Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford

OPHI WORKING PAPER NO. 130

Impact of the SADA-Northern Ghana Millennium Village Project on Multidimensional Poverty: A Comparison of Dashboard and Index Approaches

Edoardo Masset* and Jorge García Hombrados**

October 2019

Abstract

This paper assesses the impact of the SADA-Northern Ghana Millennium Village Project (MVP) on multidimensional poverty using dashboard and index approaches. Using a unique, large dataset that spans five years and contains data on multiple welfare indicators, we estimate the impact of MVP on the Millennium Development Goals (MDGs) and on the global multidimensional poverty index (global MPI). We find that the project had a limited impact on the MDGs and yet a positive impact on the global MPI. We assess the robustness of the impact of MVP on the global MPI, and we conclude that it was largely driven by the sensitivity of the index to changes in a few MDG indicators. We conclude that the MVP had a limited impact on welfare and that the global MPI should be used with caution in the evaluation of development programmes.

Keywords: multidimensional poverty; Alkire-Foster method; Millennium Village Project

JEL Classification: I32, C33

*CEDIL at the London School of Hygiene and Tropical Medicine, UK, Email: Edoardo.Masset@lidc.ac.uk.
**Universidad Autonoma de Madrid and Max Planck Institute for Demographic Research, Email: jorge.garciah@uam.es.

This study has been prepared within the OPHI theme on multidimensional measurement.

Acknowledgments
The impact evaluation of the Millennium Villages Project in northern Ghana 2012–2017 was funded by the UK Department for International Development (DFID) and was carried out by Itad (www.itad.com) in partnership with the Institute of Development Studies (IDS) (www.ids.ac.uk) and PDA-Ghana (www.pdaghana.com). We thank the Earth Institute at Columbia University for collecting and providing the data used in this paper. The design of the evaluation and the analysis of the data were informed by an evaluation advisory group (comprising the Earth Institute, DFID and independent advisors) and by an independent peer review group led by the International Initiative for Impact Evaluation (3ie). Andy Mckay, Stephen Jenkins, and the participants in a workshop on multidimensional impact evaluation organised by OPHI and held in Oxford on 1–2 March 2018 provided comments to an earlier version of this paper. The contents of the paper are the responsibility of the named authors and do not reflect the views of DFID, the UK government and other institutions, or individuals who were involved in the evaluation.

Citation: Masset, E. and García Hombrados, J. (2019): 'Impact of the SADA-Northern Ghana Millennium Village Project on multidimensional poverty: A comparison of dashboard and index approaches' *OPHI Working Paper* 130, University of Oxford.

The Oxford Poverty and Human Development Initiative (OPHI) is a research centre within the Oxford Department of International Development, Queen Elizabeth House, at the University of Oxford. Led by Sabina Alkire, OPHI aspires to build and advance a more systematic methodological and economic framework for reducing multidimensional poverty, grounded in people's experiences and values.

Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK

Tel. +44 (0)1865 271915 Fax +44 (0)1865 281801
ophi@qeh.ox.ac.uk http://www.ophi.org.uk

ISSN 2040-8188 ISBN 978-19-1229-121-2

1 Introduction

In this paper we assess the impact of the SADA[1]-Northern Ghana Millennium Village Project (MVP) on the Millennium Development Goals (MDGs) and on multidimensional poverty. The SADA-MVP was one of 15 Millennium Villages established in Sub-Saharan Africa by the Millennium Promise, the Earth Institute at Columbia University, and the United Nations Development Programme. The MVP was one of the most influential and debated development projects implemented by the international community in the last 15 years. The project was endorsed by the UN Secretary General, as well as numerous prime ministers, philanthropists, academics, and celebrities. It was for 10 years the UN's flagship anti-poverty programme. The SADA-MVP was implemented between 2012 and 2016 in two districts of Northern Ghana with funding from the UK Department for International Development (DFID).

We first assess the impact of MVP on the MDGs using a dashboard approach. In this approach, impacts on different outcomes are estimated and presented separately. Dashboard approaches have several advantages but struggle to make sense of the data when there are many outcomes or when there are conflicting results. With many outcomes, an index that summarises all impacts is appealing. Many indices of deprivation have been proposed in the literature, and in this paper we employ the global multidimensional poverty index of Alkire and Santos (2014) (global MPI for brevity), which is based on the methodology for multidimensional poverty measurement outlined in Alkire and Foster (2011).

We employ the global MPI as a 'found' poverty index (Cartwright and Bradburn, 2011) rather than constructing an ad-hoc index based on the MDG indicators, as was done in other evaluations of development interventions (Loschmann et al., 2015; Song and Imai, 2018; Mitchell and Macció, 2018). The construction of an ad-hoc index involves a number of arbitrary choices. Its results are difficult to interpret and cannot be compared across studies using other indices – so knowledge does not accumulate. On the other hand, the global MPI can be easily calculated using data on a relatively small number of indicators. It was adopted by the UNDP in 2010 for the measurement of global poverty in the Human Development Report series (UNDP, 2010), and it is currently used for ranking countries and regions in

[1]The Savannah Accelerated Development Authority (SADA) was an agency of the Government of Ghana established in 2010 with the goal of coordinating development efforts in the northern regions of Ghana. SADA was involved in the design and implementation of the MVP in Northern Ghana and the project came to be known as the SADA-Northern Ghana Millennium Village Project. SADA discontinued operations and ceased to exist in 2016.

national and international comparisons. Changes in the global MPI can be interpreted by benchmarking to changes in national trends. The global MPI has the potential to become an accepted metric in the evaluation of development programmes. To our knowledge, it has not been used in policy evaluation before and this is the first study employing the global MPI to assess the impact of an intervention.

Our study finds that the MVP had a limited impact on the MDGs and that it did not improve key welfare indicators such as expenditure poverty, undernutrition, and child mortality. However, the project had a large impact on the global MPI. These contrasting results offer an opportunity to compare dashboard and index approaches in the evaluation of social policies. The global MPI has been criticised for being too sensitive to changes in one or few dimensions of well-being and for being unable to account for changes in the distribution of deprivations (Pattanaik and Xu, 2018). On the other hand, the global MPI has been praised for being able to respond to changes in deprivations for people who are deprived in several dimensions (Aaberge and Brandolini, 2015). Our sensitivity analysis suggests that the positive impact of MVP on the global MPI is mainly driven by improvements in just two MDG indicators. In light of this, there is no conflict between the large impact observed on the global MPI and the limited impact on the MDGs. Our simulations also show that the large impact of the intervention on the global MPI is not driven by the larger weight given by the global MPI to improvements for people who are deprived in several dimensions. We conclude that the global MPI should be used with caution in the evaluation of development programmes.

The paper is structured in the following way. The next section describes the SADA-MVP project, the evaluation design, and the datasets. Section 3 presents the impact of the intervention on the MDGs. Section 4 illustrates the impact of MVP on the global MPI. Section 5 compares and reconciles the results and draws some conclusions.

2 The SADA-Northern Ghana MVP evaluation

There are few impact evaluations of Millennium Village projects. These evaluations assess impact on a limited number of MDG indicators, often using inadequate designs without baselines or control groups. Pronyk et al. (2012) found improvements in skilled birth attendance, bednet use, malaria incidence, access to sanitation, and child mortality. Their study, however, did not include a baseline and the impact on child mortality was challenged and retracted (Bump et al., 2012; Pronyk, 2012). Remans et al. (2011) found a reduction in stunting and improvements in indicators of food security, child care, and infectious disease control.

But they did not use a control group and compared beneficiaries before and after the intervention. Clemens and Demombynes (2011) compared trends in births attended by skilled professionals, ownership of mobile phones, stunting, and primary school attendance in three Millennium Villages to the same trends in rural and district areas of the same countries and showed that differences were smaller than those observed in simple before-and-after comparisons. Wanjala and Muradian (2013) found large impacts on maize yields, profits from maize production, and consumption of own-produced food in the village of Sauri in Kenya. The study, however, did not include a baseline, and the sample was very small (nine project localities and five control localities with a total of 411 observations). Mitchell et al. (2018) examined data from 10 MVP sites and from control localities of 10 different countries, and found positive impacts on 30 out of 40 'MDGs-related' outcomes, particularly in the agriculture and health sectors. The study, however, did not include a baseline, and many of the estimated outcomes were not MDG indicators.

Our study is the only impact evaluation of an MVP that employs a baseline, a valid control group, and that assesses impact on all MDGs. The evaluation was registered with the Register for International Development Impact Evaluations (RIDIE) hosted by 3ie (Masset, 2015), and a pre-analysis plan was published before conducting the analysis of the data (Masset, 2014).

The SADA-MVP was one of 15 Millennium Villages established in 10 countries of Sub-Saharan Africa since 2006. The Millennium Villages were an experimental application of the recommendations made by the Millennium Project to achieve the MDGs and end African poverty (UN Millennium Villages Project, 2005). In the plan for ending Africa's poverty (Sachs et al., 2004), the Millennium Project presents the canonical threshold model of a poverty trap, whereby African countries are trapped in poverty by extremely low levels of human and physical infrastructure. Poverty trap models have a long history in development economics and have been used to explain poverty persistence among countries and people.[2] Poverty traps are difficult to track empirically, and there is conflicting evidence on whether they exist or not (Kraay and McKenzie, 2014). The MVP, however, did not have the immediate goal of breaking the poverty trap but to prove that the MDGs could be met as an intermediate target on the road to breaking the poverty trap (Sachs et al., 2004).

The SADA-MVP was implemented between May 2012 and December 2016, with funding of £11 million from DFID. The project was implemented in a cluster of geographically contiguous communities spread across the West Mamprusi and the Builsa districts of the Northern

[2]See for example Rosenstein-Rodan (1943), Leibenstein (1957), and Jorgenson (1961) for earlier formulations of poverty trap models and Murphy et al. (1989) and Bowles et al. (2006) for more recent ones.

and Upper East regions of Ghana. It provided a package of services in agriculture, health, education, and infrastructure to 35 communities with an approximate population of 3,900 households and 27,000 individuals.

More specifically, agricultural activities included the promotion of farmers' based organizations and cooperatives; increased access to fertilizer, seeds, and tractor services; agricultural training; access to agricultural loans; strengthening of value chains; and construction of grain storage facilities. Activities in health included the construction, rehabilitation, and staffing of health clinics; deployment of community health workers in home visits; provision of basic drugs and preventative treatments (Vitamin A, deworming, iron, vaccines, anti-malarial drugs, and mosquito bed nets); registration in the national health insurance scheme; and behavioural change campaigns. Activities in education included the construction, rehabilitation, and staffing of primary schools; teachers' training; construction of school toilets; scholarships for girls attending junior secondary school; social mobilization through parents and teachers' associations and school management committees; construction of teachers' accommodations; provision of basic school supplies; and establishment of IT learning centers. Infrastructural activities included the rehabilitation and construction of roads, boreholes and water points, and home latrines and the promotion of communication technology.

The evaluation design of the SADA-MVP consisted of a difference-in-difference (DiD) analysis of matched samples of villages and households (Masset et al., 2013). Before the project started, the 35 project villages were separately matched to 34 'near' control villages and to 34 'far' control villages within the same administrative districts using a one-to-one matching algorithm based on a propensity score calculated using village-level characteristics from the 2000 and 2010 population censuses.[3] Census data were supplemented by village data collected in the field. After collecting baseline household data, project households were matched to control households to further improve the balance of project and control observations.

Baseline surveys were conducted between April and September 2012 and follow-up rounds were conducted every year – at the same time of year – for the subsequent four years, from 2013 to 2016. The survey instruments consisted of five separate tools aimed at measuring the largest possible number of MDGs. They included a household questionnaire modelled on the Living Standard Measurement Surveys: modules on demographic characteristics of household members, migration, education, employment, weather shocks, income, and ex-

[3]Project villages were separately matched to 'near' and 'far' control villages with the goal of assessing the size of spill-over effects, which were presumed to be large at the time of the study design. In this paper we do not conduct a separate analysis of impacts by distance to project villages because the analysis showed no spill-over effects of the intervention, and we therefore employ all the control observations together.

penditure; an adult questionnaire modelled on the Demographic and Health Survey (DHS) questionnaire: modules on birth histories, contraception, antenatal and postnatal care, child health and feeding, knowledge of malaria and HIV, health-seeking behavior, and participation in community life and organizations; anthropometric measurements of children under-5; blood specimens of children under-5 through finger/heel poke to assess the prevalence of malaria; and a community questionnaire collecting information on village characteristics. Table 1 shows the number observations collected at each survey round by each survey tool. Given the intensity of the data collection exercise, we administered the adult questionnaire, the anthropometric questionnaire, and the blood questionnaire only at the baseline, midterm, and endline. As a result, impacts for some indicators, such as school attendance and expenditure and income poverty, were observed every year, while impacts on other indicators, such as nutritional status and mortality, were observed only at baseline, midterm, and endline.

Table 1: Observations by survey instrument and survey round

	2012	2013	2014	2015	2016
Households	2,172	2,230	2,191	2,177	2,185
Adult females (15–49)	2,837	-	3,241	-	2,837
Adult males (15–49)	1,628	-	1,835	-	1,671
Anthropometric measurements of children under-5	1,933	-	1,670	-	1,513
Blood samples of children under-5	805	-	1,121	-	968
Community surveys	103	103	103	103	103

Attrition was very low and even at the end of the study did not exceed 5% (see Table A.1 in the appendix). Attrition can affect the validity of the estimates by changing the composition of the population in the project and control groups in different ways. We found that attrition rates were very similar in the project and the comparison group. We also analysed the characteristics of attriter households using regression analysis and we found that very few characteristics were correlated with attrition and that they did not differ in the project and the control group. We concluded that the samples were representative of the population and that impact estimates were not biased by differential attrition.

We estimate DiD effects using regression analysis, and we calculate two types of project effects: the effect of the intervention at each survey round and the average impact of the intervention. The first effect is simply the difference in the change in the outcome between the baseline and each survey round in project and control villages. The second effect is the average of the impacts estimated with respect to the baseline for each single round. The average impact provides a better estimate of the impact of the project if impacts change over time. In the regression analysis, we use cross-sectional or fixed-effects models, depending on whether panel data are available. For some of the outcomes, for example, undernutrition, school attendance, and mortality, panel data are not available because most children under-5 at the baseline are no longer in the sample at the endline. The cross-sectional regression model with five time periods ($t = 0, \ldots, 4$) is:

$$y_{it} = \alpha + \sum_{t=1}^{4} \beta_t T_{it} + \gamma P_i + \delta P_i POST_i + \sum_{j=1}^{n} \zeta_j X_{ji0} + \epsilon_{it} \tag{1}$$

where y_{it} is the outcome observed for the observation i at time $t = 0, \ldots, 4$, where 0 is the baseline and $t = 1, \ldots, 4$ are the four subsequent survey rounds. T_t are four dummy variables for each follow-up, P is a dummy variable equal to 1 if the observation is in the project group and equal to 0 if the observation is in the control group. $POST$ is a dummy variable equal to one for every observation collected after the baseline. The coefficient δ of the interaction of the project variable with the $POST$ variable is the average effect of the intervention. The X_j are baseline values of covariates that improve the balance between the project and the comparison group.

The cross-sectional model for the estimation of round-specific project effects is:

$$y_{it} = \alpha + \sum_{t=1}^{4} \beta_t T_{it} + \gamma P_i + \sum_{t=1}^{4} \delta_t P_i T_{it} + \sum_{j=1}^{n} \zeta_j X_{ji0} + \epsilon_{it} \tag{2}$$

where coefficients and variables have the same interpretation as before, except that there are now four different project effects, one for every survey round (δ_t).

With panel data we used a fixed-effects model to account for the impact of time-invariant unobservable determinants. The average project effect and the round-specific effects were estimated using the following models respectively:

$$y_{it} = \alpha_i + \sum_{t=1}^{4} \beta_t T_{it} + \sum_{j=1}^{n} \gamma_j X_{jit} + \delta P_i POST_i + \epsilon_{it} \tag{3}$$

$$y_{it} = \alpha_i + \sum_{t=1}^{4} \beta_t T_{it} + \sum_{j=1}^{n} \gamma_j X_{jit} + \sum_{t=1}^{4} \delta_t P_i T_{it} + \epsilon_{it} \tag{4}$$

Average and round-specific project effects were estimated by the δ coefficients as in the cross-sectional models. The covariates (X_j) in this case are time-varying variables that are not affected by the project and include reported weather shocks (floods and droughts) and household demographic composition.

The regression models were estimated within subclasses of project and control observations built on the values of a propensity score using the matching approach outlined by Imbens and Rubin (2015). The subclassification approach (also known as *blocking* or *stratification*) builds classes of project and control observations with a similar propensity score to estimate project effects within classes. The within-class effects are then averaged across classes and proportionally to class size to obtain the overall impact. We calculated all standard errors of project effects using 500 bootstrapped replications and village-level clustering. We opted for a bootstrap estimation of the standard errors because an analytical derivation of the variance was not straightforward. The bootstrap algorithm replicates regressions within blocks, not the whole matching procedure.[4]

The combined use of covariance adjustment (regressions) and blocking on the propensity score (subclassification) of our algorithm has a number of advantages in comparison to other matching methods (Inbens and Rubin, 2015). The regression adjustment is employed to reduce the bias of the estimates, because blocking alone does not eliminate all the bias associated with differences in the covariates. The regression adjustment also improves the precision of the estimates and increases efficiency. Finally, since regressions are estimated within subclasses with relatively small difference between the covariates, the estimates are less sensitive to different specifications of the regression function.[5]

The starting point of our matching strategy is the estimation a propensity score using a logistic regression. The regression includes *basic* covariates that are believed to affect the outcomes, based on prior knowledge and other *additional* covariates that are potential determinants of the outcomes. We include five basic covariates (household size, age of the head of household, education of the head of household, size of cultivated land, and value of total household's wealth), and we include additional covariates stepwise from a pool of 24 poten-

[4]The number of default bootstrap replications performed by Stata is 50, which was considered good enough by Efron and Tibshirani (1993) for calculating standard errors. Cameron and Trivedi (2010) recommend 400 replications when the bootstrap is used to calculate standard errors and larger in other cases. We decided to use 500 replications after observing that the standard errors changed only marginally after increasing the number of replications beyond 500.

[5]As a robustness check, we also estimated impacts using three alternative approaches: simple regression analysis, kernel matching, and coarsened exact matching. We obtained results that were similar to those obtained using the subclassification approach.

tial covariates. We then expand the model by adding stepwise squares and interactions of the included covariates. Next, we use a linearisation of the estimated propensity score to build classes of project and control observations with similar propensity score values. We then trim the sample to remove observations with extreme propensity score values with no comparable matches in the project or control group. Finally, we used the trimmed sample to re-estimate the propensity and to build again classes of project and control observations with similar propensity score values.

We judged the validity of our matching algorithm by assessing the balance of the covariates across the project and the control groups using two tests (see Table A.2 in the appendix). The first test assesses the global balance of each covariate across strata. The second test assesses the balance for each covariate within all strata. Several covariates were out of balance before matching. After matching, the Z-values of tests of differences in the covariates across strata are very low (first test), while the second test found only two covariates out of balance.

The validity of DiD analysis rests on the similarity of the trends in the outcomes in the project and control groups. We investigated the validity of the parallel trends assumption in three ways. First, we examined the occurrence of weather shocks, which affect much of economic and health outcomes in the area, and we found that trends were very similar up to five years before the baseline. Second, we investigated whether there was a difference before the intervention in the presence of government and non-government projects, and we found no statistically significant differences between the two groups. Finally, we used retrospective data collected at the baseline on school attendance, livestock holdings, and cultivated land to test differences in the trends up to two years before the baseline, and we found no differences (see Table A.3 in the appendix).

Lastly, we conduct sensitivity analysis to assess unconfoundedness. Matching methods assume the treatment is independent of the outcomes given the covariates. Practically, this requires that pre-treatment covariates remove all biases between project and control observations. This unconfoundedness condition cannot be proved or tested. We can however disprove unconfoundedness, and, if unconfoundedness is not ruled out, it becomes more plausible.

We follow two approaches to assess the plausibility of unconfoundedness. The first approach consists of estimating the impact of the intervention on pseudo-outcomes. These are outcomes that we know could have not been possibly affected by the intervention. This is the same approach used to assess the validity of the parallel trend assumption illustrated before (see Table A.3 in the appendix). While the project had an impact on school attendance, cultivated land, and livestock holdings, it did not have any impact on the same variables before the intervention.

The second approach consists of estimating the impact of pseudo-interventions. In this approach we use the control observations to build an artificial project group and use this to simulate interventions. If matching works, we should not observe impacts of interventions that did not take place. We assess the pseudo-impact of an intervention in the Builsa 'far' communities using the West Mamprusi 'far' communities as the control group. We simulate the impact on 29 MDG indicators. This test is very conservative because the two districts are different in many ways and expected to diverge over time. The impacts found are small in size and statistically significant only in two cases: the proportion of children sleeping under bed nets and household access to sanitation (see Table A.4). This is a reassuring result considering the large differences existing between the two districts.

Finally, since we were estimating a large number of project effects at the same time, we had to address the multiple hypotheses problem. When we conduct many statistical tests at the same time, we know that some null hypotheses will be rejected by chance alone. For example, with 29 hypotheses about MDG impacts, and with a statistical significance threshold of 10%, there is a 95% probability of finding at least one statistically significant effect even if the intervention has absolutely no impact (the probability is $1-(1-\alpha)^N$, where α is the level of statistical significance and N is the number of hypotheses). Is there a way to decide how many hypotheses should be rejected? Approaches to multiple testing consist of strengthening the decision rule for declaring statistical significance. For example, in one popular approach (the Bonferroni method), the critical value to establish statistical significance is set to α/N (with 10% significance and 29 hypotheses, the critical value is 0.0034). This method is very conservative and normally results in very few rejections.

A less conservative approach – the False Discovery Rate (FDR) control algorithm of Benjamini-Hochberg (1995) – calculates the proportion of false discoveries among the rejected hypotheses. The algorithm orders the p-values of each test in ascending order and indexes them by $i = 1, \ldots, N$ and rejects all the null hypotheses whose p-value is less than $\frac{i}{N}\alpha$. In the tables of results we mark with a star the coefficients whose p-values are below the the statistical significance controlled by the FDR with $\alpha = 10\%$. Coefficients marked by a star can be interpreted as the remaining statistically significant results after removing false discoveries.

Strictly speaking, the stars do not correspond to standard statistical tests and can be defined as 'interesting' results that deserve our attention (Efron and Hastie, 2016).[6]

3 The impact of the SADA-MVP on the MDGs

Our evaluation set out, first and foremost, to assess the impact of the MVP on the MDGs. The MDGs are tracked by 60 indicators, and our study assessed impact on 29 of them. Twenty-one of the 60 indicators are measured at the national or international level and cannot be calculated using household-level data.[7] Ten indicators could not be estimated with our data because the required information was not collected or because the samples were too small to estimate averages and perform statistical tests.[8] We were therefore left with a total of 29 indicators at the household or individual level. We estimated the indicators using survey data following the UN guidelines for monitoring the MDGs.[9] In some cases our indicators differ slightly from the official UN definitions, but great care was taken in reproducing the official methodology.

The estimated average impact of the intervention on each MDG indicator and the year-specific impacts are reported in Table 2. The trajectories of the indicators over time in project and control areas are depicted in Figures 1 and 2. In these, as in the following charts, MV stands for millennium villages and CV stands for control villages.

[6]The original formulation of the Benjamini-Hochberg theorem states that the FDR control algorithm is valid only provided p-values are independent of each other. In most practical applications the p-values will be correlated. However, it can be shown that, unless the correlation is severe, the FDR control is still unbiased even in the presence of correlation, so that the FDR adjustment is correct at least in expectations (Efron and Hastie, 2016).

[7]The MDGs that cannot be calculated at the household level are: 3.3 Proportion of seats held by women in national parliament; 7.1 Proportion of land area covered by forest; 7.2 CO2 emissions, total, per capita, and per $1 GDP (PPP); 7.3 Consumption of ozone-depleting substances; 7.4 Proportion of fish stocks within safe biological limits; 7.5 Proportion of total water resources used; 7.6 Proportion of terrestrial and marine areas protected; 7.7 Proportion of species threatened with extinction; and 7.10 Proportion of urban population living in slums. Similarly, 12 other outcomes (8.1 through 8.12) relate to official development assistance, market access, and debt sustainability and can only be calculated at national or international level.

[8]The household-level indicators that could not be measured with the available data are: 5.1 Maternal mortality rate; 5.6 Unmet need for family planning; 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; 6.9 Incidence, prevalence, and death rates associated with tuberculosis; 6.10 Proportion of tuberculosis cases detected and cured under directly observed treatment short course; 8.13 Proportion of population with access to affordable essential drugs on a sustainable basis; and 8.16 Internet users per 100 inhabitants.

[9]Available at http://mdgs.un.org/unsd/mdg/host.aspx?Content=indicators/officiallist.htm.

Table 2: Impact of the MVP on the MDGs

MDG	DD Impact 2013	DD Impact 2014	DD Impact 2015	DD Impact 2016	DD Average impact
Proportion of population below a per capita income of $1.25 (PPP) per day	-11.70*	-9.84*	-3.95	-9.05*	-8.65*
	(0.001)	(0.013)	(0.294)	(0.015)	(0.002)
Proportion of population below the national (expenditure) poverty line	-0.72	-1.09	0.83	5.567	1.17
	(0.793)	(0.745)	(0.826)	(0.133)	(0.676)
Poverty gap ratio (expenditure poverty line)	-0.68	-5.89	3.24	1.90	-0.38
	(0.804)	(0.054)	(0.218)	(0.573)	(0.869)
Consumption share of poorest quintile (expenditure data)	1.17	1.49	0.99	-0.01	0.87
	(0.352)	(0.300)	(0.392)	(0.990)	(0.321)
Employment to population ratio	2.14	5.22	4.31	0.80	3.06
	(0.450)	(0.075)	(0.083)	(0.800)	(0.204)
Proportion of employed people living below a per capita income of $1.25 (PPP) per day	-13.59*	-13.63*	-6.60	-9.05*	-8.65*
	(0.001)	(0.001)	(0.106)	(0.007)	(0.000)
Proportion of own account and contributing family workers in total employment	3.89	3.90	4.28	4.04	4.02
	(0.042)	(0.046)	(0.030)	(0.049)	(0.037)
Underweight prevalence (children under-5)		1.03		-2.14	-0.51
		(0.727)		(0.435)	(0.821)
Proportion of population below the food poverty line	-0.84	-9.50	8.81	-0.42	-0.55
	(0.847)	(0.078)	(0.067)	(0.933)	(0.885)
Net attendance ratio in primary education	9.56*	4.35	3.54	13.48*	7.69*
	(0.007)	(0.252)	(0.325)	(0.000)	(0.015)
Completion rate in primary education	0.90	-1.43	-1.40	-4.12	-1.62
	(0.837)	(0.725)	(0.741)	(0.300)	(0.670)
Young adults (15-24) literacy rate		-3.36		-0.19	-3.36
		(0.113)		(0.961)	(0.313)
Ratio of girls to boys in primary education	-0.29*	-0.09	-0.10	-0.26	-0.19
	(0.011)	(0.413)	(0.420)	(0.021)	(0.058)
Share of women employed in the non-agricultural sector	-10.97	0.96	-6.92	-14.54	-8.06
	(0.531)	(0.960)	(0.664)	(0.387)	(0.545)
Under-5 mortality rate		-2.09		0.41	0.41
		(0.389)		(0.842)	(0.842)
Infant mortality rate		-0.87		2.02	2.02
		(0.711)		(0.285)	(0.285)
Measles immunisation rate (children under-2)		-6.45		-3.10	-4.95
		(0.160)		(0.545)	(0.182)
Proportion of births attended by skilled health personnel		16.57*		39.08*	27.00*
		(0.001)		(0.000)	(0.000)
Contraceptive prevalence rate		5.73*		11.48*	8.50*
		(0.018)		(0.000)	(0.000)
Adolescent birth rate				-8.67	-8.67
				(0.269)	(0.269)
Ante-natal care coverage		-7.43		2.36	-2.94
		(0.129)		(0.538)	(0.468)
Proportion of young adults (15-24) with correct HIV knowledge		0.06		2.41	1.47
		(0.832)		(0.249)	(0.474)
Malaria prevalence (children under-5)		-4.50		-4.47	-5.53
		(0.333)		(0.345)	(0.196)
Proportion of children under-5 sleeping under insecticide treated bed nets		42.88*		34.60*	39.24*
		(0.000)		(0.000)	(0.000)
Proportion of children under-5 with fever treated with antimalarial		11.13		23.70	15.99
		(0.240)		(0.023)	(0.041)
Proportion of the population using an improved drinking water source		-5.89		5.50	-0.27
		(0.174)		(0.129)	(0.940)
Proportion of the population using an improved sanitation facility		1.61		61.36*	31.04*
		(0.444)		(0.000)	(0.000)
Fixed telephone subscriptions rate		0.01		0.01	0.01
		(0.707)		(0.698)	(0.675)
Mobile telephone usage rate		-5.40		-9.96	-7.60
		(0.374)		(0.059)	(0.146)

Note: Coefficients are difference-in-difference estimates expressed in per cent terms (i.e. the coefficient are multiplied by 100). Cluster-level standard errors and P-values were calculated using 500 bootstrap replication. P-values are reported in parentheses. A star * represents a statistically significant coefficient at 10% after applying a False Discovery Rate adjustment to critical values of statistical significance. Infant and under-5 mortality rates were calculated using the DHS synthetic cohort probability method using the SYNCMRATES stata package. Standard practice is calculating mortality rates over an interval of five years before the survey. As a result, the endline and the average changes in mortality rates are the same.

All MDG indicators are binary, and the impact estimates can be interpreted as percentage points differences between the project and the control group. The project had a beneficial impact on just seven MDG indicators (less than one fourth of the total): income poverty, employment rate of people below the income poverty line, net attendance rate in primary school, the proportion of births attended by skilled professionals, the proportion of women using contraceptive methods, the proportion of children under-5 sleeping under insecticide-treated bed nets, and the proportion of the population with access to improved sanitation facilities. The impacts on the use of mosquito bed nets and on access to sanitation were large, while other impacts were small (the difference was under 10 percentage points). There were no negative impacts, with the possible exception of a reduction of the gender parity ratio in primary education.

Judged against the MDGs, the project produced, at best, mixed results. Even excluding some indicators that were not explicitly targeted by the intervention, such as adult literacy rates and the number of landline connections, the project only improved about 25% of the MDG indicators (seven out of 27). It is worth noting that the largest beneficial impacts occurred on output indicators, such as births attended by skilled professionals, use of contraceptive methods, use of bed nets, and access to improved sanitation. These impacts reflect successful project implementation, but few and small beneficial impacts were observed on final outcome indicators. Crucially, some key MDG indicators, including expenditure poverty, undernutrition, and child mortality, were totally unaffected. Interestingly, the project reduced income poverty (based on a $1.25 purchasing power parity poverty line) but did not reduce expenditure poverty (based on per-adult equivalent consumption).[10]

In relation to the activities carried out by the MVP, the project achieved some limited results. Agricultural interventions brought about a reduction of income poverty and an increase in employment, but expenditure poverty remained unchanged. Education activities produced an increase in primary school attendance, but the project failed to promote school retention as measured by completion rate in primary school. Health interventions made considerable progress in fighting malaria and in providing perinatal and postnatal care. However, progress in intermediate outcomes did not affect final outcomes: child mortality, undernourishment, and malaria incidence remained unchanged. Finally, infrastructural interventions

[10] Further analysis of the data shows that households increased asset holdings during the project, suggesting that income gains were invested rather than spent. This result is consistent with the consumer behaviour predicted by the permanent income hypothesis. Consumers interpreted income changes brought about by the intervention as temporary rather than permanent and did not therefore adjust their expenditure levels – a phenomenon also observed in the evaluation of other poverty eradication programmes (Ravallion and Chen, 2005).

gave households access to sanitation facilities, but there was no improvement in access to drinking water and mobile and Internet technology. The project did not perform better in terms of equity goals. The MVP focused on extreme poverty and on gender, but it did not affect the distribution of expenditure nor did it improved gender empowerment.

Figure 1: MDGs trends in project and control areas

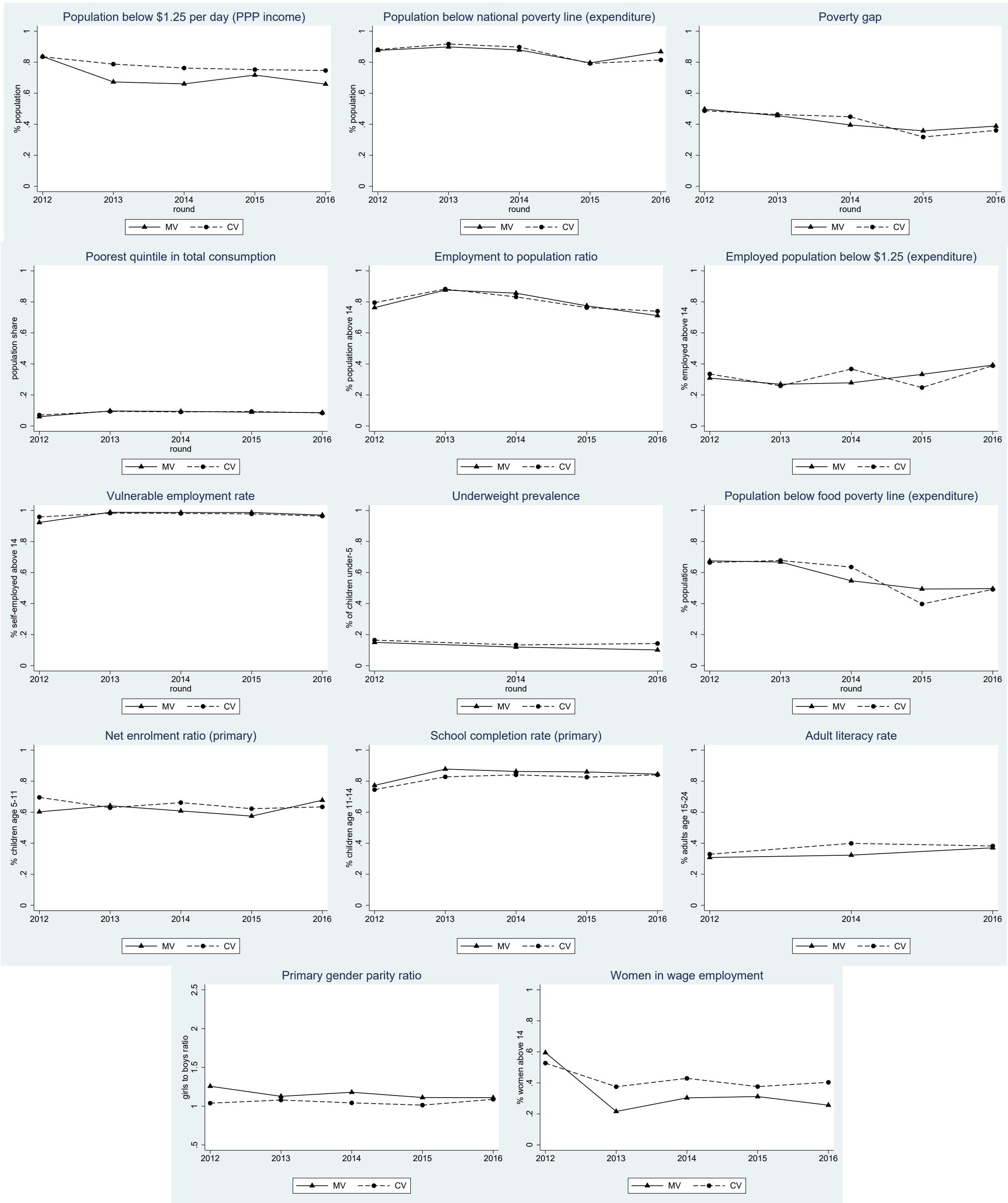
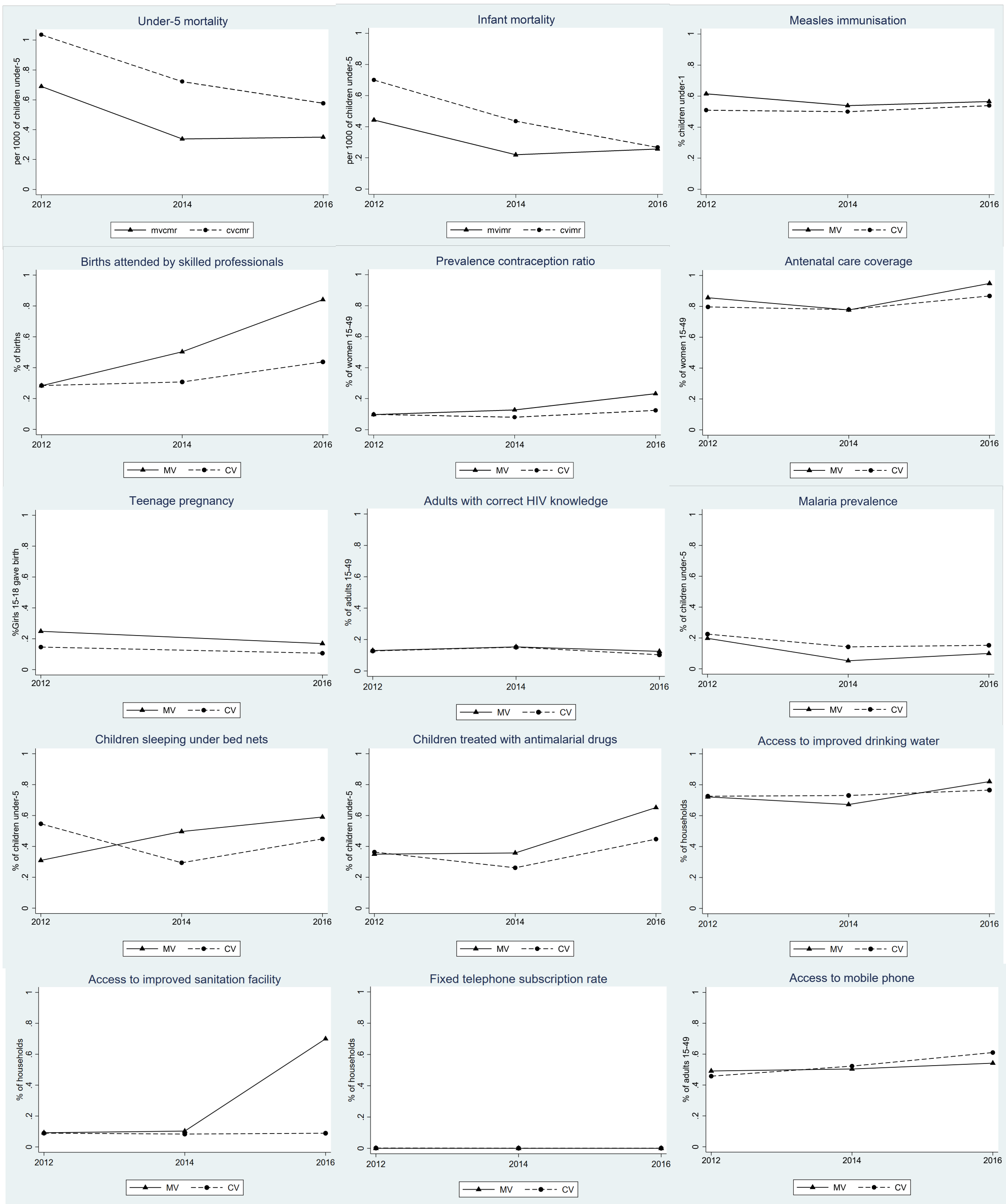


Figure 2: MDGs trends in project and control areas



4 Impact of the SADA-MVP on the global MPI

The dashboard approach struggles to make sense of heterogeneous results. The MVP had a limited impact on the MDG indicators. The project produced some positive effects in some areas but was ineffective in others, and the size of the effects varied across MDG indicators with some very large impacts and some very small ones. Counting the proportion of positive impacts is one approach to measuring the overall success of the intervention (Mitchell et al., 2018). However, counting the proportion of positive results can be misleading because it focuses attention on the statistical significance of the estimated coefficients rather than on the practical significance of their effect sizes. In addition, some indicators represent the same construct and tend to vary together, for example, the proportion of the population below a per capita income of $1.25 PPP and the proportion of the employed population below a per capita income of $1.25 PPP. Finally, simple counting of all statistically significant effects implicitly assumes that all indicators have the same relevance.

Indices solve many of these problems. They coalesce heterogeneous information into a single metric and allow straightforward comparisons between the project and the control group. Impact assessment based on an index is unambiguous and not open to different interpretations. Indices have another advantage over dashboards. They eliminate the multiple hypotheses testing problem by reducing all hypotheses tested to just one. The use of indices of families of outcomes has indeed become common in the evaluation literature following the examples of O'Brien (1984) and Kling et al. (2007).

Indices, however, have their own problems. The construction of an index is an arbitrary exercise implying a number of choices on what should be included and how it should be measured. With regards to our study, there are a number of difficulties in building an index using the MDG indicators. First, the MDG indicators are calculated over different segments of the population: households, adult women, mothers, children under 5, adolescent girls, young adults (15–24), children under 2, women of reproductive age, children under 1, and school-age children. As a result, many MDG indicators cannot be calculated for some households and cannot be included in the index. Second, some MDG indicators are population-level metrics that cannot be calculated at the household level: the poverty gap, the share of consumption of poorest quintile, and the gender parity ratio. Third, the choice of weights assigned to each indicator remains arbitrary. The construction of an index implies choices on its constituent elements and relative weights. As a result, an index of the MDG indicators could be constructed in many different ways and the advantage in the interpretation of results comes at the cost of a lack of transparency on the way the index is built.

Problems in the use of indices are not limited to the lack of transparency though. Ad-hoc indices cannot be employed by other studies using different data, thus knowledge does not accumulate. If different interventions are evaluated using different indices, there is no hope of determining which intervention is more effective or cost-effective. Indices also offer little information about how interventions operate. It is sometimes suggested that indices can be unpacked to assess the impact of an intervention in specific areas, but this seems to go against the goal of building an index in the first place. Finally, the impact of interventions on ad-hoc indices are difficult to interpret, and it is difficult to explain the practical significance of any given change in the index. How can we say whether a particular improvement in an ad-hoc index was small, medium, or large?

These problems advise against the construction of an ad-hoc index based on the MDG indicators. However, sometimes an index can be 'found' and become a useful concept beyond its originally intended use (Cartwright and Bradburn, 2011). Some indices are problematic but nevertheless useful because a) they are widely used and accepted and data are routinely collected for their calculation, b) being widely used for the evaluation of different interventions or policies, they allow the accumulation of knowledge, and c) changes in the indices can be interpreted by benchmarking in relation to observed trends and changes observed in other contexts. The Gross Domestic Product (GDP) index is a typical example. GDP is an imperfect indicator of progress, or even of economic growth (Stiglitz et al., 2009). It is not a measure of welfare. It ignores relevant economic dimensions such as, for example, the environment or unpaid work. Its measurement is based on wild imputations and is conducted with great difficulty. Despite all this, GDP is routinely and fruitfully used for the evaluation of social policies in very different contexts.

Is there a multidimensional poverty index that could serve the role played by GDP in the evaluation of public policies? We believe the global MPI of Alkire and Santos (2014) is a potential candidate for this role. Like all indices, the global MPI is not immune to criticisms. There is some arbitrariness in the selection of the dimensions, of the weights assigned to its components, and in the cut-offs used in its construction. It has been observed that the index implicitly makes undesirable trade-offs between welfare dimensions (Ravallion, 2012) and and that it may provide misleading results in the evaluation of welfare policies (Duclos and Tiberti, 2016).

The global MPI, however, shares some of the advantages of GDP. First, it is widely used. The global MPI was adopted by the UNDP in 2010 for the measurement of global poverty in the yearly Human Development Report series (UNDP, 2010). It is well accepted by the international community and data for its calculation are routinely collected by major agencies. If the global MPI were used more widely in the evaluation of public policies it would allow the cross-comparison of impacts of different projects. Second, the availability of historical series of the global MPI for all countries offers the opportunity to better understand the impact of interventions on the index via a benchmarking of the results. There are then some advantages that relate to how the global MPI is constructed. First, it is theoretically grounded in the capability approach to poverty (Sen, 1992), whereby poverty is the failure to function in a number of welfare dimensions. The global MPI does not include all MDG indicators but captures several key welfare dimensions. Second, the global MPI is constructed in such a way so that it increases when people are failing to meet basic functionings in several dimensions at the same time. The 'dual cut-off' approach to the construction of the index (Alkire and Foster, 2011) is such that the index is sensitive to the distribution of deprivations in the population. The index therefore is aligned with a system of social preferences that gives more weight to overlapping deprivations (Aaberge and Brandolini, 2015), which is closer to how most people understand poverty.

The global MPI represents deprivation in three dimensions: health, education, and living standards. These three dimensions are given equal importance (1/3), and indicators are calculated for each dimension. In particular, a household is deprived if no household member has completed five years of schooling (with a weight of 1/6); any school-age child is not attending school in years 1 to 8 (1/6); any child has died in the family (1/6); any child for which there is information is malnourished (1/6); the household has no electricity (with a weight of 1/18); the household does not have access to an improved sanitation facility (1/18); the household does not have access to improved drinking water (1/18); the household has dirt, sand, or dung flooring (1/18); the household cooks with dung, wood, or carbon (1/18); or the household does not own more than one radio, TV, telephone, bike, motorbike, or refrigerator, and does not own a car or truck (1/18). The weighted sum of the deprivation indicators produces a deprivation score with values between zero and one for each household.

Our survey questionnaires were modelled on the DHS questionnaires, which are also used in calculating the global MPI. We were therefore able to calculate the index in the same way with just two exceptions. First, our malnourishment deprivation index is based on child undernutrition only, because our surveys did not measure the nutritional status of mothers, through the body mass index. Second, we restricted the time for calculating child mortality to five years before the survey in order to be able to measure more accurately changes produced by the intervention.[11] Since not all households have children under five or children of school age, some deprivation indicators are censored. We deal with censored observations following the same procedure adopted for the global MPI (Alkire and Santos, 2014). We consider as not deprived those households for which no information is available to assess their deprivation status.

Following Alkire and Santos (2010), we use the the household deprivation score described above to calculate the following indices:

- The multidimensional poverty headcount ratio, also called the **incidence** of poverty (H). This is the fraction of the population with a deprivation score equal to or larger than one third. Note that deprivation is measured at the household level and that population-level deprivation is obtained by weighting household observations by household size.
- The average deprivation score among the poor, or the **intensity** of poverty (A).
- The average deprivation score of the poor across the whole population, also called the adjusted multidimensional poverty index (MPI). This is the **global MPI**, which is the average deprivation score after setting to zero the deprivation scores of households that are not multidimensionally poor according to the index H above. It can be shown that the global MPI is the product of the other two indices: the incidence and intensity of poverty (MPI=H*A).

As mentioned, the global MPI measures deprivation by employing a dual cut-off method. First, each household is classified as deprived or not deprived in each indicator. Second, each household is classified as multidimensionally deprived if it is deprived in at least one third of all dimensions. A consequence of the use of a dual-cutoff approach is that the global MPI considers the joint distribution of deprivations. This is a great advantage of the global MPI versus other poverty indices. Deprivation indices as the MDGs indicators, whether in isolation or aggregated in a single index, do not capture whether a household is deprived in multiple dimensions. The global MPI assumes that we prefer a society in which the same

[11]The latest version of the global MPI of 2018 has also adopted the same five-year convention for the calculation of mortality deprivation (Oxford Poverty and Human Development Initiative , 2018).

deprivations are equally distributed, to one in which they are concentrated in few households (Aaberge and Brandolini, 2015).

The global MPI has three other properties (Alkire and Santos, 2014). First, it satisfies dimensional monotonicity. It increases if a person becomes more deprived in an additional dimension. Second, the index can be decomposed by population subgroups, thus enabling poverty comparisons between subgroups of the same population. Third, the index can be broken down by deprivation indicator. The contribution of dimension-specific deprivations to overall global MPI deprivation can be calculated, thus providing information on the main sources of deprivation.

The MVP had a positive impact on multidimensional poverty. The charts in Figure 3 show the impact of the MVP on the global MPI, the incidence of multidimensional poverty, and the intensity of multidimensional poverty. The values of the three indices were nearly identical at the baseline in the project and control groups, and a t-test showed that the differences were not statistically significant. After the intervention, multidimensional poverty decreased in the control group, but it decreased at a faster rate in the project group.

Figure 3: Impact of MVP on the MPI

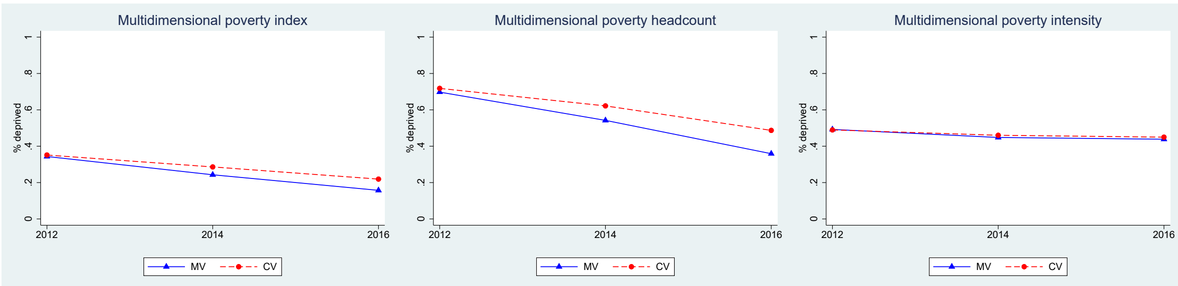


Table 3 presents the DiD impact estimates. The MVP produced a statistically significant reduction in the global MPI and in poverty incidence. The impact was larger at the endline than at the midterm, pointing to a continuous impact of the intervention over time. Finally, the project decreased multidimensional poverty intensity only marginally, and the effect was not statistically significant. The different impact of the MVP on incidence and intensity of poverty is interesting from a policy perspective. Recall that the global MPI is the product of poverty incidence and of poverty intensity. If the global MPI is reduced via a large reduction in poverty incidence and a small reduction in poverty intensity, it means that it was mainly reduced by improving the conditions of those poor people who had lower intensities of poverty. Conversely, changes relatively more favourable to the poorest of the poor would reduce the global MPI by reducing relatively more poverty intensity than poverty incidence. In interpreting changes in poverty intensity, it should be noted, however, that a small reduction in intensity does not necessarily mean that the conditions of the poorest of the poor did not improve. It may also be the case that their conditions improved, driving poverty intensity down, at the same time as the conditions of the less poor also improved, driving poverty intensity up, in such a way that the two effects cancelled each other out.

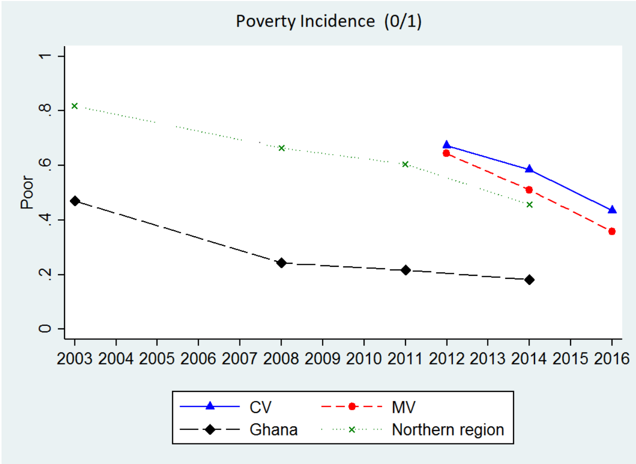
Table 3: Impact of the MVP on multidimensional poverty

	Baseline in CV areas	Baseline difference in MV	DD impact 2014	DD impact 2016	Average DD impact
Multidimensional poverty index	35.16	-0.81	-3.64*	-5.48**	-4.56**
		(0.793)	(0.070)	(0.017)	(0.019)
Multidimensional poverty incidence	71.82	-2.04	-6.42*	-11.37**	-8.90**
		(0.675)	(0.083)	(0.015)	(0.015)
Multidimensional poverty intensity	48.96	0.24	-1.74	-0.96	-1.50
		(0.857)	(0.116)	(0.470)	(0.115)

Note: Coefficients are difference-in-difference estimates using a cross-sectional model estimated using sub classification on a trimmed sample. Standard errors calculated using 500 bootstrap replications. P values in parentheses based on cluster standard errors. Stars represent statistical significance levels, whereby * is 10%, ** is 5%, and *** is 1%.

Was the impact of the MVP on the global MPI large? Interpreting results measured with indices is difficult. Without additional information we are unable to interpret the practical significance of a change in the value of an index. The relevance of an impact however can be assessed by benchmarking the results. We compared poverty incidence trends in project and control areas to the same trends in Ghana and in the Northern Region (see Figure 4). To do this we calculated the global MPI using the four DHS datasets available for Ghana over the last 15 years. Since no DHS data were available to calculate the global MPI at the project endline, we extrapolated linear trends (for the Northern Region) and log-linear trends (for Ghana). At baseline, poverty incidence in the study area was four times the incidence in Ghana and larger than in the Northern Region. After the intervention, progress occurred in both project and control areas, but at different rates. After four years, the MVP reduced the gap with the rest of Ghana by 50%, while the gap reduction in the control area was only 30%. After four years, the MVP area caught up with the Northern Region, while the control area was still poorer than the Northern Region. At current trends, the control areas will catch up with the rest of the country in 2024 while the project areas would catch up in 2022, implying the MVP produced a two year acceleration in current trends.

Figure 4: Multidimensional poverty in Ghana and in the study area



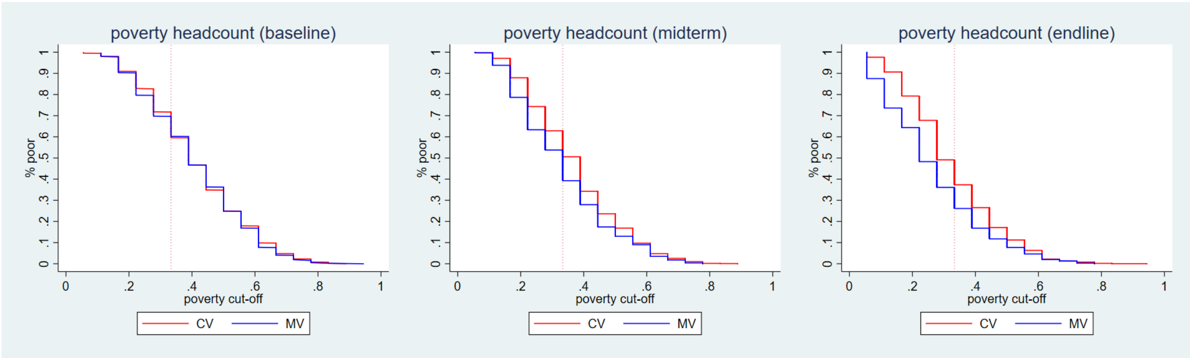
5 Sensitivity analysis

In this section we assess the robustness of the results obtained in the previous section. In particular, we assess the sensitivity of the global MPI to the choice of the poverty cut-off, we analyse to what extent the changes observed in the global MPI are driven by changes in one or two of its component elements, and we measure to what extent they reflect changes in the distribution of deprivation in the population.

The impact of the MVP on multidimensional poverty was independent of the poverty cut-off. We assessed the sensitivity of the results to the cut-off with a simple stochastic dominance analysis. The global MPI employs a 1/3 cut-off, meaning that a household (and all its members) are poor if their deprivation score is equal to or larger than a third. Using a 1/3 cut-off we found that poverty was lower in project areas after the intervention. It would be problematic if we were to find that poverty was lower in control areas using a different cut-off of, for example, 1/2 or 1/5, because this would imply that the difference depends on the cut-off used.

The charts in Figure 5 plot multidimensional poverty incidence for all possible poverty lines for the three survey rounds separately. When the multidimensional poverty line is 0, everybody is poor. Poverty decreases as we increase the poverty cut-off, and when the cut-off is 1 (a household has to be deprived in all dimensions to be classified as poor) very few households are poor. At baseline, the poverty distributions in the project and control areas overlap and poverty incidence is nearly identical for all possible poverty lines. At the midterm and endline poverty is unequivocally lower in project areas. At no poverty cut-off the lines are crossing. It cannot be argued that the impact is particularly large at the 1/3 cut-off. Using different cut-offs may or may not have resulted in smaller impact estimates.

Figure 5: Stochastic dominance analysis



The global MPI is a composite index of 10 deprivations, and it is useful to see in what dimensions most improvement occurred. The index can be broken down by deprivation, and impacts can be analysed separately. The charts in Figure 6 show the percentage of deprived population for each deprivation category in project and control villages over time. The figure indicates a clear impact on sanitation, but no other large effects are visible.

We tested the impact of the MVP on each deprivation using DiD analysis (see the results in Table 4). Deprivations were fairly similar at the baseline in the project and control groups with only one indicator (child mortality) showing a statistically significant difference. The project reduced all deprivation indices, with the exception of child mortality and the use of cooking fuel. The impacts, however, were small (less than 10 percentage points), with the exception of a large impact on sanitation, and they were statistically significant only in the case of school attendance and sanitation.

Figure 6: MPI deprivation trends in project and control areas

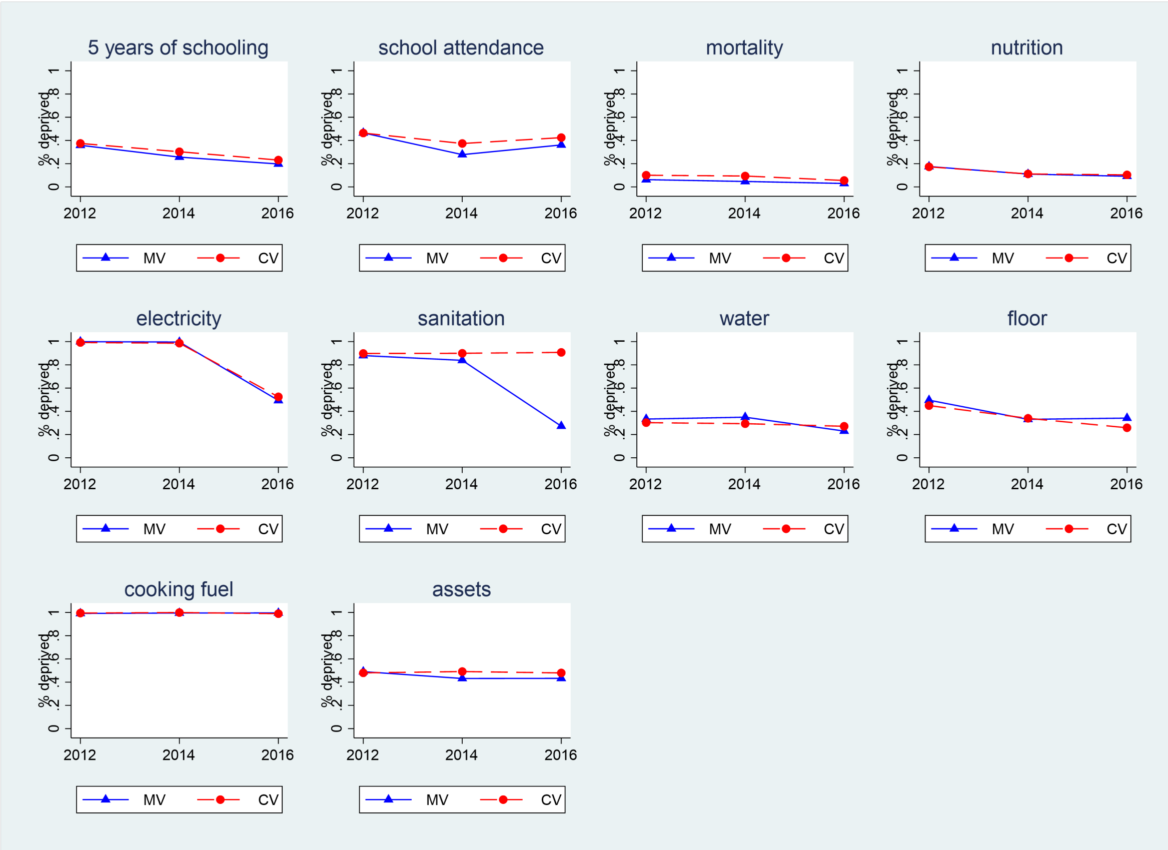


Table 4: Impact of the MVP on deprivation indices

	Baseline in CV areas	Baseline difference in MV	Contribution to index	DD impact 2014	DD impact 2016	Average DD impact
Years of schooling	37.48	-1.60	17.03	-2.82	-1.78	-2.30
		(0.778)	(0.390)	(0.573)	(0.445)	
School attendance	46.39	0.11	22.22	-9.13**	-6.49	-7.83**
		(0.980)	(0.011)	(0.113)	(0.021)	
Child mortality	10.07	-3.80**	2.99	-1.11	1.51	0.15
		(0.021)	(0.578)	(0.384)	(0.932)	
Nutrition	17.27	0.31	7.99	0.53	-1.05	-0.26
		(0.915)	(0.857)	(0.693)	(0.914)	
Electricity	99.23	0.75	11.28	0.36	-5.43	-2.52
		(0.151)	(0.533)	(0.620)	(0.640)	
Sanitation	89.78	-1.85	10.73	-3.61	-62.26***	-32.74***
		(0.631)	(0.233)	(0.000)	(0.000)	
Water	30.25	3.06	4.16	2.91	-7.32	-2.20
		(0.603)	(0.595)	(0.138)	(0.646)	
Floor	44.95	4.46	6.47	-6.34	1.08	-2.65
		(0.563)	(0.119)	(0.857)	(0.559)	
Cooking fuel	99.51	0.50	11.17	-0.03	0.97	0.47
		(0.531)	(0.965)	(0.153)	(0.421)	
Assets	48.03	0.95	5.9	-7.49	-6.37	-6.97
		(0.825)	(0.182)	(0.222)	(0.160)	
TOTAL	100	100	100	100	100	100

Note: The first column is the deprivation at each dimension in the control group. The second column is the per cent difference at baseline between the project and the control group. Columns 3 to 5 are per cent DD effects of the project on each deprivation index. Standard errors calculated using 500 bootstrap replications. P values in parentheses based on cluster standard errors. Stars represent statistical significance levels, whereby * is 10%, ** is 5%, and *** is 1%. The last column figures are the per cent contributions of each deprivation index to the global MPI.

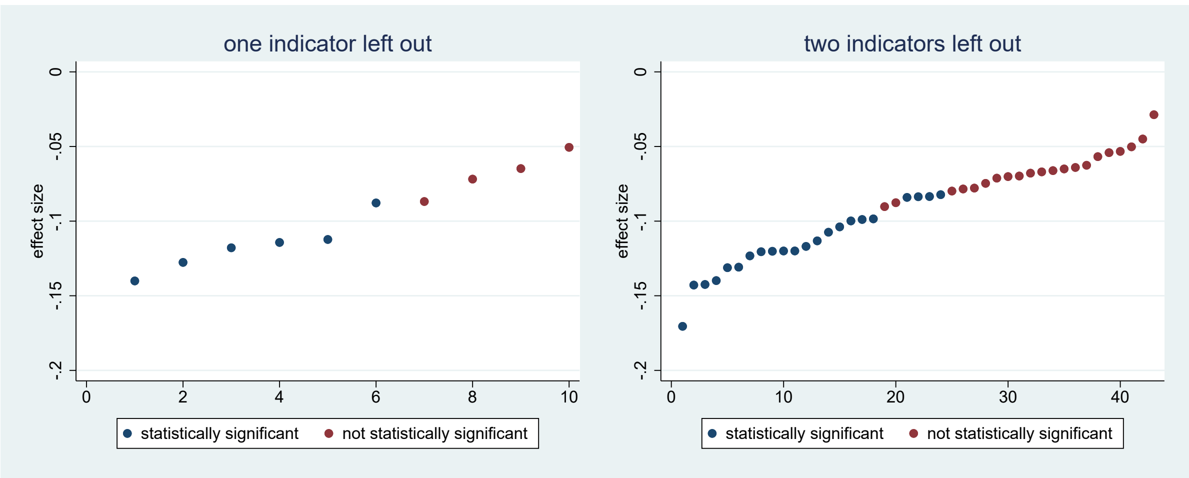
The third column of Table 4 shows the contribution of each deprivation to the global MPI at the baseline. Contributions are obtained multiplying each censored deprivation (setting to zero the deprivation index for households that are not multidimensionally poor) by the weight assigned to each deprivation and dividing by the global MPI. The percentage contributions tell us what are the main drivers of overall deprivation. At the baseline, nearly 40% of total deprivation was caused by failures in education. Another 20% was driven by failures in sanitation and electricity. One interpretation of the large impact of the MVP on multidimensional poverty is that it happened to have a large impact on two deprivations (school attendance and sanitation), which together accounted for more than 30% of total deprivation.

We further investigate the sensitivity of the results by estimating impacts without each indicator in turn. We estimated poverty after leaving one of the index components out at a time.

This is equivalent to setting the weight of the left-out indicator to zero. We then redistributed the weight of the left-out indicator in a 'nested' way, that is, we reassigned the weight of the missing indicator to the other indicators within its welfare dimension rather than across all dimensions (for example, after excluding nutrition, the weight for child mortality increases from 1/6 to 1/3, while all other weights remain unchanged). With 10 dimensions, this exercise produces 10 new poverty estimates. We also conducted the same exercise leaving out two indicators at a time using the same nested procedure, but avoiding combinations that would result in a removal of an entire welfare dimension (for example, we did not leave out years of schooling and school attendance at the same time). For the same reason, we did not extend this exercise to more than two components, because this would lead to the removal of entire welfare dimensions, which is against the rationale for building a multidimensional index.

The results of these simulations are shown in the charts of Figure 7. Effect sizes at each run are reported in increasing order. Blue dots are statistically significant results, while red dots are not statistically significant. Given the large number of tests, statistical significance was assessed against critical values adjusted by the False Discovery Rate. All estimations show a positive impact of the MVP on multidimensional poverty. Note, however, the large variety of results, ranging from 15% to 5% when leaving one indicator out, and ranging from 17% to 3% when leaving two indicators out. In addition, impact is statistically significant only in 60% of cases when leaving one or two indicators out and only in 50% of cases when leaving out two indicators. These simulations suggest that both the effect size and the statistical significance of the intervention are highly sensitive to the removal of one or two indicators.

Figure 7: Sensitivity of the global MPI to 1 or 2 indicators



The use of a dual cut-off in the construction of the global MPI allows the identification of households suffering joint deprivations. This sensitivity of the index to joint deprivations is one of its most attractive characteristics because it corresponds to the way people normally think about poverty. The global MPI can improve even if the population-level deprivations remain unchanged. For example, if the project changed the distribution of deprivations in favour of the most deprived, the global MPI would improve even with small changes in deprivations. Could the large impact of the MVP on the global MPI be explained by a large change in the distribution of deprivations in favour of the most deprived?

To test this hypothesis we conducted a simulation exercise whereby we randomly shuffled the switches in deprivation at the midterm and at the endline, separately for the project and control observations. For example, if 20 households that were deprived in one indicator at the baseline in the project group became not deprived at the midterm, we randomly shuffled these changes within the project group at the midterm. Similarly, we shuffled the changes occurring in the other direction (from not deprived to deprived) – in both the project group and control group. By randomly reallocating the deprivation switches in this way, we preserved the population effect of the intervention and we broke the correlations between switches. The simulations deliver the impact of the intervention produced by uncorrelated switches to which the actual impact of the intervention can be compared. The observed impact will be larger than the simulated impact if it reduces the correlation between deprivations among the poor.

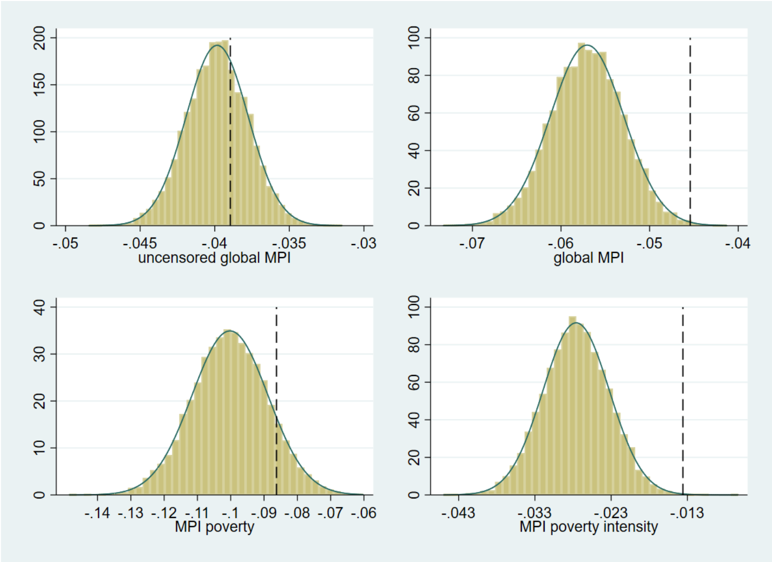
Table 5: Simulated impacts of uncorrelated changes in deprivations

	Observed impact	Simulated impact	95% interval
Mean deprivation score	-4.28	-3.98	[-4.39 , -3.58]
Global MPI	-4.56	-5.71	[-6.52 , -4.89]
MPI incidence	-8.90	-10.0	[-12.26 , -7.78]
MPI intensity	-1.50	-2.76	[-3.61 , -1.91]

Note: The first column shows the estimated coefficients of the impact of the MVP as reported in Table 3. The second column reports the average estimated impacts after 10,000 reshufflings of changes in deprivation. The last column is the 95% interval of the simulated normal distribution.

The results of the simulations are shown in Figure 8. The estimated impact of the MVP on the global MPI and other indices are compared to simulated impacts in Table 5. As expected, the actual and simulated impacts on the mean deprivation score are very similar and the actual impact is within the 95% normal distribution of the simulated impact. All actual

Figure 8: Simulation of impact of the MVP with uncorrelated changes in deprivation



Note: Histograms of 10,000 simulated impacts of MVP after reshuffling changes in deprivations. The dashed vertical lines are impacts estimated with the observed data.

impacts are lower than the simulated ones and outside the 95% normal interval, with the exception of the impact on poverty incidence. The differences are not large but point to the conclusion that the reductions in deprivations produced by the MVP did not reduce the correlations in deprivations among the poor. On the contrary, the MVP improved relatively more the conditions of those relatively less deprived, a fact that is in agreement with the lack of observed change in the intensity of multidimensional poverty. Hence, the large observed impact of the MVP on the global MPI was not a result of a relatively larger improvement in the conditions of households suffering multiple deprivations, because, on the contrary, the project improved relatively more the conditions of households not suffering from multiple deprivations. The simulations show that the change in the global MPI would have been even larger had the project not favoured the less poor relatively more.

6 Conclusions

The use of a dashboard of MDG indicators and of the global MPI in the evaluation of the MVP produced different results. The project had a limited impact on the MDGs, but the impact on the global MPI was substantial. Why were the results so different?

First, the MDG indicators and the global MPI are not measuring the same welfare construct. The global MPI is related to the MDGs but it is not an aggregation of MDG indicators, and only about half of the items used to build the index are also used to monitor the MDGs. Second, the global MPI is sensitive to small changes in deprivations. Because of the dual cut-off, a household can switch from a poor to a non-poor status simply as a result of a change in one or two deprivations. We showed that the MVP particularly improved school attendance and access to improved sanitation, which in turn drove much of the improvement in the global MPI. The global MPI is also sensitive to changes in deprivation that benefit relatively more the multidimensionally poor. We investigated whether the large impact observed was driven by a relatively larger improvement in the living conditions of households that were multidimensionally poor. We found that, on the contrary, the project improved relatively more the conditions of the less poor, and our simulations show that the global MPI would have improved even more if it were constructed in such a way to give equal weight to improvements among the poor and the non-poor.

We conclude with some reflections on the effectiveness of the MVP and on the reliability of the global MPI in the evaluation of welfare policies. Was the MVP successful? And should we use the global MPI in the evaluation of development programmes?

The MVP produced modest results, improving just a quarter of the MDG indicators and failing to improve key welfare outcomes. In addition, the largest beneficial impacts were observed on project outputs rather than final outcomes. It is true that the MVP reduced multidimensional poverty as measured by the global MPI, and this reduction was substantial in comparison to current trends. However, much of this impact was driven by improvements in just two deprivation indicators. It is also worth mentioning that changes in school attendance and access to sanitation are changes in outputs rather than outcomes. They suggest that the project was successfully implemented but do not necessarily represent welfare improvements. An increase in school enrolment does not ensure that children's literacy and numeracy skills are improving. Similarly, an increased access to sanitation facilities does not imply that the same facilities are used or that there has been a reduction in morbidity.

Index approaches have a number of advantages over dashboards in the evaluation of programmes with multiple outcomes. First, they summarise heterogeneous information in a single metric and allow straight comparisons between groups and subgroups. Second, by summarising impacts on multiple outcomes, they remove the problem of testing multiple hypotheses. Finally, they prevent the selective reporting of results.

Despite these advantages, our conclusion is that the global MPI should be employed with caution in the evaluation of development policies. First, some of these advantages are only apparent. The problem of testing multiple hypotheses can be addressed by using appropriate statistical methods such as, for example, the False Discovery Rate correction. As for the risk of selective reporting, this can be prevented through the use of pre-analysis plans and higher transparency in conducting research.

More importantly, the global MPI is sensitive to changes even when these changes occur in just a few deprivations. It is true that the global MPI can be disaggregated into its components and that the impacts on specific deprivations can be analysed in a transparent way. However, if the global MPI has to be decomposed into its components to be fully understood, then we would prefer to consider a wider dashboard of indicators that are currently not included in the global MPI such as, for example, expenditure poverty, employment, and gender equality.

We are not suggesting, however, that the global MPI should incorporate all MDG indicators, nor that impact evaluations should build ad-hoc indices based on the expected outcomes of specific programmes. In fact, our interest in testing the use of the global MPI in evaluation was driven by a desire to prevent the proliferation of indices that are not comparable to each other and whose observed changes have limited practical meaning. We tested the use of the global MPI in evaluation as a 'found' object for the assessment of development interventions with multiple outcomes. The global MPI has the advantage of being simple and transparent, as well as capturing fundamental dimensions of well-being. The data needed for its construction are minimal and easy to collect. Our analysis found that the global MPI is very sensitive to changes in some deprivations and that it does not include important dimensions affected by development programmes. However, we do not suggest that the global MPI should not be used in evaluation or policy analysis, rather that its use should be further tested and that this might help its continuous refinement.

References

Aaberge, R. and Brandolini, A. (2015). *Multidimensional Poverty and Inequality*, vol. 2, Amsterdam: Elsevier, pp. 141–216.

Alkire, S. and Foster, J. (2011). 'Counting and multidimensional poverty measurement', *Journal of Public Economics*, vol. 95(7-8), pp. 476–487.

Alkire, S. and Santos, M.E. (2010). 'Acute multidimensional poverty: A new index for developing countries', *OPHI Working Paper No. 38, University of Oxford*.

Alkire, S. and Santos, M.E. (2014). 'Measuring acute poverty in the developing world: Robustness and scope of the Multidimensional Poverty Index', *World Development*, vol. 59, pp. 251–274.

Benjamini, Y. and Hochberg, Y. (1995). 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society Series B*, vol. 57(1), pp. 289–300.

Bowles, S., Durlauf, S. and Hoff, K. (2006). *Poverty Traps*, Princeton: Princeton University Press.

Bump, J.B., Clemens, M., Demombynes, G. and Haddad, L. (2012). 'Concerns about the Millennium Villages Project report', *The Lancet*, vol. 379(May 26), p. 1945.

Cameron, A. and Trivedi, P. (2010). *Microeconometrics Using Stata*, College Station: Stata Press.

Cartwright, N. and Bradburn, N. (2011). *A Theory of Measurement*, Washington: National Academies Press.

Clemens, M. and Demombynes, G. (2011). 'When does rigorous impact evaluation make a difference? The case of the millennium villages', *Journal of Development Effectiveness*, vol. 3(3), pp. 305–339.

Duclos, J.Y. and Tiberti, L. (2016). *Mulitidimensional Poverty Indices: A Critical Assessment*, New York: Oxford University Press.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*, New York: Cambridge University Press.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, New York.

Inbens, G. and Rubin, D.B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*, New York: Cambridge University Press.

Jorgenson, D. (1961). 'Development of the dual economy', *Economic Journal*, vol. 71, pp. 309–334.

Kling, J.R., Liebman, J. and Katz, L. (2007). 'Experimental analysis of neighborhood effects', *Econometrica*, vol. 75(1), pp. 83–119.

Kraay, A. and McKenzie, D. (2014). 'Do poverty traps exist? Assessing the evidence', *Journal of Economic Perspectives*, vol. 28(3), pp. 127–148.

Leibenstein, H. (1957). *Economic Backwardness and Economic Growth*, New York: Wiley.

Loschmann, C., Parsons, C. and Siegel, M. (2015). 'Does shelter assistance reduce poverty in Afghanistan?', *World Development*, vol. 74, pp. 305–322.

Masset, E. (2014). *Northern Ghana Millennium Village Impact Evaluation: Analysis Plan*, ITAD.

Masset, E. (2015). *Impact Evaluation of the SADA Northern Ghana Millennium Village Project*, RIDIE.

Masset, E., Acharya, A., Barnett, C. and Dogbe, T. (2013). 'An impact evaluation design for the Millennium Villages Project in Northern Ghana', *Journal of Development Effectiveness*, vol. 5(2).

Mitchell, A. and Macció, J. (2018). 'Evaluating the effectiveness of housing interventions on multidimensional poverty: The case of TECHO-Aargentina', *OPHI Working Paper No. 120, University of Oxford*.

Mitchell, S., Gelman, A., Ross, R., Chen, J., Bari, S., Huynh, U.K., Harris, M.W., Sachs, S.E., Stuart, E.A., Feller, A., Makela, S., Zaslavsky, A.M., McClellan, L., Ohemeng-Dapaah, A., Namakula, P., Palm, C.A. and Sachs, J.D. (2018). 'The Millennium Villages Project: A retrospective, observational, endline evaluation', *The Lancet*, vol. 6, pp. e500–e513.

Murphy, K., Shleifer, A. and Vishny, R. (1989). 'Industrialisation and the big push', *Journal of Political Economy*, vol. 97(5), pp. 1003–1026.

O'Brien, P. (1984). 'Procedures for comparing samples with multiple endpoints', *Biometrics*, vol. 40(4), pp. 1079–1087.

Oxford Poverty and Human Development Initiative (2018). *Global Multidimensional Poverty Index 2018: The Most Detailed Picture to Date of the World's Poorest People,* University of Oxford, UK.

Pattanaik, P. and Xu, Y. (2018). 'On measuring multidimensional deprivation', *Journal of Economic Literature*, vol. 56(2).

Pronyk, P. (2012). 'Errors in a paper on the Millennium Villages Project', *The Lancet*, vol. 379(May 26), p. 1946.

Pronyk, P.M., Muniz, M., Nemser, B., Somers, M.A., McClellan, L., Palm, C.A., Huynh, U.K., Ben Amor, Y., Begashaw, B., McArthur, J.W., Niang, A., Sachs, S.E., Singh, P., Teklehaimanot, A. and Sachs, J.D. (2012). 'The effect of an integrated multisector model for achieving the Millennium Development Goals and improving child survival in rural Sub-Saharan Africa: A non-randomised controlled assessment', *Lancet*, vol. 379(9832), pp. 2179–88.

Ravallion, M. (2012). 'Mashup indices of development', *World Bank Research Observer*, vol. 27(1), pp. 1–32.

Ravallion, M. and Chen, S.H. (2005). 'Hidden impact? Household saving in response to a poor-area development project', *Journal of Public Economics*, vol. 89(11-12), pp. 2183–2204.

Remans, R., Pronyk, P.M., Fanzo, J.C., Chen, J., Palm, C.A., Nemser, B., Muniz, M., Radunsky, A., Abay, A.H., Coulibaly, M., Mensah-Homiah, J., Wagah, M., An, X., Mwaura, C., Quintana, E., Somers, M.A., Sanchez, P.A., Sachs, S.E., McArthur, J.W. and Sachs, J.D. (2011). 'Multisector intervention to accelerate reductions in child stunting: An observational study from 9 Sub-Saharan African countries', *American Journal Clinical Nutrition*, vol. 94(6), pp. 1632–42.

Rosenstein-Rodan, P. (1943). 'Problems of industrialization of Eastern and Southeastern Europe', *Economic Journal*, vol. 53, pp. 202–211.

Sachs, J., Mcarthur, J., Shmidt-Traub, G., Kruk, M., Bahadur, C., Faye, M. and McCord, G. (2004). 'Ending Africa's poverty trap', *Brookings Papers on Economic Activity*, vol. 1(2004), pp. 117–240.

Sen, A. (1992). *Inequality Reexamined*, Cambridge: Harvard University Press.

Song, S. and Imai, K. (2018). 'Does the hunger safety net program reduce multidimensional poverty? Evidence from Kenya', *OPHI Working Paper No. 124*, University of Oxford.

Stiglitz, J., Sen, A. and Fitoussi, J.P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*, https://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report.

UN Millennium Villages Project (2005). 'Investing in development: A practical plan to achieve the Millennium Development Goals'.

UNDP (2010). *Human Development Report*, New York: Oxford University Press.

Wanjala, B. and Muradian, R. (2013). 'Can big push interventions take small-scale farmers out of poverty? Insights from the Sauri millennium village in Kenya', *World Development*, vol. 45, pp. 147–160.

Appendix

This appendix includes tables discussed in the text and mostly related to the matching approach used in the study. Table A.1 presents the attrition rates observed in the samples of project and control observations. Table A.2 shows the results of statistical tests of the difference in covariates in the project and the control group before and after matching. Table A.3 shows the result of our first assessment of the plausibility of the unconfoundedness assumption. The results refer to tests of the impact of the intervention on values of the outcomes before the baseline. The nonrejection of the null hypotheses can also be interpreted as stating the plausibility of the parallel trends assumption made in DiD analysis. Table A.4 is our second assessment of the plausibility of the unconfoundedness assumption. It shows the results of assessing the impact of an intervention that did not take place by artificially splitting the control observations into a project and a control group.

Table A.1: Household attrition in project and control areas

Sample	2012	2013	2014	2015	2016
Households panel (project)	711	707	697	689	684
attrition		0.6%	2%	3.1%	3.8%
Household panel (control)	1461	1,454	1,424	1,391	1,389
attrition		0.5%	2.5%	4.8%	4.9%
All households	2172	2,161	2,121	2,080	2,073
attrition		0.5%	2.3%	4.2%	4.6%

Table A.2: Test of covariance balance

Covariate	Unadjusted T-test	Z-value test across strata	F-value test within strata
Household size	-0.83	-0.13	0.71
Age of head	-1.50	0.25	1.42
Education of head	-0.98	-0.24	1.14
Cultivated land	-2.63	-0.31	0.47
Wealth	-1.78	0.01	0.69
Remittances	-4.19	-0.03	2.19
Millet farm	-2.82	-0.10	0.75
Rice farm	3.92	0.09	0.43
Drought shocks	3.02	-0.07	3.75
Flood shock	-3.25	-0.03	1.40
Isolated household	-3.56	-0.03	0.44
Months food insecure	3.62	0.12	0.44
Farmer household	2.52	0.27	1.07
Bank access	-3.13	-0.06	1.02
Metal roof	0.50	-0.09	0.52
Distance to drinking water	-1.49	0.00	0.59
Groundnut farm	-1.95	0.46	1.26

Note: The first column includes t-statistics of tests of the differences in the covariates at baseline. The second and the third columns show the values of test statistics after applying our matching algorithm.

Table A.3: Impact of the MVP on pseudo-outcomes

	Pseudo-effect t-1	Pseudo-effect t-2	Project effect 4-year average
Net attendance rate (primary)	0.012		0.077
	(0.015)		(0.032)
Cultivated land (ln of acres)	0.002	0.029	0.092
	(0.026)	(0.034)	(0.062)
Livestock holdings (ln of value in Cedis)	0.128	0.032	0.333
	(0.193)	(0.134)	(0.123)

Note: Cluster-level standard errors in parentheses obtained by running 500 bootstrap replications.

Table A.4: Impact of pseudo-intervention on the MDGs

MDG	DiD (Control Far Builsa vs Control Far West Mamprusi)
Proportion of population below $1.25 (PPP) per day	0.006 (0.919)
Proportion of population below the national poverty line	0.019 (0.621)
Poverty gap ratio	-0.024 (0.559)
Consumption share of poorest quintile	0.014 (0.152)
Employment to population ratio	-0.047 (0.222)
Proportion of employed people living below $1.25 (PPP) per day	0.004 (0.956)
Proportion of own account and contributing family workers in total employment	-0.016 (0.100)
Underweight prevalence (children under-5)	0.088 (0.086)
Proportion of population below the food poverty line	0.010 (0.878)
Net attendance ratio in primary education	0.063 (0.197)
Completion rate in primary education	-0.035 (0.390)
Young adults (15–24) literacy rate	0.024 (0.704)
Ratio of girls to boys in primary education	-0.034 (0.782)
Share of women employed in the non-agricultural sector	0.105 (0.649)
Under-5 mortality rate	0.053 (0.201)
Infant mortality rate	0.037 (0.274)
Measles immunisation rate (children under-2)	-0.172 (0.016)
Proportion of births attended by skilled health personnel	0.086 (0.241)
Contraceptive prevalence rate	-0.025 (0.502)
Adolescent birth rate	-0.055 (0.545)
Antenatal care coverage	0.028 (0.600)
Proportion of young adults (15–24) with correct HIV knowledge	-0.007 (0.853)
Malaria prevalence (children under-5)	-0.148 (0.142)
Proportion of children under-5 sleeping under insecticide treated bed nets	**0.358** (0.000)
Proportion of children under-5 with fever treated with antimalarial drugs	0.066 (0.655)
Proportion of the population using an improved drinking water source	0.184 (0.011)
Proportion of the population using an improved sanitation facility	**0.107** (0.002)
Fixed telephone subscription rate	0.000 (0.992)
Mobile telephone usage rate	0.014 (0.837)

Note: P-values based on cluster-level standard errors in parentheses after 500 bootstrap replications. Coefficients were declared statistically significantly different from 0 at 10% if P-values were lower than critical values adjusted by the False Discovery Rate algorithm.