methods in Multicentre Economic Evaluation: the Use of Multiple Imputation for Unit Costs. Health Economics. 2009;10(10002):1532.

[88] Schulenburg JM. The influence of economic evaluation studies on health care decision-making: a European survey. Amsterdam, Holland: IOS Press; 2000.

[89] Wordsworth S, Ludbrook A, Caskey F, Macleod A. Collecting Unit Cost Data in Multicentre Studies: Creating Comparable Methods. The European Journal of Health Economics. 2005;6(1):38–44.

[90] Guba EG, Lincoln YS. Fourth generation evaluation. Newbury Park, CA: Sage Publications; 1989.

[91] Oakley A, Strange V, Stephenson J, Forrest S, Monteiro H, RIPPLE Study Team. Evaluating Processes: A case study of a randomized controlled trial of sex education. Evaluation. 2004;10(4):440–462.

[92] Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. British Medical Journal. 2008 June;336(7656):1281–3.

[93] Wedeen L. Reflection on Ethnographic Work in Political Science. Annual Review of Political Science. 2010;13:255–272.

[94] Denzin NK, Lincoln YS, editors. Handbook of Qualitative Research. 1st ed. Sage Publications; 1993.

[95] Rabinow P. Reflections on Fieldwork in Morocco. University of California Press; 1977.

[96] Stocking GW. Observers Observed: Essays on Ethnographic Fieldwork. University of Wisconsin Press; 1983.

[97] Fabian J. Time and the Other: How Anthropology Makes its Object. Columbia University Press; 2002.

[98] Kumar K. Conducting Key Informant Interviews in Developing Countries. Agency for International Development; 1989.

[99] QSR International. NVivo Qualitative Data Analysis Software: Version 8. Cambridge, MA, USA: QSR International; 2008.

[100] Brady HE, Collier D. Rethinking Social Inquiry: Diverse Tools, Shared Standards. Rowman and Littlefield; 2010.

[101] Sustainable Engineering Lab. SharedSolar; 2014. Available from: `http://sharedsolar.org/` [cited October 2014].

[102] Sustainable Engineering Lab. formhub; 2012-2013. Available from: `http://formhub.org/`.

[103] Dimagi, Inc . CommCare; 2014. Available from: `http://www.commcarehq.org/home/`.

[104] Lohr SL. Sampling: Design and Analysis. 2nd ed. Cengage Learning; 2010.

[105] Särndal CE, Swensson B, Wretman J. Model Assisted Survey Sampling. New York: Springer-Verlag; 1992.

[106] HemoCue Worldwide. HemoCue. Angelholm, Sweden: Radiometer Group; 2014. Available from: `http://www.hemocue.com/en/products/hb-301-kit`.

[107] Access Bio Inc. Somerset, NJ, USA; 2012. Available from: `http://www.accessbio.net/english/product/01.asp` [cited October 2014].

[108] Koita OA, Doumbo OK, Ouattara A, Tall LK, Konaré A, Diakité M, et al. False-negative rapid diagnostic tests for malaria and deletion of the histidine-rich repeat region of the hrp2 gene. American Journal of Tropical Medicine and Hygiene. 2012 February;86(2):194–198.

[109] WHO product testing. Malaria Rapid Diagnostic Test Performance: Results of WHO product testing of malaria RDTs: Round 5 (2013). World Health Organization; 2014. ISBN 978 92 4 150755 4. Available from: `http://apps.who.int/iris/bitstream/10665/128678/1/9789241507554_eng.pdf`.

[110] World Health Organization. Information note on recommended selection criteria for procurement of malaria rapid diagnostic tests (RDTs). World Health Organization; 2014. Available from: `http://www.who.int/malaria/publications/atoz/rdt_selection_criteria/en/`.

[111] UNICEF. Height/Length Measuring Boards. UNICEF;. 18. Available from: http://www.unicef.org/supply/files/Height_Length_Measuring_Boards.pdf.

[112] US Census Bureau. CSPro: Census and Survey Processing System, Version 5.0.3. Washington DC, USA: US Census Bureau, Macro International, and Serpo, S.A.; 2013. Available from: http://www.census.gov/population/international/software/cspro/.

[113] StataCorp. Stata Statistical Software: Release 12. College Station, Texas: StataCorp LP; 2011.

[114] International Initiative for Impact Evaluation. The Registry for International Development Impact Evaluations; 2013 [cited July 2014]. Available from: http://ridie.3ieimpact.org/.

[115] Donner A, Klar N. Pitfalls and controversies in cluster randomization trials. American Journal of Public Health. 2004 March;94(3):416–422.

[116] Sorenson G, Emmons K, Hunt MK, Johnston D. Implications of the Results of Community Intervention Trials. Annual Review of Public Health. 1998;19:379–416.

[117] Bicego G, Ahmad OB. Demographic and Health Surveys: Infant and Child Mortality. Calverton, Maryland: Demographic and Health Surveys; 1996. Available from: https://dhsprogram.com/pubs/pdf/CS20/00FrontMatter00.pdf.

[118] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986;51:1173–82.

[119] Green DP, Ha SE, Bullock JG. Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose. The ANNALS of the American Academy of Political and Social Science. 2010;628(1):200–208.

[120] Blattman C. Do the Millenium Villages work?; 2007. Blog post. Available from: http://chrisblattman.com/2007/12/28/do-the-millenium-villages-work/.

[121] Innovations for Poverty Action. Ultra Poor Graduation Program;. Available from: http://www.poverty-action.org/ultrapoor [cited March 2014].

[122] Duflo E, Glennerster R, Kremer M. Using randomization in development economics research: a toolkit. In: Schultz TP, Strauss J, editors. Handbook of Development Economics. vol. 4. North Holland; 2008. p. 3895–3962.

[123] Haushofer J, Shapiro J. Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya; 2013. Working paper.

[124] Blattman C, Fiala N, Martinez S. Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda; 2013. Quarterly Journal of Economics, Forthcoming.

[125] United Nations Development Group. Indicators for monitoring the Millennium Development Goals: Definitions, rationale, concepts and sources. New York, NY: United Nations; 2003. ST/ESA/STAT/SER.F/95.

[126] UNESCO Institute for Statistics. Global Education Digest 2010: Comparing Education Statistics Across the World. Montreal, Quebec: UNESCO Institute for Statistics; 2010. Available from: http://www.uis.unesco.org/Library/Documents/GED_2010_EN.pdf.

[127] World Health Organization. Health situation and trends assessment;. Available from: http://www.searo.who.int/entity/health_situation_trends/data/mdg/measles/en/.

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

[129] FAO. Introducing the Minimum Dietary Diversity - Women (MDD-W): Global Dietary Diversity Indicator for Women. FAO; 2014. Available from: http://www.fao.org/fileadmin/templates/nutrition_assessment/Dietary_Diversity/Minimum_dietary_diversity_-_women__MDD-W__Sept_2014.pdf.

[130] Singh I, Squire L, Strauss J. Agricultural Household Models: Extensions and Applications. Baltimore, Maryland: Johns Hopkins University Press for the World Bank; 1986.

[131] Grosh ME, Glewwe P. A Guide to Living Standards Measurement Study Surveys and their Datasets? Washington: The World Bank; 1995. LSM120.

[132] Deaton AS. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy (World Bank). Baltimore, Maryland: Johns Hopkins University Press for the World Bank; 1997.

[133] UNECE. Rural Households' Livelihood and Well-Being: Statistics on Rural Development and Agriculture Household Income. New York and Geneva: United Nations Economic Commission for Europe; 2007. E.07.II.E.14. Available from: http://www.fao.org/docrep/015/am085e/am085e.pdf.

[134] Agresti A, Coull BA. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician. 1998 May;52(2):119–126.

[135] Collett D. Modelling Survival Data in Medical Research. Second edition ed. CRC Press; 2003.