Reconciling evaluations of the Millennium Villages Project[1]

Andrew Gelman, Shira Mitchell, Jeffrey D. Sachs, Sonia Sachs

Original submission: 30 Oct 2020

This revision: 10 Dec 2021

Abstract. The Millennium Villages Project was an integrated rural development program carried out for a decade in 10 clusters of villages in sub-Saharan Africa starting in 2005, and in a few other sites for shorter durations. An evaluation of the 10 main sites compared to retrospectively chosen control sites estimated positive effects on a range of economic, social, and health outcomes (Mitchell et al., 2018). More recently, an outside group performed a prospective controlled (but also non-randomized) evaluation of one of the shorter-duration sites and reported smaller or null results (Masset et al., 2020). Although these two conclusions seem contradictory, the differences can be explained by the fact that Mitchell et al. studied 10 sites where the project was implemented for 10 years, and Masset et al. studied one site with a program lasting less than 5 years, as well as differences in inference and framing. Insights from both evaluations should be valuable in considering future development efforts of this sort. Both studies are consistent with a larger picture of positive average impacts (compared to untreated villages) across a broad range of outcomes, but with effects varying across sites or requiring an adequate duration for impacts to be manifested.

---

[1] To appear in *Statistics and Public Policy.* Corresponding author: Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu. We thank Avi Feller and several reviewers for helpful comments and the U.S. Office of Naval Research for partial support of this work.

1. Background

In 2000, the United Nations set "Millennium Development Goals" (MDGs) for reducing extreme poverty in the world. The Millennium Villages Project (MVP) was launched in 2005 by Columbia University's Earth Institute with the aim of demonstrating the feasibility of achieving the MDGs using an integrated rural development strategy based on proven economic, social, health, and infrastructure interventions that could ultimately be sustained globally within the promised aid budget of 0.7 percent of GDP of the world's donor countries (Sachs and McArthur, 2005). The MVP was applied in clusters of villages in 10 countries of sub-Saharan Africa from 2005 through 2015, and in a few other sites for shorter durations.

The MVP has been controversial, both in its conception and in evaluation of its effects. The starting point for the controversy was the project's approach of economic and social development catalyzed by foreign aid, which has been criticized as a doomed-to-fail relic of a bygone paternalistic era (see, for example, Easterly, 2014). In addition, the MVP was criticized for not being designed as a randomized controlled trial. Clemens and Demombynes (2011) review the difficulty of estimating the impacts of the MVP given its lack of prospective control group. As discussed by de Souza Leão and Eyal (2019), recent decades have seen a resurgence of enthusiasm for randomized controlled trials to study the effect of interventions in international development, as underscored by the 2019 Nobel Prize in economics. The MVP stands out as a high-profile project organized by an academic economist that did *not* include such a control group.

At the inception of the MVP, two reasons were given for not designing the MVP as a randomized controlled trial. First, the MVP used a basket of many interventions that had

already been shown to work, often through previous controlled trials. The main focus of the MVP was on the feasibility of implementing the package of proven interventions within the specified budget and timeline, a concern for which a control group is not relevant. Second, the MVP did not have an adequate project budget to engage systematically with control sites, especially to be able to offer those other sites the package of interventions at a later date. From a pragmatic, political, and ethical point of view, the MVP was therefore wary of identifying and engaging actively with non-project sites.

A related debate is over cost-effectiveness: To the extent that the MVP has been shown to demonstrate an effective low-cost intervention, this provides encouragement for larger-scale programs of this sort; conversely, if any positive effects of these innovations could be achieved using more efficient, inexpensive, and scalable approaches, this would point policymakers to alternative strategies for poverty reduction.

The Earth Institute conducted a retroactive impact evaluation of the MVP's first five years (Pronyk et al., 2012), reporting positive effects on some indicators and not others. The paper made an erroneous claim regarding progress on under-5 mortality relative to the national rural average that was pointed out by Bump et al. (2012) and acknowledged by Pronyk (2012). A few years later, the Earth Institute performed an entirely new evaluation of the full ten-year project (Mitchell et al., 2018), reporting positive impacts in a wide range of poverty and health outcomes, compared to retrospectively-chosen control villages.

More recently, Masset, Hombrados, and Acharya (2020) performed a separate analysis at a single MVP site in operation for 4 years 7 months, in the Savannah Accelerated Development Authority (SADA) region of northern Ghana, and reported mostly small or null results. The

Masset et al. study is based on the results of an independent evaluation of the SADA project managed by Itad (Barnett et al., 2018), funded by the UK Department for International Development (DFID).

The purpose of the present paper is to assess the apparent discrepancy between Mitchell et al., who report consistent positive effects, and Masset et al., who are more pessimistic in their conclusions.

The present authors were involved in the Millennium Villages Project in different ways: Jeffrey Sachs, an economist and former director of the Columbia Earth Institute, was the coordinator and leader of the MVP; Mitchell, a statistician, was brought into the project in 2014 to design and conduct a quantitative evaluation of the program; Gelman, a statistician at Columbia who is also affiliated with the Earth Institute, provided guidance in this effort; and Sonia Sachs, an MD and MPH, oversaw the public health interventions. All of us were among the authors of Mitchell et al. We do our best to assess the evidence and claims of the two papers impartially, while recognizing our involvements in the MVP and its evaluation.

2. Comparison of two evaluations of the Millennium Villages Project

Mitchell et al. (2018) summarize:

> Averaged across the ten project sites, we found that impact estimates for 30 of 40 outcomes were significant (95% uncertainty intervals [UIs] for these outcomes excluded zero) and favoured the project villages. . . . The MVP had favourable impacts on outcomes in all MDG areas, consistent with an integrated rural development approach.

> The greatest effects were in agriculture and health, suggesting support for the project's emphasis on agriculture and health systems strengthening.

In contrast, Masset et al. (2020) conclude:

> Our study finds that the impact of MVP on the MDGs was limited, and that core welfare indicators such as monetary poverty, child mortality and under-nutrition were not affected. . . . despite some positive impacts, we found mostly null results, suggesting that the intervention was ineffective.

Both of these were serious studies conducted by comparing outcomes in Millennium Villages to matched control villages, attempting to adjust for pre-treatment differences between treated and control groups. So how can we understand the starkly different conclusions? In this paper, we consider several differences between the studies. First, we summarize the methods in both papers.

Methods in Mitchell et al. (2018). Mitchell et al. aimed to estimate the MVP's impact in the 10 main sites, where the project was applied from 2005 to 2015. These sites are clusters of 3 to 28 villages, and were chosen non-randomly, without random assignment into treatment versus control. In 2015, they retrospectively selected comparison villages that in 2005 best matched project sites on possible confounding variables. They chose 5 comparison villages per project site, balancing statistical power and budget.

They collected cross-sectional survey data on 40 outcomes of interest from both the project and the comparison villages. They randomly sampled 300 households in each site and comparison group. They captured household-level data by a household survey. Within these

sampled households, they captured person-level data by a sex-specific adult survey, malaria and anemia testing, and anthropometric measurements.

They report raw differences between project and comparison for each outcome and site. They also took standardized averages across related outcomes to create 8 outcome indices. They fit a Bayesian hierarchical model to obtain site-specific and outcome-index-specific estimates based on information from all sites and outcomes. This model includes parameters that vary by country and village, accounting for random shocks at these levels, with weak priors on the hyperparameters so that the amount of partial pooling was determined by the data; for details see the appendix of Mitchell et al. (2018).

Methods in Masset et al. (2020). Masset et al. aimed to estimate the MVP's impact in the Northern Ghana site, a cluster of 35 villages, where the project was applied from 2012 to 2016. This site was also chosen non-randomly, without random assignment into treatment versus control. In 2012, they prospectively selected comparison villages based on village-level characteristics from the 2000 and 2010 censuses, along with additional field data. They chose two comparison villages per project village, with one near the project and the other far from the project. They then did further matching at the household level.

In 2012, they collected baseline data from sample of 755 project households and 1496 comparison households. They collected follow-up rounds each year from 2013-2016, with less than 5% attrition. Similarly to Mitchell et al., they captured household-level data by a household survey, and person-level data by a sex-specific adult survey, malaria testing, and anthropometric measurements.

They estimate impacts using a difference-in-difference regression within subclasses of the propensity score (see, e.g., Angrist and Pischke, 2009).

Differing time horizons. As noted above, Mitchell et al. analyzed effects of a program applied from 2005 to 2015, whereas the program studied by Masset et al. ran from 2012 to 2016. Comparing time periods is a challenge without further data analysis (for example, one might want to look at outcomes after just the first five years of the main MVP study), but we might expect much larger impacts on some metrics from a 10-year program than from one that ran for less than 5 years. The first two to three years of the MVP involved the construction of schools, clinics, roads, and other basic infrastructure, and recruitment and training of personnel in health, education, agriculture, and infrastructure management. Since the MVP was based on implementing and operating public systems in many sectors for which the basic infrastructure is a necessary starting point, it is natural that these systems take several years to bring into operation and even longer to refine those operations in line with experience. Future work could attempt to compare metrics that are more linked to infrastructure demands versus those that are not.

When the SADA MVP was launched, none of the major participants (including DFID, the MVP, and the government of Ghana) expected that 5 years would be sufficient to achieve the MDGs. But all parties agreed to move forward, as it was felt that even the shorter project would benefit the SADA region in light of its impoverishment.

Different numbers of sites. In 2005–2006, the Millennium Villages Project was initiated at 14 different sites in Africa. Mitchell et al. analyzed results from 10 of these sites; the other four were not scaled up or were discontinued because of funding constraints or regional conflict.

Masset et al. analyzed the final (15th) Millennium Village site added to the project, located in northern Ghana (not the same location as the Ghana villages which were one of the 10 locations analyzed by Mitchell et al.). To get a handle on the effect of considering just one location compared to 10 locations, we start with Figure 1, which displays separate estimates for each site, from Mitchell et al. (2018). We see substantial site-to-site variability in treatment effect estimates across outcomes.

In general, distributions of outcomes differ by geography, regardless of treatment. To account for this, the model in Mitchell et al. includes varying coefficients for villages and countries in a multilevel regression. Masset et al. account for the hierarchical structure of the data by computing standard errors that are clustered by village.

In both studies, villages are either entirely treated or not, and treatment villages are matched to control villages within the same country. So while country effects are shared across treatment and control groups, village effects are not. As Imbens (2014) points out, with only one treatment village and one control village, the treatment effect cannot be separated from the difference in village effects. Luckily, in both studies, there are more than one village per treatment group. In Mitchell et al., there are 3 to 28 villages per country-treatment-group and 10 countries, and in Masset et al. there are 35 to 68 villages per country-treatment-group and one country.

One could be concerned that in both studies, the treatment villages are relatively close to each other. The project describes these as a "cluster." Neither study takes into account spatial correlations beyond village (and country) effects. This could mean that both studies

underestimate statistical uncertainty, and that their discrepancies could be attributed to statistical imprecision.

If there were a cluster-level effect, the 10-site study in Mitchell et al. would be better-equipped to estimate the overall average treatment effect, while the one-site study in Masset et al. would be stuck with the lack of identification discussed by Imbens (2014). Without such cluster-level effects, the site-to-site variation seen in Mitchell et al. could be attributed to treatment effect variation (to the extent we believe unconfoundedness). Thus, differences between the two studies could arise both from site-to-site variation in treatment effects and from cluster-level effects. In that context, the apparent null findings from Masset et al. can be attributed to their using less data.

Prospective or retrospective design. Both studies assigned treatment non-randomly, but a key strength of the study conducted by Masset et al. (2020) is that it is prospective: control villages were chosen at the start. In contrast, Mitchell et al. (2018) conducted a retrospective study, imitating as best as possible a prospective design by matching treated and control villages only based on information that could have been available in 2005 at the start of the intervention or which could not have been affected by the intervention. Masset's prospective approach enabled them to collect more baseline data to adjust for possible confounding. Therefore, confounding may account for some of the difference in results between the two studies.

Another advantage of Masset et al.'s prospective design is that they were able to collect data at each site in each year. Even if we have disagreements of how they analyzed these data, it is a strength of that study that yearly estimates of outcomes in treated and control villages are

available, including for additional analyses. It is a tradeoff that this prospective study was only performed at one location covering a short time period, making it difficult to detect effects that are variable.

Choices in modeling and inferential summaries. We have concerns with the difference-in-difference regressions of Masset et al., which specifies a treatment effect that does not vary by time (see their equations (3.1)–(3.2)), hence if the program has cumulative effects that vary over time, as would be expected, the result would be to underestimate the effect over the full period.

Furthermore, the difference-in-difference assumptions may be less attractive than assuming unconfoundedness given the baseline outcome (Imbens and Wooldridge, 2009, p.70). As mentioned above, Mitchell et al. (2018) did not have adequate baseline data with which to use either difference-in-differences or unconfoundedness given the baseline outcome. Instead they selected comparison villages that best matched project sites on available baseline data from Demographic and Health Survey (DHS) and geographic information system (GIS) databases. These possible confounding variables were only available at the area level, limiting the number of data points available to estimate propensity scores. Instead, they matched on indices of related variables. In contrast, Masset et al. (2020) used their richer baseline data to estimate propensity scores. They then used these propensity scores for subclassification, a type of matching method (Stuart, 2010).

Thus, both Mitchell et al. (2018) and Masset et al. (2020) combine matching with regression, using the data they have available. As mentioned above, Masset et al. (2020) has richer data to adjust for possible confounding. However, we think their difference-in-differences regression

could be improved by allowing treatment effects to vary over time, including baseline outcome as a covariate, and using hierarchical modeling to better describe statistical uncertainty.

Given the inherent noisiness in estimates for a single site over a short time period, we feel it was a mistake for Masset et al. to summarize their findings in terms of statistical significance (for example, "the count of statistically significant impacts is low") or to report non-significant comparisons as if they were zero (for example, "we found mostly null results, suggesting that the intervention was ineffective"); this latter is a statistical fallacy, as discussed by Gelman, Carlin, and Nallamothu (2019). These concerns do not invalidate the study as a whole, just the interpretations of some of the results.

Masset et al. use a statistical procedure to control the false discovery rate at the 10% level. As they say, this results in fewer reports of statistical significance. They then interpret non-significant comparisons as if they are zero, a statistical fallacy. Mitchell et al. address the issue of multiple comparisons as recommended in Gelman et al. (2012), considering countries and outcomes jointly to reduce statistical uncertainty through combining data. They fit a Bayesian hierarchical model to obtain site-specific and outcome-index-specific estimates based on information from all sites and outcomes, and then report estimates and uncertainty intervals rather than using a significance threshold.

Framing the interpretation of results. Much of the difference in the conclusions of the two reports can be explained by differences in framing. On one hand, the report from the Millennium Villages team found improvements in 40 different outcome measures, even if those improvements did not always reach the MDG target; on the other hand, the outside group reported that impacts were limited. Is it a plus that "the project conclusively met one third of its

[MDG] targets" (Mitchell et al., 2018) or a minus that "the impact of MVP on the MDGs was limited" (Masset et al., 2020)?

Much depends on expectations. If we consider the MVP as "a plan for meeting the Millennium Development Goals" (Sachs and McArthur, 2005), then it is indeed a shortfall that after ten years it only met on third of its targets, justifying Masset et al.'s description of the project as "aiming high and falling low." If we consider the MVP as a study of feasibility of implementing a realistic integrated approach to aiding low-income rural areas, then consistently positive average effects across multiple sectors are encouraging, even if the outcomes are variable enough that improved outcomes do not appear in all locations for all measures.

Look again at Figure 1, which shows estimates of effects on the Millennium Villages, compared to retrospective control villages, on several different indexes. The overall positivity of the comparisons can be taken as a sign of the success of the program, but the positive outcomes fell short of the ambitious MDG targets. n any case, the variation across sites on particular outcomes also suggests the importance of local context.

Masset et al. suggest that the MVP is a test of the "big push" solution for Africa recommended by Sachs et al. (2004). Yet they acknowledge that the MVP "was not meant to address all potential sources of the poverty trap," especially those arising at the "macro level," such as national infrastructure required for villages to be connected to the national economy. In fact, the MVP was not designed as a test of the big push hypothesis, but of something much

more limited: the feasibility of integrated rural development, in the face of long-standing skepticism by some that integrated development projects are too complex to implement.

This is the perhaps the main achievement of the MVP: the successful implementation of a multi-sector strategy at low cost. It is notable that such a multi-sector strategy could be implemented at a very local scale even when the country as a whole was unable to mobilize the resources for national-level infrastructure (roads, power, water, health, education and other areas) needed for national success in meeting the MDGs. Masset et al. report the broad scope of activities carried out by the project across health, education, infrastructure, and agriculture, and the background evaluation (Barnett et al., 2018) presents data on the high level of community engagement in the project. Masset et al. criticize the program for using "a parallel structure [to government] to manage its activities," but this can be viewed in a positive light given that the aim was to demonstrate to the SADA government (and governments across Africa in the full project) how to undertake such a village-based program, in close consultation with local and regional officials. It was a demonstration project and training ground for governments to implement such projects through their own structures.

An important difference in interpretation arises from claims about poverty reduction. Both papers report a non-significant impact on household consumption (consumer expenditures). Yet Masset et al. (2020) also reports a significant positive impact on income (see Figures 2a and 2b). While Mitchell et al. (2018) did not have high quality income data and so did not report on incomes, Mitchell et al. instead reported on asset ownership data, finding a positive impact on assets. Masset et al. does not report on assets, though they were measured in their evaluation, as Barnett et al. (2018) reports: "The analysis gives some credence to the notion that income

gains were spent on durable goods, saved in cash or invested in livestock and assets." The implication of both studies, therefore, is that the project achieved gains in income that were translated into saving in consumer durables and other assets. The evaluation in Barnett (2018) notes clear reductions in multi-dimensional poverty (that is, a measure of deprivation across several dimensions beyond income): "MVP produced a considerable reduction in the multidimensional poverty index, and by implication, on multidimensional poverty."

Cost comparisons. Masset et al. suggest that the MVP was not cost effective because of the relatively high spending per impact. They acknowledge, however, that they only have spotty evidence of cost comparisons. We believe their cost analysis does not support their conclusions. The MVP spending of $88 per person per year in the SADA site covered interventions across multiple sectors (health, education, roads, power, water and sanitation, agriculture, community engagement, and others). In the ten sites, MVP spending per person per year averaged $66 per year in the first five years and $25 per year in the second five years. We are not aware of other projects that have delivered this package of core services at lower cost. Assessments of the cost-effectiveness of this spending will depend on estimates of effectiveness in the medium and long term, which returns us to the general point that impacts do not show up consistently in a single site during a short time period, and therefore do not provide the basis for assessing cost-effectiveness.

3. Conclusions

In this paper, we considered several differences between two evaluations of the Millennium Villages Project. Without more data, we cannot identify exactly which study differences explain

how the two studies arrive at disparate conclusions. Nevertheless, we think it is useful to clarify researcher degrees of freedom to aid in interpretation and inform future study design.

The two apparently contradictory evaluations of the Millennium Villages Project are both consistent with a larger picture in which the MVP has positive average effects (compared to untreated villages) across a broad range of outcomes, but with effects that are variable across sites and that require several years to take effect, given that the first few years are focused on infrastructure building, and recruitment and training of staff, before systems implementation.

Different policy implications can be derived from evidence for effects that are positive but small on average but variable in particular instances.

First, expectations should be realistic regarding effect sizes and variability over time and across sites. A program should be highly attuned to local contexts, provide the needed time for implementation, and not be expected to provide a one-shot solution to long-term problems.

Second, analysts should be aware of the potential for learning from multiple sites when performing experimental or quasi-experimental evaluations of interventions and policy choices (Mitchell et al., 2018, Meager, 2019).

The enduring controversy about the evaluation of the Millennium Villages Project suggests that it was a shortcoming of the project not to include a control group in the design from the beginning. Barnett et al. (2018) and Masset et al. (2020) demonstrate how a prospective control group can be built in from the start in future studies, acknowledging the political, practical, and ethical complexities of including control sites in such intervention projects and the need to receive from project donors an adequate program budget for control groups and program evaluation.

References

Angrist, J. D., and Pischke, J. S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Barnett, C., Masset, E., Dogbe, T., Jupp, D., Korboe, D., Acharya, A., Nelson, K., and Eager, R. (2018). The impact evaluation of the Millennium Villages Project: Endline summary report. UK Department for International Development.

Bump, J. B., Clemens, M. A., Demombynes, G, and Haddad, L. (2012). Concerns about the Millennium Villages project report. Lancet 379, 1945.

Clemens, M. A., and Demombynes, G. (2011). When does rigorous impact evaluation make a difference? The case of the Millennium Villages. Journal of Development Effectiveness 3, 305–339.

de Souza Leão, L., and Eyal, G. (2019). The rise of randomized controlled trials in international development in historical perspective. Theory and Society 48, 383–418.

Easterly, W. (2014). Aid amnesia. Foreign Policy, 23 Jan. https://foreignpolicy.com/2014/01/23/aid-amnesia/

Gelman, A., Carlin, J., and Nallamothu, B. (2019). Objective Randomised Blinded Investigation With Optimal Medical Therapy of Angioplasty in Stable Angina (ORBITA) and coronary stents: A case study in the analysis and reporting of clinical trials. American Heart Journal 214, 54-59.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness 5, 189-211.

Imbens, G. (2014). Introducing 'Ask Guido.' World Bank Development Impact blog, 19 Feb. https://blogs.worldbank.org/impactevaluations/introducing-ask-guido

Imbens, G. W., and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, 5-86.

Masset, E., Hombrados, J. G., and Acharya, A. (2020). Aiming high and falling low: The SADA-Northern Ghana Millennium Village Project. Journal of Development Economics 143, 102427.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. American Economic Journal: Applied Economics 11, 57–91.

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, S., Namakula, P., Palm, C. A., and Sachs, J. D. (2018). The Millennium Villages Project: A retrospective, observational, endline evaluation. Lancet Global Health 6, e500–e513.

Pronyk, P. M. (2012). Errors in a paper on the Millennium Villages project. Lancet 379, 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 379, 2179–2188.

Sachs, J. D., and McArthur, J. W. (2005). The Millennium Project: A plan for meeting the Millennium Development Goals. Lancet 365, 347–353.

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science 25, 1-21.

United Nations General Assembly (2000). Resolution 55/2. United Nations Millennium Declaration. United Nations, 18 Sep.

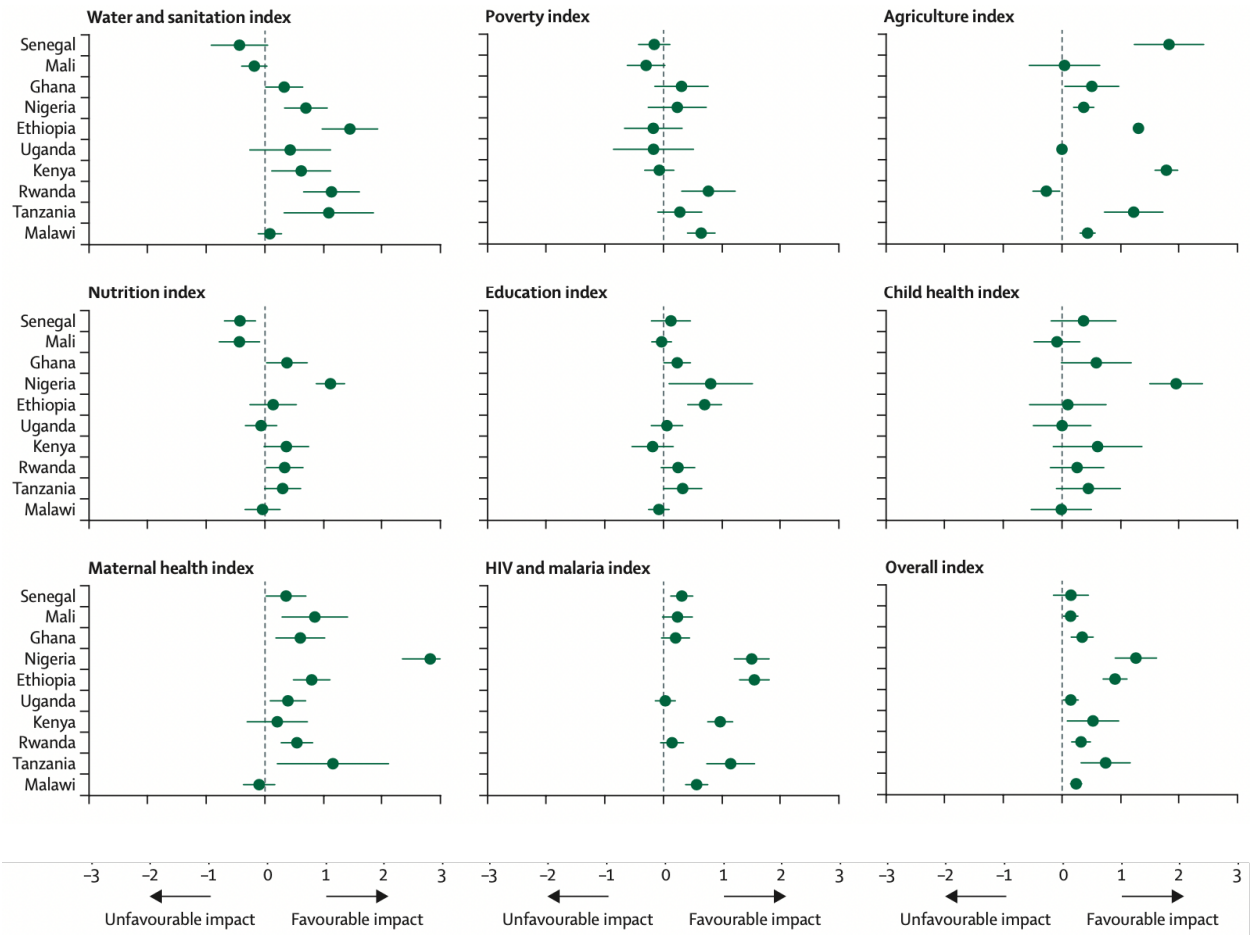

Figure 1: Estimates and uncertainties for the effect of the MVP on eight different indexes and an overall summary, for each of 10 locations, from Mitchell et al. (2018). These graphs show how a positive average effect will not necessarily show up clearly at each site.

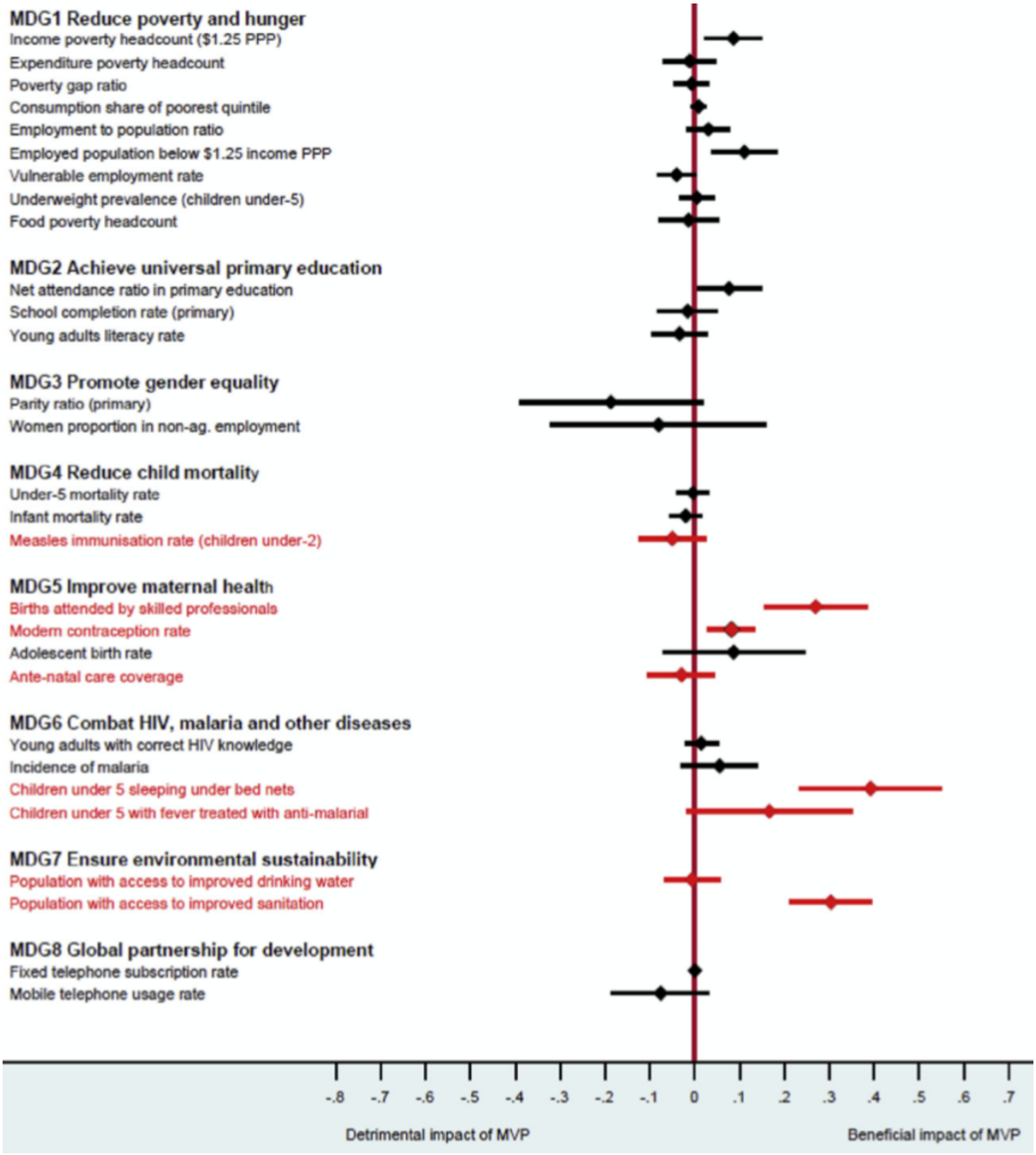

Figure 2a: Estimates and uncertainties for the effect of the MVP on a range of outcomes from Figure 3 from Masset et al. (2020), based on a 5-year intervention in the northern Ghana location. The estimates have high uncertainties, which is expected given that they are based on data from just one site.

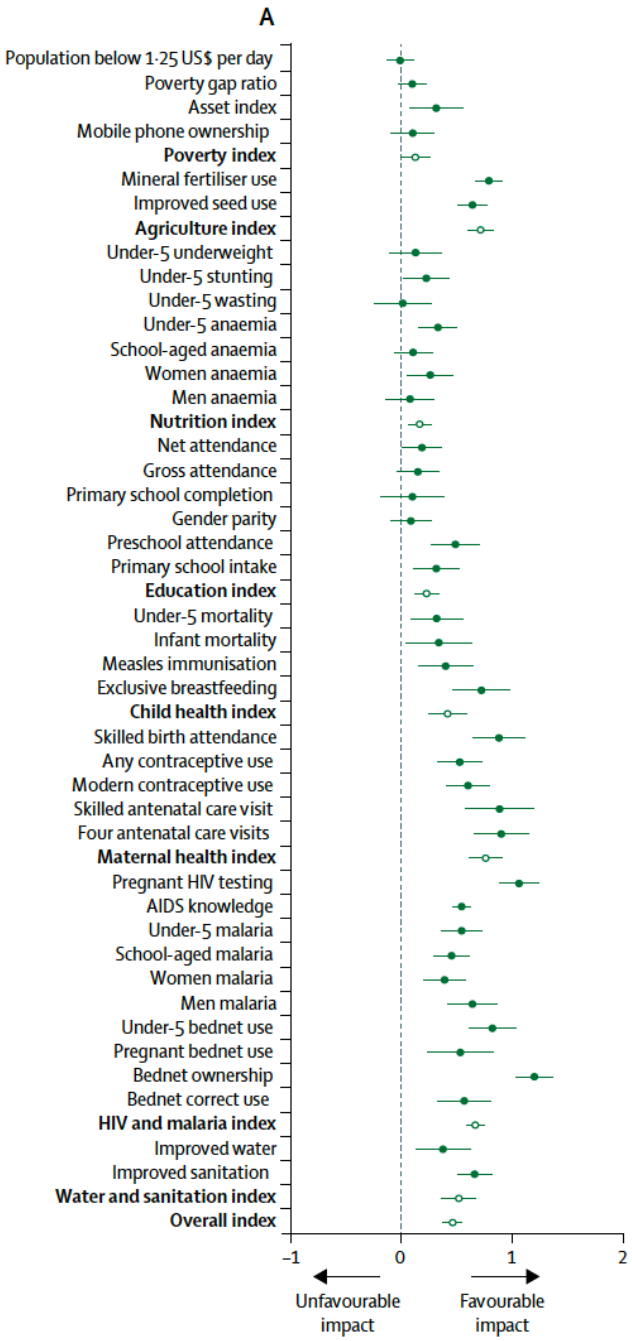

Figure 2b: Estimates and uncertainties for the effect of the MVP on a range of outcomes from Figure 3 from Mitchell et al. (2018), based on a 10-year intervention, averaged across 10 sites.