Reconciling evaluations of the Millennium Villages Project[1]

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

2 March 2020

Abstract

The Millennium Villages Project (MVP) 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). That study was conducted by researchers from Columbia University's Earth Institute under the oversight of an independent body, the African Population and Health Research Center. More recently, an outside group, Itad, performed a controlled evaluation of one of the other sites, in the Savannah Accelerated Development Authority (SADA) region of northern Ghana, and reported smaller or null results (Masset et al., 2020). Although these two conclusions seem contradictory, the differences between the two studies 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 framing of the results and in measurement of poverty. Insights from both evaluations should be valuable in considering future development efforts of this sort. Both studies are consistent with a larger picture in which the MVP had positive average impacts (compared to untreated villages) across a broad range of outcomes, but with effects varying across sites and requiring an adequate duration for the impacts to be manifested.

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

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

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 controlled comparison.

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 controlled comparison is not relevant. Second, the MVP did not have an adequate project budget to engage systematically with control or comparison 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 operations 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 Mitchell et al. and Masset et al.

Mitchell et al. summarize:

> 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. The project conclusively met one third of its targets.

In contrast, Masset et al. 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.

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 should expect much larger impacts 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.

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). The overall effects shown are positive, but in any given location, one sees a range of positive and negative estimates. Consider, for example, Senegal, where three of the indexes have negative estimates and most of the positive effects are not statistically significant. And these are indexes; results for individual components of each index are even more variable.

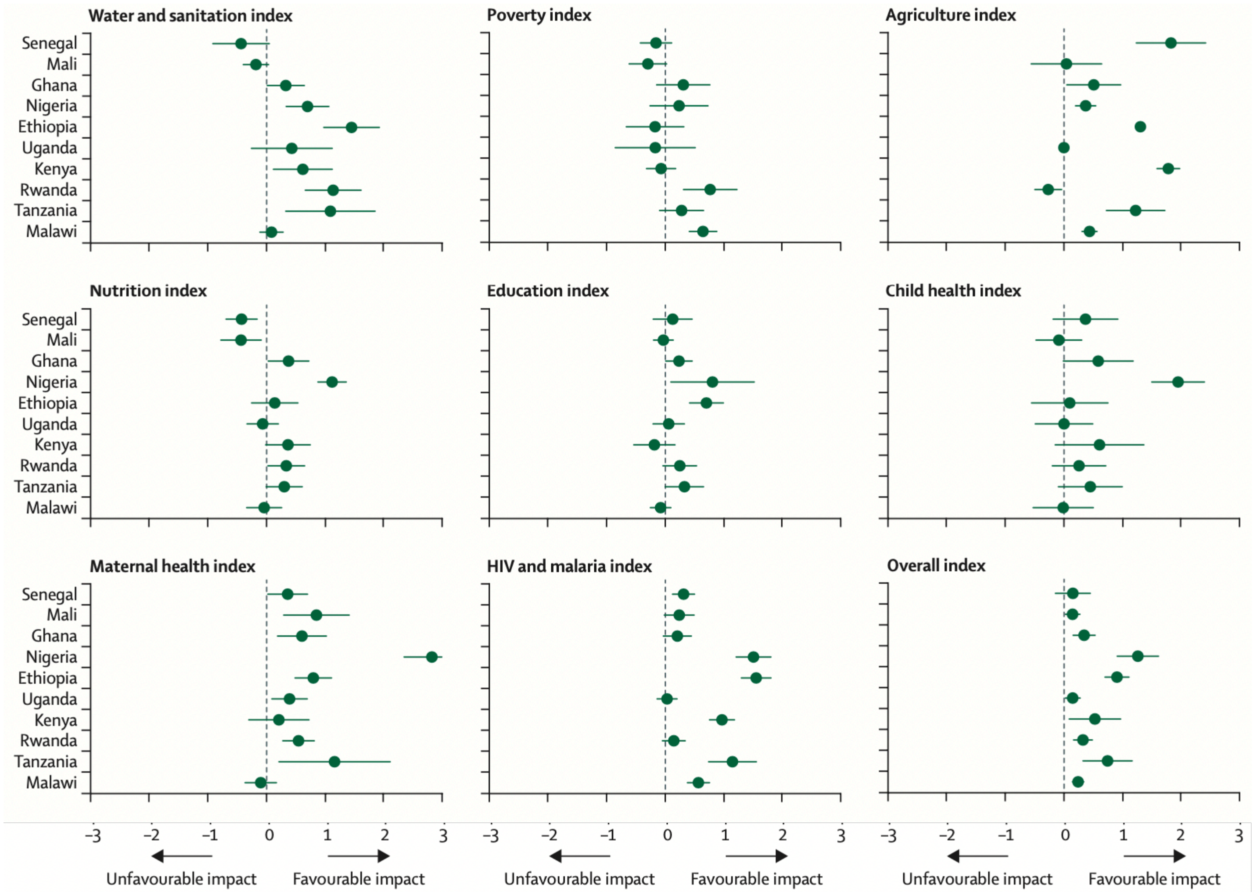

Figure 1: Estimates and uncertainties for the effect of the MVP on eight different indexes and an overall summary, for each of 10 locations, from Mitchell et al. (2018). These graphs show how a positive average effect will not necessarily show up clearly at each site.

For another demonstration of this point, in Figure 2 we reproduce a graph from Masset et al. (2020) that shows multiple-comparisons-adjusted 95% intervals for a range of outcomes estimated from its one site over the five-year period. The wide uncertainties in these graphs demonstrate the challenge of estimating average effects from a small sample.

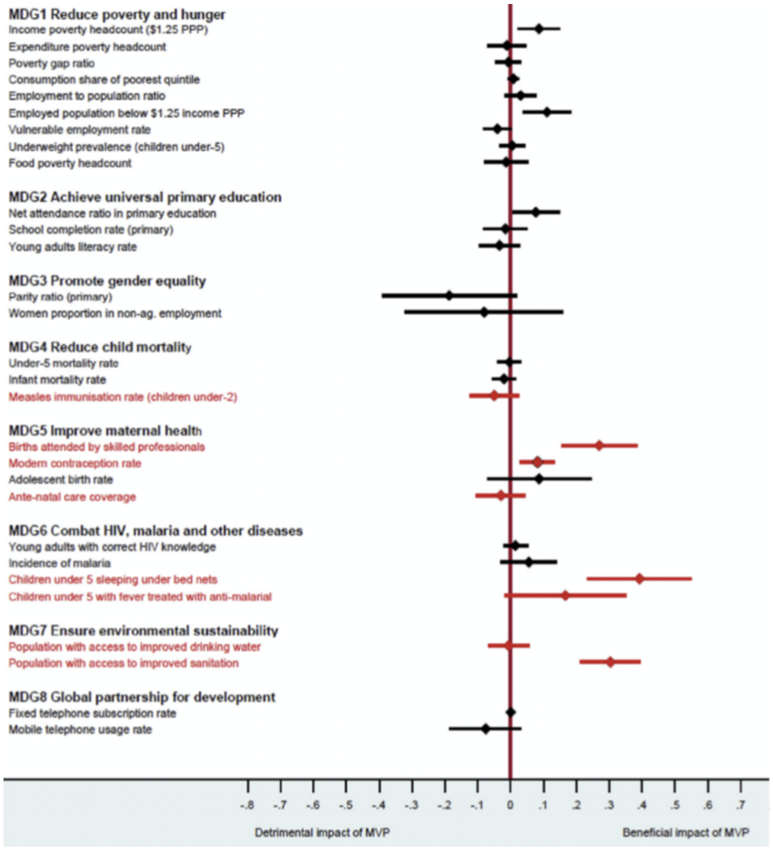

Figure 2: Estimates and uncertainties for the effect of the MVP on a range of outcomes from Masset et al. (2020), based on a 5-year intervention in the northern Ghana location. The estimates have high uncertainties, which is expected given that they are based on data from just one site.

Prospective or retrospective design. A key strength of the study conducted by Barnett et al. (2018) and Masset et al. (2020) is that they are prospective: control villages were chosen at the start, and they were monitored along with the treated villages as the program progressed. 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 we do not believe could have been affected by the intervention. As coauthors of that MVP-organized analysis, we can attest to the care that went into the construction of these retrospective controls; however, we recognize that a prospective design is preferred when possible.

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.

Choices in modeling and inferential summaries. We have concerns with the time series model 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.

In addition to issues with statistical modeling, 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; 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.

Glass half empty or half full. 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" or a minus that "the impact of MVP on the MDGs was limited"?

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 disappointment that after ten years it only met on third of its targets, perhaps 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 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, though the positive outcomes often fell short of the ambitious MDG targets. In 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. 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 how to undertake such a 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.

A final difference in interpretation arises from claims about poverty reduction. Masset et al. report no improvement in household consumption, but do not report on household income, though it was also measured in the ITAD evaluation. While Masset et al. state that "monetary poverty" was not affected by the project, the Itad evaluation shows clear gains in household income: "While consumption (expenditure) did not increase in MV areas more than in CV areas, incomes increased substantially" (Barnett et al., 2018, p. 67). See also Figure 3.

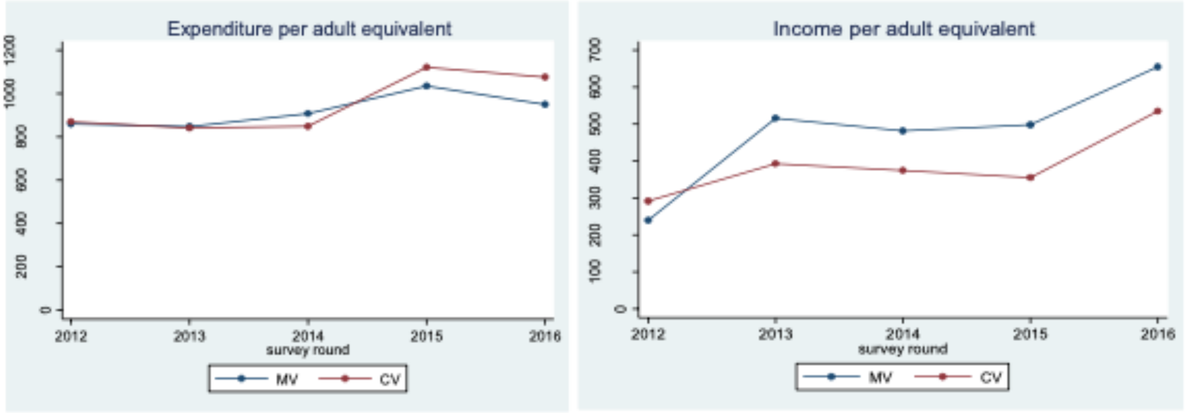

Figure 3: Expenditure and income per adult household in the northern Ghana site. Figure 17 from Barnett et al. (2018).

It appears from a comparison of the income and consumption data in the SADA MVP that most or all of the increase in income was saved in capital assets, notably in livestock, rather than spent on non-durable consumption. This would make sense for households in view of the limited time horizon of the project. Similarly, Mitchell et al. found a positive impact on the MVP on household capital assets. (Mitchell et al. did not report on household income.)

Masset et al. also do not report the Itad conclusion that the MVP significantly reduced multidimensional poverty, a measure that includes deprivation regarding primary schooling, child deaths, malnutrition, and access to electricity, sanitation, improved drinking water, and various household assets. As Barnett et al. (2018) summarize in their report on the SADA site, “MVP produced a considerable reduction in the multidimensional poverty index, and by implication, on multidimensional poverty.”

Cost comparisons. Masset et al. suggest that the project 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 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.

3. Going forward

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 on average but variable in particular instances.

First, expectations should be realistic regarding variability over time and across sites. Programs should be highly attuned to local contexts, and should provide the needed time for implementation.

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 controlled comparison in the design from the beginning. Barnett et al. (2018) and Masset et al. (2020) demonstrate how an ex ante control comparison 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 comparisons and program evaluation.

References

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.

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.

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