A Discussion of Best Practices in Impact Evaluation of Agricultural Projects

By Mackenzie Hickman, Oleg Firsin and Alexandre Monnard

Background:

First, establishing a clear definition of impact evaluation is both necessary and valuable, as some debate exists in the field of development on what exactly constitutes an impact evaluation. One side, which includes Abdul Latif Jameel Poverty Action Lab and a large number of generally large development organizations, is adamant that impact evaluation must tackle the issue of attribution by rigorously identifying a counterfactual to those who receive the benefits of the program. Attribution can be summarized as tying the changes observed in indicators to the actual intervention. Attribution is important, as many other factors can generate changes in indicators; therefore, establishing a clear link allows for an understanding of the causal chain and the success or failure of a given program. This expectation of tackling the issue of attribution means that experimental or quasi-experimental methods are crucial as they allow, when conducted properly, for a reliable comparison between those receiving the benefits of the program and those who do not receive the benefits. Further, they allow for the control of other factors that may have influenced a program's impact.

The other side of the debate is comfortable with conducting impact evaluations that often make little attempt to attribute change in the chosen indicators to the intervention[1]. Thus, that side opposes lifting experimental and quasi-experimental methods above less rigorous and more qualitative methods that are widely used in the evaluation field. This difference in definition is central to the controversy surrounding impact evaluation yet neither of the definitions is right nor wrong in essence; they are fundamentally different. Both definitions have led to useful studies in the field of development. Yet there have been rising concerns that we lack understanding and evidence of what works in development and at

[1] Howard White, "Some Reflections on Current Debates in Impact Evaluation", 2009, p. 7.

what cost. This is particularly critical with the worldwide push for increased transparency in public spending and in development funding. Beyond use of funds, policy decisions benefit greatly from reliable and quantifiable information regarding the effect of a given program.

Because of this discrepancy in definition, many entrenched disagreements have emerged on the methodology of impact evaluation. One such disagreement centers on the reliance on quantitative measures in impact evaluation. During the 1960s and 1970s, it was common for projects to be subject to cost-benefit analysis, which is explicitly based on establishing what outcomes would have been without the project through the use of comparison groups. However, in the 1980s, the focus shifted with the rise of participatory evaluation[2], which rejects evaluation based on criteria set by foreign implementers. It was also often claimed that projects were not suitable to be evaluated using strong quantitative analysis. This change in focus has had some consequences, in that little reliable evidence was produced on how much benefit development spending was generating. A review by 3ie, a non-profit organization focused on impact evaluation, uncovered that out of the 339 evaluation studies found in the evaluative reports database of the Active Learning Network for Accountability and Performance in Humanitarian Action (ALNAP), not one could be classified as a rigorous impact evaluation[3].

This realization, corroborated by other research from a multitude of sources and concerning other repositories of knowledge, has helped to place the emphasis back on the use of quantitative measurements in the field. In recent years, it has been more common to place great emphasis on rigorous impact evaluation. As an example, the World Bank, despite having an early start with the creation of the Independent Evaluation Group (IEG) thirty years ago by Robert McNamara, has added other initiatives fostering the advancement of impact evaluation. For example, it crafted the initiative Development Impact Evaluation (DIME) to provide technical support for operational staff interested in including an impact evaluation in their project design.

For our part, we define impact evaluation for the purpose of this report as the systematic analysis of the significant or lasting changes in people's lives brought about by a given set of actions in relation to a counterfactual. These changes can be positive or negative and intended or unintended. We therefore side with a more constrained and demanding definition that includes the need for a counterfactual, but at the same time, we find value in using qualitative tools to complement a quantitative analysis. It is therefore a relatively centrist view that we hope can find support from both sides of the argument. This view of

[2] Ibid., p. 8.
[3] Ibid., p. 9.

impact evaluation will be expended upon in this report, as we cover the varying methodologies suitable for an impact evaluation, when they should be used and what their advantages are.

The Millennium Challenge Corporation:

Since this guide to best practices in impact evaluation for agricultural projects is written for the Millennium Challenge Corporation (MCC), it is important to spend some time looking at the organization and its defining features. The MCC is a relatively new player in the world of developmental aid. Created in 2004 by an act of Congress, it is the MCC's main mission to promote economic growth and eliminate extreme poverty in low- and lower middle- income countries. To achieve this mission, it forms partnerships with some of the world's poorest countries, as long as they are committed to good governance, economic freedom and investing in their citizens. These partnerships take the form of large-scale grants called Compacts, with a maximum duration of five years. Several things stand out when comparing the MCC with other aid organizations. First, the MCC has a competitive selection process where its Board examines a country's performance on 17 independent and transparent policy indicators to identify which countries are eligible for a grant. It is important to note that countries must maintain strong policy performance during the implementation phase of the Compact, or face a termination of the Compact.

Another standout feature of the MCC is that countries must identify their priorities for achieving sustainable economic growth and poverty reduction. Once a country has written its initial proposal, MCC teams provide support in refining the program and making it more compatible with MCC's goals and objectives. The prioritization of areas for investment by the MCC is different for each country largely for this reason. Continuing on the logic that countries are aware of their own issues and shortcomings, and often better-suited at finding solutions, the MCC also requires countries to lead the implementation of the projects detailed in the Compact, while providing oversight on the use of funds. Through this focus on country ownership and other features in the delivery of aid, the MCC attempts to be in line with the key elements of the Paris Declaration on Aid Effectiveness. This international agreement, ratified in 2005, calls for signatories to meet 56 different partnership commitments around the principles of ownership, alignment, harmonization, results and mutual accountability.

In part because of this system of country-led implementation and in part to increase aid effectiveness in general, the MCC is strongly committed to having a strong M&E and impact evaluation component to all its projects. This means tracking performance on processes and outputs at the beginning of a Compact's life and then to continue to track high-level outcomes and impact to the end of the

Compact. The analytical framework used by the MCC starts with a technical analysis to identify the impediments to growth. This is followed by a benefit-cost analysis, where expected benefits and costs of the program are compared, and the expected economic rate of return (ERR) of the project is calculated. The MCC tends to prefer projects with an ERR of at least 10%. As mentioned earlier, benefit-cost analysis generally requires the definition of a counterfactual, since we need to know what would have happened without MCC's investment. This is followed by a distributional analysis, which looks at how the benefits of the project are anticipated to be distributed within the population.

The MCC, in defining the impact evaluation, states that "an impact evaluation measures the changes in individual, household or community income and well-being that result from a particular project or program." It adds that "the distinctive feature of an impact evaluation is the use of a counterfactual, which identifies what would have happened to the beneficiaries absent the program." It is therefore clear that the MCC falls in the group defining impact evaluation more narrowly and requiring a counterfactual. Importantly, our definition of impact evaluation provided above is very similar and highly compatible with the MCC's definition, which should decrease problems of terminology throughout this report.

Whereas many development organizations share a strong commitment to M&E, it is rare for an organization to have an impact evaluation component in all of its projects. Even organizations like the World Bank, which have been very involved in the debate on impact evaluation, tend not to use comprehensive impact evaluation in all its projects. While the share of projects with impact evaluation has been rising at a steady pace in these organizations, few can match the intense commitment that the MCC has made. This is therefore a defining feature of the MCC and illustrates its position of leader in the field. In terms of spending, the MCC estimates that about 2% of the total Compact amount is spent towards monitoring and evaluation, including impact evaluation.

Methodology

Discussion of Impact Evaluation Methods

As earlier indicated, impact evaluation (IE) is the systematic analysis of the significant or lasting changes in people's lives brought about by a given action or a series of actions in relation to a counterfactual. There are several criteria to consider when choosing statistical methods to estimate impact:

Internal Validity. Internal validity refers to the ability to assert (with a certain degree of confidence) that a program has caused measured results, given other plausible alternative explanations.

Among the threats to internal validity, **selection bias** is perhaps the most significant, referring to systematic difference between the treatment and control groups due to selection, which may also affect the outcome variable. Additional hazards include: **contamination,** including spillovers—positive or negative effects of the intervention on control or comparison group—and effects of programs by other agencies on treatment and/or the control group; **compliance--**non-intake of treatment by some of the individuals who are offered treatment or intake by those not eligible; **selective attrition—**dropping out of the program with systematic correlation with relevant variables**.**

External validity. External validity is the degree to which findings of a given study are generalizeable to other studies and populations. External validity may be threatened by interactive relationship of the intervention with context-specific factors.

Costs. Impact evaluation cost is undoubtedly one of the main factors deterring many organizations from conducting IE rigorously and systematically. The costs include modifications to implementation, required by design specification to enable evaluation, conducting surveys and data gathering, necessary monitoring & evaluation and the actual IE analysis.

Ethical concerns. Given that development projects aim to increase welfare of certain populations, the fairness of aid intervention is a critical concern. Some key ethical concerns are whether it is acceptable to give aid to some people but not to others, and what the criteria for selection should be, as well as how aid is administered. Horizontal and vertical equity are two of the principles considered when designing IE.

Difficulty of implementation and statistical power. Since aid interventions take place in the context of complex interaction between political, economic, and socio-cultural factors, some are more feasible and easy to implement than others. The ability to collect reliable data is of importance. The sample size needed and statistical power—capability of a test to detect significant results, in probability terms—are also of key concern.

There can hardly be an objective algorithm for measuring and weighing different evaluation methods on the totality of criteria noted above. Different organizations may have varying limitations and justifiably focus on different aspects. For small organizations, costs may be, and often are, the deciding factor. Absence of proper expertise may also limit the possibility of implementing certain tools. In some contexts, feasibility of data collection and sample size may be crucial. An organization of MCC's size and with an emphasis on accountability may afford to focus to a large extent on internal validity and difficulty of implementation, which are key for quality impact evaluations. Methods and techniques discussed

below are not necessarily mutually exclusive and can be used in different configurations as complements, substitutes and robustness enhancers; furthermore, different methods can be used for different parts of given projects.

Costs, external validity and ethical concerns perhaps require less discussion here than other factors. The cost of evaluation depends on several factors, including the number of indicators collected, frequency and quality of information sought, and others; whereas development organizations do tend to say that more rigorous methods are also more expensive, there does not seem to be a clear and significant cost differential between experimental and quasi-experimental methods.[4] Additionally, organizations committed to the most rigorous impact evaluations possible, such as the MCC, tend to prioritize evaluation quality over costs. In terms of ethical concerns, randomized control trials have been criticized considerably, but randomized rollout is more ethically acceptable, given that all eligible farmers are supposed to receive treatment in the end, although this has not always happened[5]. PSM and DD have been criticized to a lesser degree, since they do not require an initial group of individuals from which only some would receive immediate treatment; regression discontinuity can often be the most ethically acceptable, since all eligible individuals (above or below the cutoff line) receive the treatment, especially if the selection criteria is income, bolstering horizontal equity. In terms of external validity, it depends to a large extent on the type of intervention and selection criteria for farmers, and is to a lesser degree affected by statistical design. However, it would be understandably hard to replicate PSM in different environments, given the likely variation in the contribution of various factors to selection probability. Internal validity and feasibility of implementation concerns are discussed in more detail below.

Experimental Methods: Randomized Control Trials

Randomization seeks to solve the fundamental problem of establishing the difference between the treatment and control group, and randomized control trials have been called the "gold standard" of impact evaluation. Randomization does not in itself ensure complete internal validity but addresses perhaps the greatest impediment to it—selection bias—when done properly. Differences between the groups may be the result of selection or self-selection into a given program, and may consist of characteristics that may also have impact on the outcome variables. Randomization addresses this problem by making the selection independent of any original characteristics—randomly. Sometimes, randomization is done through computer software, when a program is asked to generate random values for the group from which treatment and control groups are to be drawn. Other times, it may be done ostensibly by picking papers or

[4] "Monitoring and Evaluation: Some Tools, Methods & Approaches," The World Bank, 2004, p. 24.
< www.mfcr.cz/cps/rde/xbcr/.../WB_Evaluation_ME_Tools_2004_pdf.pdf>.
[5] Cancellation of MCC's land titling program in Nicaragua discussed later on is one such example.

balls with names from a hat, or using a similar observable operation. Thus, akin to a coin flip, the outcome of selection is expected to be unrelated to any characteristics of individuals, and they should be approximately equal across both groups, given large numbers of eligible units of observation over which the randomization is to be done.

Special cases of randomization include randomized phase-in (or rollout), when all eligible candidates receive treatment, but the timing is randomly allocated, with later groups being the controls. Within-group randomization is randomization on a smaller scale, with some individuals within each targeted area receiving treatment. An alternative, encouragement design, instead of treatment, randomizes announcements or incentives, with them being used as instruments for take up.

The MCC attempted to use randomized rollout in all of the countries with agricultural projects. The method had to be canceled, however, in a number of projects and activities, as will be discussed later on.

Internal Validity. As mentioned, randomization helps solve the selection bias problem within the class of internal validity concerns**.** This is the single greatest contribution of randomized control trials, since selection bias is perhaps the single greatest concern to validity of results. On the other hand, randomization does not in itself solve contamination and other problems. Contamination is especially dangerous when the treatment is easily transferable.

Such is the case with many of the MCC agricultural projects, which have training components within them. Thus, within-group randomization is not a good option, and randomization needs to be done at a large-enough level to minimize potential contamination effects. Using a larger **unit of randomization** is also reflective of ethical concerns, as it would be difficult to justify giving treatment to some farmers but not other similar farmers within the same community. Additionally, it is significantly cheaper to provide treatment to larger groups of individuals, rather than smaller groups or individuals. Another argument for a greater unit of randomization is that in some cases group interactions and collaboration are parts of what produces the outcome, as is the case in many agricultural projects. A further benefit of using larger groups is strengthening the unit homogeneity assumption: since individuals within the group are given treatment in almost the same way, we can be more certain that the treatment is the same than if the treatment was received individually at different times; of course, the danger that treatment somewhat varies between groups remains.

The sample size concern was reflected in several MCC projects. The unit of randomization had to be as small as possible to maximize independent variance between the units and increase statistical power

(or decrease it by a lesser amount), but it had to be large enough to forestall potential contamination and address the issues discussed above. Since the optimal unit of randomization varies by activity and sector, MCC randomized groups of varying size. In Armenia, the unit of randomization was village, and villages were also grouped into clusters of between 1 and 5 villages, based on geographic proximity. In El Salvador, randomization was done at levels from productive group to municipality, depending on the activity. In Nicaragua, randomization was conducted at the village level. In Ghana, randomization was at the FBO, Farmer-Based Organization, level, while Honduras saw randomization at an "aldea," a farming community level.

Difficulty of implementation and statistical power. Perhaps the key disadvantage of randomized control trials is that they are sometimes not feasible and usually harder to implement. Randomized rollout requires implementers to postpone aid to eligible groups other than the treatment group. This may be difficult to accomplish. First, timing is crucial in proper implementation of randomization. Given a limited amount of time that is allocated to projects, due to set Compact duration, inability to timely distribute treatment to selected individuals may make it impossible to properly compare effects. Second, there may be, and often is, dissatisfaction among farmers with not receiving aid at the same time as others, which may create complications for implementers working with them. Third, if implementation is done by government agencies, they may have incentives to release control groups earlier, before elections, for example. Fourth, if aid is tied to governance, unexpected changes may halt the program. This is likely to affect randomized rollout more than other designs since its last stage includes the control group receiving treatment, which may prolong the project and extend the time in which unfavorable political developments may take place and affect the implementation. Additionally, the assumption of similarity between treatment and control groups comes from the law of large numbers, which requires a large enough sample size. Since both treatment and control groups are drawn from the eligible population, requirements for its size are high and often hard to satisfy. A large sample size needed for impact evaluation with Randomized Control Trials (RCTs) becomes an especially acute issue when randomizing at a group level and using clusters.

The MCC ran into some difficulties in implementing randomized rollout in several countries. Nicaragua illustrates the difficulty presented by timing of randomized rollout in two ways. First, unexpected delay in titling distribution made it impossible to proceed with the original IE design.[6] Second, within the scheduled time of the Compact, elections that were deemed undemocratic took place and induced termination of the property regularization activity. In El Salvador, randomization has

[6] Patricia Toledo and Michael Carter, "Impact of Business Services on the Economic Wellbeing of Small Farmers in Nicaragua," p. 6.

remained the IE methodology for the Production and Business Services activity, but was discontinued due to insufficient demand for Investment Support and Financial Services activities. This situation highlights both the sample size demands of randomization and the need to conduct demand analysis, as discussed in the recommendations. In Georgia, it was originally planned to use randomization, but due to program changes, it was discontinued, switching to quasi-experimental designs instead; randomization design problems in Georgia will also be discussed in the recommendations.

Quasi-Experimental: Propensity Score Matching

Propensity score matching is based on selection of a group most similar to the treatment group in terms of probability of being selected, which is derived from accumulated contributions to selection probability by observed characteristics. Being a technique for constructing a comparison group, PSM is not a complete IE methodology in and of itself, and is matched with different evaluation components, such as single difference or double difference. PSM is often considered the second-best strategy after randomization. The method relies on two conditions: sizeable overlap in propensity scores of treatment and control groups and conditional independence (unobservables do not affect probability of being selected).

Internal validity. Propensity score matching diminishes possible selection bias but does not eliminate it. It addresses selection bias that can come from observable characteristics that are included in the specifications used. As long as only observed and accounted for characteristics affect the probability of being selected and the outcome variable, this method is an appropriate way to find a comparison group. If unobserved heterogeneity (ex. motivation, personality, intelligence) affects both probability of selection and outcomes, estimates may be biased. When using the method, one has to hope or assume that unobservable factors are at least not very significant. It is more justified to do this in some types of projects than in others, and in some specific interventions more than in others. For example, in the further discussed Georgia Agricultural Development Activity, farmers had to have sufficiently high scores on a test, which may be correlated with intelligence and entrepreneurial ability. If instead of randomization, PSM was to be used on observable characteristics to match those that scored between 70 and 85 with a comparison group, there would likely be an omitted variable bias. On the other hand, in Nicaragua, selection criteria included observable factors, such the number of cows of milk producing age, access to water, possession of or access to a farm and others, so matching them with others of similar characteristics could be more, although not completely, justified. Contamination concerns apply here, similarly to randomized control trials, and one has to choose units of analysis sufficiently large and far

apart, and try to measure spillover effects when possible. When doing this, however, the danger arises that localities may be affected by different economic or political dynamics, which may bias results.

Difficulty of implementation and statistical power. PSM may be statistically complex, and may be easier or more difficult to implement than other methods. PSM relies primarily on observable characteristics, and may make extensive use of secondary information sources. It is also not as susceptible to timing problems at the beginning and at the end of a given project, such as the need to select treatment and control groups before any treatment is given and the need to protract the program until all the individuals receive treatment (as in randomized rollout). On the other hand, PSM is dependent on a sizeable overlap in propensity scores of treatment and control groups, which may not be present. Statistical power of PSM and the sample size needed rely on the range of propensity score deviation that can be used to match the treatment group units--the greater the range, the greater the statistical power, but the lower the validity of estimates.

Difference-in-Differences

Difference-in-differences can be measured in two ways, yielding the same result. The first is to estimate the difference in the outcome variable between the treatment and the comparison groups before the intervention and after, and then subtract the former from the latter. The second is to estimate the change in the value of the outcome variable before and after the intervention for the treatment group, and the change in values for the same period for the comparison group, and subtract the latter from the former. The assumption is that without the intervention, the change would be the same for the treatment and the comparison group, so the difference is the result of the treatment. Double difference is not a clear alternative to other methods, as it says nothing about the selection of the control/comparison group. It is often used in conjunction with propensity score matching, where DD is estimated for the matched groups, for which one needs panel data and a large set of observed characteristics.

Internal validity. Double difference does not eliminate selection bias, but assumes that potential unobserved heterogeneity between treatment and comparison groups is time invariant. When this does not hold, DD introduces a bias in estimation. The assumption of the same trends in the change of the outcome variable for the treatment and control groups without the intervention may also not hold due to preexisting trend differences, suggesting the need to analyze earlier trends.

Difficulty of implementation and statistical power. Complexity of implementation depends to a large extent on the way a comparison group is selected. It may be easier to do DD without randomization or PSM if the comparison group is chosen with looser criteria. In this case, the statistical power of DD is

greater and the sample size needed is smaller than with other methods, but the quality of the comparison group may be unsatisfactory.

Regression Discontinuity

Regression discontinuity relies on comparison of groups just below and just above a threshold of eligibility. The groups clustered closely to the threshold of eligibility are expected to be almost indistinguishable in other characteristics and, therefore, provide a good basis for comparison. The treatment group is the one just meeting the selection criteria and the comparison is the one just outside of it. One advantage of this method is the relative straightforwardness of selection of the treatment and control group along with few additional assumptions, although this depends on the range selected around the threshold that is being utilized.

Internal validity. The main threat to internal validity, when using RD, is that the groups on the different sides of the cutoff are not actually identical. This can happen when the distance to the cutoff is large enough and is correlated with other variables.

Difficulty of implementation and statistical power. The main requirement for the use of this method is the presence of a clear selection threshold—which is adhered to—and a large enough number of individuals sufficiently close to this threshold. Often, it is difficult to find a large enough sample and to have a strict cut-off for one or several criteria that determine selection. The need for a larger sample size needed for this method flows from its low statistical power, as even in a nonclustered design one needs a sample size at least 2.75 times greater than is used in randomization, which generally carries over and is even greater for clustered design.[7]

Literature Review Part I: Organizational Overview

Most large international aid organizations conduct impact evaluations of at least some projects. Some of those conducting quality impact evaluations and extensive research on best practices in impact evaluation include the World Bank (WB), the International Food Policy Research Institute (IFPRI), the International Fund for Agricultural Development (IFAD), the Food and Agriculture Organization (FAO), the Abdul Latif Jameel Poverty Action Lab (J-PAL), the Innovations in Poverty Action (IPA), the Center of Evaluation for Global Action (CEGA), and the International Initiative for Impact Evaluation (3ie).

[7] Peter Z. Schochet, "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations," Institute of Education Sciences, August 2008, p. 33. <ies.ed.gov/ncee/pdf/20084026.pdf>.

J-PAL not only strongly advocates randomized design for impact evaluation, citing it as the best way to create a statistically identical comparison group and inform poverty alleviation programs,[8] but it can be stated that promotion and implementation of Randomized Evaluations (REs) is the main reason for J-PAL's existence. The Lab not only conducts REs itself but also assists other organizations in conducting REs. Since it is a young organization, established in 2003, most of its impact evaluations are ongoing, but some have already been completed.

Finding Missing Markets (and a Disturbing Epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya[9]

The project was conducted by a Kenyan NGO DrumNet with assistance from J-PAL. Designed as a randomized trial, the project sought to help farmers export high-value crops. The intervention was two-fold: service package with extension and marketing help, and agricultural credit. The design allowed for measuring both activities together and separately. DrumNet assisted farmers through training, helping them open bank accounts for commercial transactions and providing assurance to both producers and purchasers that the other party can be trusted. Researchers found positive but not overwhelming one-year impacts from this intervention, as farmers were 19 percentage points more likely to be growing export crops, with increased production and lower marketing costs. Impact on income was not statistically significant for the entire sample, but was statistically and economically significant (32%) for first-time growers of export-oriented crops. The effect of credit was found to be insignificant. They also measured impact on prices of non-export crops (through switching to exports), and were lead to believe there was no significant impact. One year after the intervention ended, the purchaser of horticulture produce stopped buying it because of a lack of compliance with European export requirements (EurepGap). Consequently, the program collapsed, as farmers had to undersell the produce, with unsellable crops rotting, and farmers being forced to default on their loans.

Key lessons learned through evaluating this intervention were that it is imperative to examine the market where exports would be directed, as well as new requirements forthcoming and whether they could be met. The IE also suggests that impact on welfare of those without previous access to export markets is much greater, and may be worth focusing on, as compared to those already exporting. Finally,

[8] "Methodology: Why Randomize," Abdul Latif Jameel Poverty Action Lab, <http://www.povertyactionlab.org/methodology/why/why-randomize>.

[9] Nava Ashraf, Xavier Gine and Dean Karlan, "Finding Missing Markets (and a Disturbing Epiloegue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya," Abdul Latif Jameel Poverty Action Lab, 2009 <http://www.povertyactionlab.org/sites/default/files/publications/Finding%20Missing%20Markets.pdf>.

access to credit did not appear to have a significant impact, possibly because those already exporting found other ways of exporting.

How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya[10]

Beginning in July 2000, a series of six field trials were conducted to test the profitability of fertilizer on farms in a region of Western Kenya. The project was conducted in cooperation with International Child Support (ICS), a Dutch NGO. Farmers were randomly selected from lists of parents with students enrolled in local schools. ICS paid for fertilizer and hybrid seeds, delivered materials, helped farmers apply fertilizer and seeds, and assisted them with the harvest, with each control farm being right next to the comparison plot, farmed by regular techniques. There were 4 different types of fertilizer applications: A) ¼ tsp. of Calcium Ammonium Nitrate two months after planting, B) and C)—1/2 tsp and 1 tsp, respectively, both in 2 months after planting, and D), hybrid seeds and 1 tsp. of fertilizer at planting and another 2 months later; the latter being recommended by the Ministry of Agriculture. IE found that interventions increased yield by 28-91%, but that the rate of return was 8.4% for A, 69.5% for B, -17.8% for C and -48.2% for D. Another finding was that offering farmers to buy fertilizer immediately after the harvest leads to a 33% increase in the proportion of farmers using fertilizer, whereas no such increase ensues when offering fertilizer at the time when it is optimal to apply it, suggesting an element of non-fully rational behavior, otherwise taken as given.

Lessons learned from this IE include that fertilizer can have a significant positive or a significant negative return, depending on the amount and time of application. It is also evident that the government agency recommended the worst application strategy, suggesting a need for caution when evaluating government recommendations. Additionally, given that farmers do not tend to use fertilizer at optimal times, detailed instruction as to the application is highly valuable.

The World Bank

Being the largest international donor with the greatest number of projects, the World Bank is able to conduct impact evaluation on only a portion of its projects. When possible, the WB sees randomization as the most accurate evaluation design.[11] The IEG (WB's Independent Evaluation Group) notes, however,

[10]Esther Duflo (MIT), Michael Kremer (Harvard), and Jonathan Robinson (UCSC), "How High are Rates of Return of Fertilizer? Evidence from Field Experiments in Kenya," January 2008, p. 3. <http://www.povertyactionlab.org/sites/default/files/publications/Duflo%2C%20Kremer%2C%20Robinson-%20How%20High%20Are%20Rates%20of%20Return%20to%20Fertilizer%20in%20Kenya-%202008.pdf www.worldbank.org/ieg/ecd/conduct_qual_impact_eval.html>.

[11] Michael Bamberger, "Conducting Quality Impact Evaluations Under Budget, Time and Data Constraints," The World Bank Independent Evaluation Group, 2006, p. 3. <www.worldbank.org/ieg/ecd/conduct_qual_impact_eval.html>.

that often it is not possible to conduct randomization. As the next best option, it encourages propensity score matching (PSM) with pre and post measurements for both treatment and control groups, which are cheaper and simpler to implement. In some projects, the Bank uses regression discontinuity and multiple comparison group designs.[12] The most common evaluation structure is that used in Program Performance Assessment Reports (PPARs), conducted on approximately 1 in 4 projects, where programs are rated on 3 main criteria: relevance, efficacy, and efficiency. Assessed aspects also include outcome, risk to development outcome, bank performance and borrower performance. Programs and components are evaluated using categorical terms, such as highly satisfactory, satisfactory, unsatisfactory, and highly unsatisfactory. More rigorous evaluations of economic impact, such as ERR, are also conducted. Despite some attention to impact, PPARs in their essence seem to be process evaluations, rather than impact evaluations.

Project Performance Assessment Report, Republic of Azerbaijan. Farm Privatization Project. Agricultural Development and Credit Project[13]

Whereas the Farm Privatization Project is more area-specific, focusing on accelerating land privatization program, the Agricultural Development and Credit Project (ADCP) is more similar to programs undertaken elsewhere in the world. It was the first phase of a three-phase Adjustable Program Loan. It sought to raise agricultural productivity by consolidating a land reform, agricultural extension and credit. Both programs were used as a package, with privatization bolstered by extension services.

The Bank, among other tools, used difference-in-differences to estimate the impact. However, M&E included a lot of output indicators, but not outcome indicators such as income. Consequently, even though program farmers increased their production by 45% compared to 10% increase for non-program farmers, an impact on welfare could not be established. Likewise, it was not possible to estimate an economic rate of return. Despite this, program outcome was rated satisfactory.

As lessons learned, World Bank mentions the success of packaging agricultural services with land reform, benefits of a rapid single-step change rather than gradual phased implementation, the importance of client focus, transparency, and stakeholder involvement, as well as strong commitment from the government.

---

[12] Ibid., p. 11.
[13]Vinod Thomas, "Project Performance Assessment Report, Republic of Azerbaijan. Farm Privatization Project. Agricultural Development and Credit Project," The World Bank Independent Evaluation Group, July 2008. <http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2008/08/20/000333038_20080820011126/Rendered/PDF/448310PPAR0P0410Box334040B01PUBLIC1.pdf>.

Agricultural Extension: The Kenya Experience. An Impact Evaluation[14]

Compared to PPARs, this is a longer report, more focused on impact. It examines efficacy and efficiency of the Training & Visit (T & V) system, which aimed to achieve two goals: institutional development of the extension service and sustained increases in agricultural productivity. The evaluation adopted a theory-based approach, linking inputs, key indicators at various stages of the results chain and outcomes. The main evaluation method was multivariate regressions, comparing those farmers that had access to extension services and those that did not (without randomization), estimating effects of the frequency of staff visits, farm size, and other factors. The evaluation found limited institutional development impact, some impact on increased geographical coverage, and improved staff quality. Overall, however, the evaluation found the system to be ineffective and inefficient in delivering the needed services to farmers. A positive rate of return was not found. Further, it was found to be financially unsustainable. Notably, an overwhelming proportion (80%) of the operational budget was consumed by staff salaries. Cost-effectiveness was found to have been reduced by providing general information to farmers, which they mostly already knew, instead of more advanced and context-specific knowledge. The evaluation reports that a very large proportion of farmers who were introduced to even the more complex practices adopted them, indicating that lack of information is an important constraint, even allowing for possible credit constraint. Inefficiency stemmed from allocating staff to previously more productive areas, while the greatest growth was in the previously less-productive areas.

The main lessons that Operations Evaluation Department drew from the program were the need for more efficient targeting to areas with the greatest marginal impact (through identifying gaps between average and best practices), the need for timely information flows and continuous evaluation for timely feedback, the need for extension services to be suited to particular circumstance (in particular, reallocation of staff to specific areas so as to increase geographic coverage), that uniform methodology in all circumstances is unlikely to be effective, and the need to fully incorporate client focus, including decentralization and responding to farmer demand, potentially through cost sharing, with its proper incentives and budgetary respite (contingent valuation through willingness to pay is mentioned as one way to estimate demand for services).

[14]Madhur Gautam, "Agricultural Extension: The Kenya Experience. An Impact Evaluation," July 2000. <http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/08/19/000094946_00080705302026/Rendered/PDF/multi_page.pdf>.

IFPRI conducts extensive research on multiple areas of agricultural development, including impact evaluation. Like J-PAL, it prefers more rigorous experimental and quasi-experimental methods to non-experimental methods, along with measuring social benefits and social costs of projects, rather than simply outputs.

Assessing the Impact of the National Agricultural Advisory System (NAADS) in the Uganda Rural Livelihoods[15]

The NAADS program is an innovative public-private extension service delivery approach that became operational in 2001. It promoted development of farmer organizations and empowered them to procure advisory services, manage linkage with marketing partners and conduct demand-driven M&E of advisory services and their impact. The main goal was promoting market-oriented production by empowering farmers to demand and control agricultural advisory and information services. Communities and households were selected using a two-stage stratified random sampling, with strata based on NAADS rollout phases.

The impact was analyzed by examining the change between 2000 and 2004 across the three strata, controlling for several factors, including inflation and temporal monetary and fiscal trends, and weighting the village clusters by parish-level human population data. IE showed positive effect on new crop adoption, as over 50% of NAADS sub-counties had adopted at least 1 new crop, compared to only 32% of non-NAADS counties; total area under cultivation also increased for NAADS sub-counties more than for the control group. Additionally, the rate of adoption of new technology was higher for NAADS farms among those whose rate of awareness of new technologies improved. However, for most crops, there was no significant difference in yields. Analysis of income effects showed that NAADS sub-counties were positively impacted by the program, as in the face of general income decline in rural areas of Uganda between 2000 and 2004, the decline in these sub-counties was the smallest—15%, compared to 32% and 28% in the control groups, with differences being statistically significant at $P<0.05$ level. Given the absence of significant difference in crop yields, researchers conclude that the probable reason for avoiding significant decline in income was due to diversification into profitable new farming enterprises. One notable finding was that farmer groups seemed to achieve economies of scale by pooling resources together. Researchers find it unsurprising since group formation was an old and established

[15] Samuel Benin et al., "Assessing the Impact of the National Agricultural Advisory Services (NAADS) in the Uganda Rural Livelihoods," October 2007. <http://www.ifpri.org/sites/default/files/publications/ifpridp00724.pdf>.

concept and NGOs and programs had been contributing to strengthening this type of organization in Uganda.

International Fund for Agricultural Development

IFAD aims to construct its evaluations in line with the major agreements from the Paris Declaration on Aid Effectiveness.[16] As the World Bank, IFAD uses three main criteria for evaluation: relevance, effectiveness and efficiency. Specific criteria used to gauge rural poverty impact include five main domains: household income and assets; human and social capital and empowerment; food security and agricultural productivity; natural resources and the environment; and institutions and policies. Sustainability and innovation are also evaluated. Evaluated projects fall into two main categories—satisfactory and unsatisfactory—with six more specific assessment scores: 6—highly satisfactory, 5—satisfactory, 4—moderately satisfactory, 3—moderately unsatisfactory, 2—unsatisfactory, 1—highly unsatisfactory. Like other organizations, IFAD sees random sampling as the most rigorous way of creating a counterfactual,[17] and sees quasi-experimental methods as the next best method.

Literature Review Part II: Best Practices Literature

There are several major, pervasive themes that are stressed in recent literature on best practices in impact evaluation. Firstly, while there is an acknowledgement of the benefits of random assignment in addressing the problem of selection bias, researchers note that simply using random assignment does not guarantee the best possible practices, and additional requirements apply. In particular, the literature stresses theory-based impact evaluation and the use of mixed methods.

In "Theory-Based Impact Evaluation: Principles and Practice," executive director of the International Initiative for Impact Evaluation (3ie) Howard White, explains the benefits and principles of implementation of theory-based impact evaluation.[18] He stresses that the mantra of supporting the move to better impact evaluation is not just to understand what works, but also why. He quotes the 3ie impact evaluation guide saying, "Studies should clearly lay out how it is that the intervention (inputs) is expected to affect final outcomes, and test each link (assumption) from inputs to outcomes (sometimes referred to

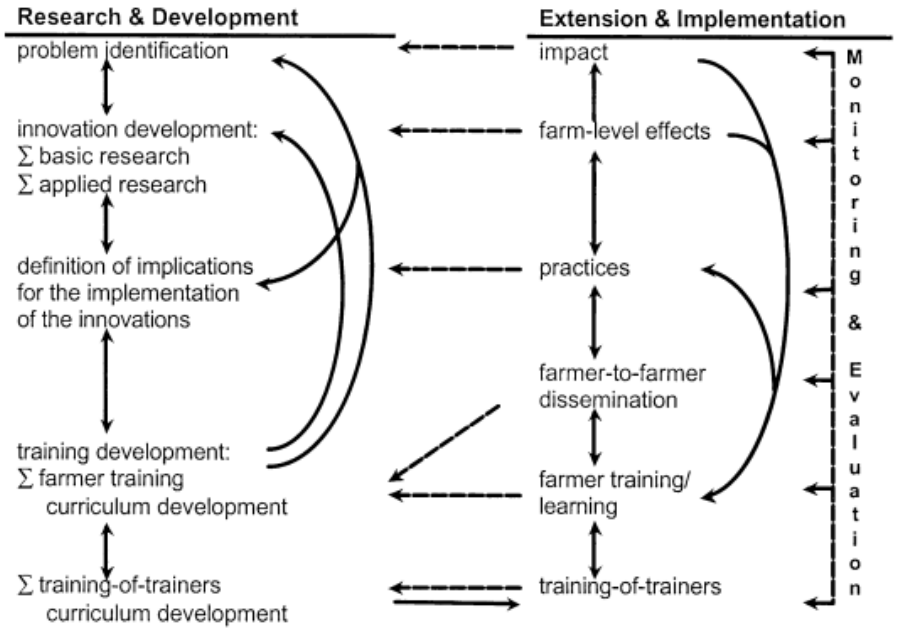
[16] "Evaluation Manual: Methodology and Processes," Office of Evaluation IFAD, April 2009, p. 8. <http://www.ifad.org/evaluation/process_methodology/doc/manual.pdf>.
[17] Ibid., p. 19.
[18] Howard White, "Theory-Based Impact Evaluation: Principles and Practice," International Initiative for Impact Evaluation Working Paper 3, June 2009. <http://www.3ieimpact.org/temp2.php?path=pdfs_papers/51.pdf&name=Theory-Based%20Impact%20Evaluation%20by%20Howard%20White>.

as the program theory). The evaluation design should incorporate analysis of the causal chain from inputs to impacts."[19]

While White notes that theory-based impact evaluation (TBIE) is not new in principle, few studies to date meet the requirements of such an evaluation. To bridge the gap between theory and application, he outlines main steps and principles behind theory-based impact evaluation. The six key principles are: 1) map out the causal chain, linking inputs to outputs and outcomes. When doing this, theory should not be static, but be ready to adapt to surprises. Preliminary participatory analysis is crucial for predicting unintended outcome, as is literature on previous experiences with similar interventions; 2) understand context, include social, economic and political factors related that may affect the causal chain steps; 3) anticipate heterogeneity, predicting varying impact based on beneficiary or geographic characteristics (this step may also inform external validity assessment); 4) rigorous evaluation of impact using a credible counterfactual, through application of experimental or quasi-experimental methodology, and attention to potential spillovers, contagion and contamination 5) rigorous factual analysis, especially targeting analysis as to who benefits from the program (for this, White recommends cross-tabulation as opposed to multivariate regression); 6) use mixed methods, combining qualitative and quantitative approaches. Qualitative methods can vary, from "development tourism" (spending time in the field) to imbedding anthropologists on the ground.

While White provides examples of application of theory-based impact evaluation, they do not relate to agriculture. Examples of application of theory-based impact evaluation are provided by Douthwaite et al. in "Impact Pathway Evaluation: an Approach for Achieving and Attributing Impact in Complex Systems."[20] The study examines in detail the application of theory based impact evaluation in two projects in Nigeria and Indonesia. The latter constituted a project by the International Potato Center (CIP) to develop an integrated pest management approach (IPM) to sweet potato cultivation. Based on detailed assessment of farmers' needs, the project was scaled up to develop an Integrated Crop Management (ICM) approach and promoted through farmer field schools (FFS), initially established for rice cultivation. Researchers and farmers together evaluated effects of ICM and developed consequent steps. The figure below provides an overview of the different stages of impact pathway, all of which can be assessed to establish which links work and to what degree, as well as to address those links that failed and attempt to find solutions.

[19] Ibid.
[20] Boru Douthwaite et al., "Impact Pathway Evaluation: an Approach for Achieving and Attributing Impact in Complex Systems," 2003. <http://boru.pbworks.com/f/ag_syst_IPE.pdf>.



In a NONIE paper "Of Probits and Participation: the Use of Mixed Methods in Quantitative Impact Evaluation," Howard White discusses principles and application of the mixed methods approach to impact evaluation, which best practices literature combines with theory-based impact evaluation. White suggests that quantitative evaluation addressing selection bias when using mixed methods, defined essentially as an active use of both quantitative and qualitative methods, needs to be supplemented by qualitative data of three sorts: reading relevant contextual literature, commission of field work using Participatory Rural Appraisal, and regular visits to the field at different stages of the project. He also outlines 3 main ways of combining quantitative and qualitative approaches: "1) integrating methodologies; 2) confirming/reinforcing, refuting, enriching, and explaining the findings of one approach with those of the other; and 3) merging the findings of the two approaches into one set of policy recommendations"[21]. As an illustration of the effectiveness of mixed methods, White cites the famous example of Bangladesh Integrated Nutrition Project (BINP), a project aimed at decreasing the number of severely underweight children under the age of two by enrolling their mothers into nutritional counseling sessions. First, anthropological literature pointed to widespread existence of joint families and limited decision-making power of women living with their mothers-in-law. It was then confirmed by participatory fieldwork and focus groups. These findings guided quantitative analysis, which further confirmed the predictions. Subsequently, impact evaluation revealed a significant negative influence of

[21]Howard White, "Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation," January 2008, NONIE Working Paper No. 7, p. 4. <http://www.worldbank.org/ieg/nonie/docs/WP7_White.pdf>.

mothers-in-law on nutritional outcomes and the policy implication of targeting a broader audience than just mothers ensued.

Another significant theme in the literature on best practices in impact evaluation is impact trajectory. The concept sounds similar to impact pathway, but the two are fundamentally different. Whereas impact pathway refers to causal chain from inputs to outcomes and impact, impact trajectory is the change over time of impact itself. Michael Woolcock from the Brooks World Poverty Institute calls the latter "functional form," reflecting underlying "technology"[22]. He criticizes current practices of impact evaluation for assuming a linear relationship between factors and expecting immediate effects, linking the latter to incentives evaluators and project managers face in delivering speedy results. He calls for greater utilization of previous experience, data, and sound theory for the accumulation of knowledge about impact trajectory for different types for project, pooling experiences of different projects and organizations. Based on previous research, he suggests that impact trajectory shape for governance improvement projects tends to be in a step rather than smooth shape, expecting that there is no change for some time before there is a significant improvement. For pest control projects, the expected trajectory is accelerating benefits up to a maximum point, after which they decrease. The impact of infrastructure projects such as bridges he expects to continue increasing, but at a decelerating rate. Unfortunately, Woolcock provides little guidance on specific methods of impact trajectory estimation. On the other hand, Howard White responds to criticism of positivism and linearism that current IE practices face. He accepts the fact that often projects do not have a linear and uniform impact, but stresses that quantitative research can control for major non-linearities, including logarithmic, quadratic or spline function model specifications.[23]

Countries Investigated in this Report:

Further analysis makes extensive use of specific projects and IEs conducted by MCC and its contractors. We were provided with information on six countries in which it has Compacts with an agricultural component. These countries are Armenia, Ghana, Georgia, El Salvador, Nicaragua and

[22] Michael Woolcock, "Toward a Plurality of Methods in Project Evaluation: a Contextualized Approach to Understanding Impact Trajectories and Efficacy," Journal of Development Effectiveness, Vol. 1 (1), 2. <http://www.informaworld.com/smpp/content~db=all~content=a909219743~frm=titlelink>.
[23] Howard White, "Some Reflections on Current Debates in Impact Evaluation," International Initiative for Impact Evaluation Working Paper 1, April 2009, 13. <http://www.3ieimpact.org/temp2.php?path=pdfs_papers/11.pdf&name=Some%20Reflections%20On%20Current%20Debates%20In%20Impact%20Evaluation%20by%20Howard%20White>.

Honduras. A brief description of each project arranged in order of Compact signing date, followed by summary tables will be presented in this section.

First, we believe it is valuable to provide information about the documents we were provided by the MCC and which created the foundation of the analysis. The table below offers a snapshot of the documents made available.

	M&E Plan	Design Report	Midterm IE Report	Working Notes	Baseline Survey Instruments
Armenia	X	X	-	-	X
El Salvador	X	X	-	-	X
Georgia	X	X	-	X	-
Ghana	X	-	-	-	-
Honduras	X	X	-	-	-
Nicaragua	X	X	X	-	-

Honduras:

One of MCC's first Compacts to be completed was the agriculture Compact with Honduras, which was signed in June 2005. Two main projects were implemented, one focusing on agriculture development and one focused on transportation. This Compact provided funding for a variety of sub-projects in the realm of agriculture development, including assisting farmers with technical training in business skills and agricultural capacity, granting farmers access to credit, and building farmer-to-market roads. Specifically, MCC invested $72.2 million in the rural development project, with $30.4 million being invested in farmer training and development. The Hondurans put particular emphasis on receiving training in the proper use of pesticides, as well as providing funds to local institutions to research pest control and coffee varieties, both of which could significantly increase rural incomes. Additionally, local farmers received technical training in order to increase crop yields and earnings. The Compact with

Honduras ended on September 28, 2010; however, because of delays at the beginning of the project, farmers will continue to receive training through December 2010. A notable exception to this extension of the agricultural development component is the early termination of the farm-to-market roads activity in October 2009 due to the removal from power of President Zelaya, the democratically elected leader of Honduras, through a coup d'état involving the military and civilian institutions. MCC's Board of Directors decided that such action contradicted several of MCC's criteria for initial and continued eligibility. With the closing of the Compact, results illustrate that more than 7,400 farmers received technical training in better crop management, irrigation techniques, business skills, marketing, and post-harvest handling. Out of these 7,400 farmers, 6,000 farmers increased their production of crops and increased their earning to $2,000 per hectare (prior to the training, farmers were earning approximately $1,100 per hectare). These results, recorded in the monitoring data, illustrate that the farmers have successfully adopted the new techniques needed to continue to earn a higher income. Furthermore, more than 8,200 hectares of land are under cultivation by farmers harvesting high-value horticultural crops and more than $9 million in loans have been given to farmers, agribusinesses, and other producer in the horticulture industry, allowing them to buy equipment, seeds, and tools to better their farms, grow more crops, and earn more profits. Finally, over the next 20 years, MCC expects that its investment in Honduras will benefit more than 1.7 million people and incomes will be increased by $240 million. The IE methodologies used for this project included randomized rollout and difference-in-differences estimation to establish the impact.

Georgia:

In 2008, the Georgian government signed a Compact with the MCC, with the aim of reducing poverty through economic growth by minimizing key constraints to development in Georgia. This goal is achieved through rehabilitation of dilapidated infrastructure, including roads and energy production, and investment in Small and Medium Enterprises (SMEs) and agribusinesses. The program targets the whole country but focuses in particular to areas outside of Tbilisi, the capital, where poverty and lack of infrastructure is much more significant. The part of the Compact that is particularly relevant to agriculture is the Agribusiness Development Activity (ADA) program, which focuses on the provision of grants and technical assistance to farmers and agribusinesses and aims at improving production capabilities for both raw and processed products. The hope is to reduce the dependence on imports of agricultural and agriculture-derived products and to increase the potential for profitable exports of such products. In 2009, a new initiative was added, Machinery Rings Initiative (MRI). Designed to increase the mechanization of agricultural techniques, the program provides additional grants to Farm Service Centers created by the ADA project. The underlying targets are for increased farmers' agricultural income and laborers' wages

along with increased revenue and wage for value-added enterprises and service providers. These targets are achieved through two main approaches; first, by mitigating the issues of incomplete information, credit constraints and risk management, and second, by increasing and facilitating coordination between each link of the agricultural value chains. It is expected that there will be roughly 76,000 total beneficiaries including 3,800 direct beneficiaries who either receive grants or are employed by grantees of the ADA program. Impact evaluation for the ADA component relied initially on a randomized methodology for the Primary Producers and the Value-Adders. However, due to the low number of applicants for these two services, and Farm Services Centers being too few to be randomized, these activities will be assessed using a quasi-experimental method, matching the surrounding community with others that did not have a center but are as similar as possible otherwise.

Nicaragua:

In July 2005, MCC signed a five-year, $175 million Compact with Nicaragua to support the people living in the region of León and Chinandega, by increasing incomes of rural farmers and entrepreneurs. Originally, the MCC planned to invest in three projects that would reduce transportation costs, improve access to markets, strengthen property rights, increase investments, and raise incomes for farms and rural businesses. However, both the transportation project and the property regularization project were terminated. Because of these terminations, the MCC funding available to Nicaragua was reduced from $175 million to $113.5 million. The Rural Business Development Project aims to increase profits and wages in farms and non-farm businesses that help develop higher-profit agriculture and agribusiness enterprises. Specifically, this project provides business development services, disseminates market information, and develops improved production techniques. Furthermore, it provides technical assistance to small and medium farms and agribusinesses to help them transition to higher profit activities. Finally, it provides grants to improve water supply for farming and forest production. The Compact is scheduled to end in May 2011, with the final evaluation report being submitted in September 2011. For impact evaluation, originally, randomized rollout varying the timing of both land titling and business services was to be implemented, creating four distinct groups and allowing for a detailed understanding of the interactions between these two activities. It was modified, however, due to unexpected delay in land titling activity in some regions.

Armenia:

The $235.65 million Compact between MCC and the Government of Armenia, signed in March 2006, targets a reduction of rural poverty through increased economic productivity in the country's

agriculture sector. Specifically, the project focuses on strategic investments in rural roads, irrigation infrastructure, and technical and financial assistance to farmers and agribusinesses. Because of rising global construction costs and currency fluctuations, the Compact has been restructured from its original form. Furthermore, due to concerns about the democratic governance of the state, it was decided to terminate the funding for further road construction and rehabilitation. However, the irrigated agriculture project continued as planned. Specifically, the irrigated agriculture project aims to increase the productivity of farm households through improved water supply, higher yields, higher-value crops, and a more competitive agricultural sector. Additionally, as part of the Water-to-Market Activity, the farmers will receive training in on-farm water management and high-value agriculture, as well as the possibility to access credit for those farmers who complete the training. Training is provided both in classroom setting and on demonstration farms. In particular, two to three days of the training will be theoretical lessons taught in classrooms, and these theoretical lessons will be supplemented with practical lessons taught on demonstration farms. These demonstration farms can serve anywhere from 1-5 villages.

To enhance the rigor of IE, MCC designed the intervention in the randomized roll-out form. In this form, all of the eligible farmers will receive training, but at various times. Out of approximately 60,000 farmers, some received services in the second year of the Compact, some—in years 3 or 4, and others will benefit from services in year 5. The earliest group became the treatment group, while the latest—the control group. The assignment of time for the administration of project benefits was random; therefore, it is expected that unobserved factors did not influence the timing. The first year of the Compact was a pilot phase, with 2,000 farmers given assistance. The unit of random assignment is a village, and villages are grouped in clusters of 1-5 villages, based on geographic proximity. Generally, villages across different regions in Armenia had similar chances of being assigned to either the treatment or control group.

Ghana:

In 2006, MCC and the government of Ghana signed a Compact with the aim of reducing poverty through economic growth driven by agricultural development. This aim is pursued by increasing the production and productivity of high-value crops, along with enhancing the competitiveness of high value cash and food crops in local and international markets. Multiple regions of the country have been selected, for a total of 30 districts. The agricultural section of the project comprises mostly technical assistance in areas such as commercial skills training, land tenure assistance and the development of basic irrigation infrastructure along with knowledge on how to use it. It is hoped that the programs will have about 300,000 beneficiaries. The current status of the agriculture section of the Ghana Compact is good.

In the last 6 months, 32,000 starter packs were provided to farmers as part of the commercial skills development program as the technical assistance program began. Irrigation development is in the construction phase and is scheduled to be completed by early 2011. Monitoring and Evaluation is expected to account for roughly 2.5% of the total outlays of the Compact.

A broad approach to impact evaluation was chosen for this project, with two levels of assessment, one at the district and national level to measure growth and development of the overall Compact and one at the household and Farmer Based Organization (FBO) level to assess the effect of activity-specific interventions. A randomized rollout of training is planned for the Farmer and Enterprise Training in Commercial Agriculture Activity where 1,200 FBOs are surveyed in two batches with each batch being surveyed over two periods. Spillover and recall errors will be measured through the expanded Ghana Living Standards Survey 5 (GLSS5+) and through additional surveys.

El Salvador:

The Compact signed between MCC and the government of El Salvador focuses on the Northern Zone of the country, which includes one-half of El Salvador's poorest communities and has great sustainable development potential.[24] There are three main projects comprised in the Compact, and we focus on Productive Development which aims at developing profitable and sustainable business ventures within seven key value chains: horticulture, dairy, beekeeping, forestry, handicraft, coffee and tourism. To achieve this target, several services are offered, such as business development support through training and technical assistance and the provision of investment capital to those who qualify and are affected by insufficient collateral. The Productive Development project is expected to directly benefit about 11,000 producers, which translates to roughly 55,000 beneficiaries, and has funding totaling $87 million. Only three values chains will be subject to impact evaluation, due to the diversity of the productive sectors targeted. These values chains are handicraft, dairy and horticulture and each will be evaluated by randomized rollout design, using a specifically tailored survey. In order to provide a measure of the overall impact of the Productive Development project, the impact estimates from the three value chains will also be pooled together. Another element of Productive Development, the Investment Support activity, will be assessed through a case study since it wasn't feasibly to use a randomized rollout method due to lack of control group and limited sample size. As of the middle of 2010, the baseline surveys for all three value chains have been completed and $2.4 million in loans have been provided to 13 high-impact business ventures and financing about 900 loan guarantees.

[24] El Salvador Program Monitoring and Evaluation Plan (March 2009).

Below is a table offering a summary of the main features, such as the name of the project implemented, the Compact start date, and the evaluation designer, of the agriculture-related components of Compacts in each of the countries discussed above.

Country	Project	Compact Start Date	Evaluation Designer
Honduras	Rural Development	September 2005	National Opinion Research Center
Georgia	Enterprise Development	April 2006	National Opinion Research Center
Nicaragua	Rural Business Development	May 2006	Millennium Challenge Corporation
Armenia	Irrigated Agriculture	September 2006	Mathematica Policy Research
Ghana	Development of Agricultural Productivity and Value-added	February 2007	National Opinion Research Center
El Salvador	Productive Development	September 2007	Mathematica Policy Research

The next table provides an overview of the budget allocated the number of beneficiary and estimated benefit of the intervention for the main agriculture projects in Compacts for all the countries discussed above.

Country	Budget Allocation	Estimated Beneficiaries	Estimated Benefits
Honduras	$68 million	471,417	$132 million
Georgia	$52 million	30,499	$76 million
Nicaragua	$33 million	21,985	$51 million
Armenia	$152 million	421,407	$434 million
Ghana	$228 million	878,121	$294 million
El Salvador	$87 million	55,000	$95 million

The following two tables focus on the details of the impact evaluations planned or conducted for agricultural projects supported by MCC. The first provides the methodologies selected for impact evaluation along with some of the more interesting outcomes to be monitored, for each of the country project listed previously.

	Methodology for IE	Selected Outcomes
Armenia	Randomized rollout	Adoption of practices, Agri. Production, Household Consumption, Agri. costs, Non-farm income
Ghana	Randomized rollout	Net Income, Crop income, Poverty gap, Hectares under production, Exports, Technology adoption
Georgia	Randomized rollout, Matching	Household net income, Jobs created, # of beneficiaries, Firm income,
El Salvador	Randomized rollout, Case study	Employment, Income from value chains production, Technology/practices adoption, Diversification
Nicaragua	Randomized rollout	New investment in region, # of beneficiaries with business plans, Time for land transaction
Honduras	Randomized rollout, Double Difference	# connected to irrigation system, Loans disbursed, # of business plans prepared, Produce sales

The second table offers more details, identifying the unit used for randomization in the randomized rollout and the target sample size for the impact evaluation.

	Unit of Randomization	Sample size
Armenia	Villages, Clusters of villages	5, 000 households
Ghana	FBO	6, 000 FBO members
Georgia	Beneficiary, Geographic location	-

El Salvador	Productive group, Municipality	750 in handicrafts, 647 in horticulture, 595 in dairy
Nicaragua	Village	1,600 farmers
Honduras	Farming communities	5,550 households

Recommendations

Increase the use of mixed methods during planning and as part of impact evaluation

At the nexus of our recommendation regarding survey instruments and having a better grasp of the conditions on the ground is the possibility of using qualitative, field-based methods. These methods, such as Participatory Rapid Appraisal, can provide valuable insights into the situation on the ground at three stages of the Compact. First, during planning, they can improve awareness of the situation and find unanticipated links or exogenous disruptions in the impact pathway that may need to be either accounted for or minimized. In addition, these methods create new and relatively independent channels of communication with beneficiaries, something we believe MCC could really benefit from as it is relatively removed from much of the Compact execution due to country-led implementation. Second, they can be used side-by-side with surveys to obtain information that may either not be easily collected or complementing and confirming survey finding.[25] Last, mixed methods should allow for triangulation of the impact evaluation findings, which is the comparison of data sources to improve its validity and reliability. Triangulation in this case would reinforce the findings of the evaluation by assessing how the treatment was perceived in other ways than close-ended survey questions, and it could provide important insight of what could be improved, regardless of the outcome of the impact evaluation. Mixed methods may also be well-suited to confirm and further beneficiary analysis, which, as performed by MCC, aims to describe the expected project impact on the poor and other important demographic groups, including the women, children and the aged.[26] While none of the Compacts we have studied contained formal beneficiary analysis, assessing, through qualitative methods, the changes that certain subgroups have experienced can discern more factors than are possibly assessed by the outcomes chosen in the impact evaluation. This stance on mixed methods and their use in combination with a strong impact evaluation

[25] "Rapid Appraisal Methods for the Assessment, Design, and Evaluation of Food Security Programs", IFPRI, Gilles Bergeron, p. 6.
[26] "Guidelines for Economic and Beneficiary Analysis," Millennium Challenge Corporation, p. 10.

including a counterfactual has been one of the key underlying themes of the literature review we conducted and is strongly recommended by NONIE.[27] In the end, it all boils down to having redundancy at all critical stages of the impact evaluation to minimize disruptions from unexpected events. Clearly, mixed methods do add cost to the overall M&E and impact evaluation, but we, like NONIE and many others, believe that the added costs will pay off in the future through more targeted and successful interventions. In addition, this idea of redundancy and improved depth should be highly compatible with MCC's goals.

Pre-implementation demand assessment

Estimating demand for projects and project components seems like an obvious idea, but it does not seem to have been sufficiently utilized. When project funds are allocated based on predicted number of beneficiaries, if the actual number turns out to be larger, the donor may run out of funds; if the demand is insufficient, the sample size for estimating impact may be too small.

The latter problem was thoroughly illustrated in El Salvador Productive Development Project. Whereas initially two components of the program were planned to be executed through randomized rollout, much fewer people were willing to take loans (which were quite large, at least $50, 000) under the Investment Support activity, with the Financial Services activity also seeing insufficient demand. As a consequence, the IE design had to be downgraded to a case study for investment support, while Financial Services activity was not evaluated at all. Had demand been estimated initially, either the expenditures of additional funds could have been averted, or attempts to generate demand could have been made.

Similar lessons were learned from the World Banks Kenya agricultural development program discussed earlier. The main reason for T&V program cancellation was that it was financially unsustainable. Over 80% of the cost was staff salaries. The problem was that staff time was allocated highly inefficiently, without respect to local demand. In some cases, the staff number per village was significantly larger than necessary, while in others staff visits were so frequent that there was no new information to be conveyed during each new visit. Additionally, some people were willing to pay for services, which could have mitigated the costs. Thus estimated demand for services, potentially including contingent valuation, can often be useful for better program implementation and impact evaluation.

On the other hand, pilot projects are used quite often, and they can play a role in predicting demand for the scaled-up program. Within the El Salvador PBP, PBS activity underwent a pilot phase, with over 3,600 people receiving assistance. However, the other two activities did not undergo a pilot

[27] "Impact Evaluations and Development: NONIE Guidance on Impact Evaluation", Frans Leeuw & Jos Vaessen, p. 39.

phase. Whereas pilot projects can be expensive, they are likely less expensive than more extensive main-phase projects that do not succeed.

Survey instruments could benefit from more careful design and implementation

Survey instruments have been at the basis of the vast majority of the data collection effort supporting both the M&E and impact evaluation in all the countries covered by this report. It is therefore worth looking at some steps that may be taken to improve their ability to deliver consistent and reliable data, leading to the best impact evaluation possible. The first issue, identified in the context of the baseline survey in Armenia for example, is changing some of the questions in one way or another once implementation, and therefore M&E, has started. In Armenia, several changes were made as a result of the findings from the baseline survey. Some were relatively minor, such as clarifying the definition of "head of household" because it was found that most households surveyed were multigenerational families including at least one grandparent.[28] The elder in the household was occasionally listed as head of household, but was rarely the active farmer of the family. This problem was identified because those conducting the survey were asked to speak specifically to the person primarily responsible for farming, and collected data on the respondents, which later were compared with the results of the head of household questions. The average age for head of household was 57.3 years old, while it was 49.2 years old for the person primarily in charge of farming, illustrating that the question was likely misinterpreted in many cases.

Other changes were much more far-reaching and problematic. For example, questions aiming at establishing the spending habits of farmers referred to the previous month as a reference for their answer.[29] While this may appear minor, some of the surveys were conducted around the New Year, when spending patterns of the previous month may not be representative. The questions are now referring to a "typical month". For items like health care and education spending, one month out of a year may not be representative and be subject to seasonal trends. The same issue emerged when inquiring about wages, and two scenarios were created to attempt to compensate,[30] but it remains unclear how successful this correction was. The survey was therefore modified to refer to the past twelve months. These changes mean that comparison between future surveys and the baseline on these specific items will be impossible. In the case of Armenia, the baseline survey was used for impact evaluation, and so this is may have consequences on the reliability of the data and the ability to compare and draw conclusions. Yet these

[28] "Baseline Report on Farming Practices Survey", Mathematica Policy Research, p. 11.
[29] Ibid., p. 40.
[30] Ibid., p. 44.

issues could and should have been anticipated, especially the questions regarding the two reference periods.

Beyond anticipating these issues, the use of more robust tools for data collection, such as diaries covering a whole year and relying less on recall is a much better solution, despite the costs involved. These two elements, a narrow timeframe of coverage, and issues related to recall, must be minimized in the future. Since so much depends on the quality of the survey instrument, it is critical that it is dependable and reliable. In certain cases, such as in the Georgia Compact, we saw evidence that the survey instrument was pre-tested before ever being used in the framework of the Compact in order to ensure it would function as planned.[31] This is an excellent idea, and it should be done at least on a small scale, but it might not always be possible to test on a large scale like in Georgia depending on project timeline. Despite methodological discrepancies between instruments, it can also be valuable to compare survey results with those of other, unrelated, surveys in order to identify problems that may otherwise go undetected. Such procedure was undertaken in Armenia, for example where the FPS was compared with the ISLS.

Moving past survey issues, it may be valuable to consider what is not currently measured in the survey instruments for the six countries we have analyzed. Assessing the spillover effects that could occur during Compact execution can be valuable, as the data may be used in statistical methods to account for these effects on the control group. A survey would have to target the control group before they are exposed to the intervention to see if they have heard or learned anything that relates to the intervention. Out of our six countries, this was done at least in Georgia but it is unclear how successfully.[32] This is an excellent idea and may be highly valuable in strengthening the case of attribution by ensuring that the control group remained untouched until it received the treatment. Another measure that was not found consistently, but is again very valuable, is attempting to measure the treatment effect on poverty. The Armenia Compact includes a look at poverty in the baseline report, but because the report does not list the outcome measures used in the impact evaluation, and the evaluation report we have is older than the baseline report, we cannot tell categorically if this measure will be included in the impact evaluation. We believe such measure is very valuable and may provide an additional indicator of the impact of a project beyond changes in income, profits and consumption, but only when the poverty line measures are reliable and available. A last measure that may be worth looking into are measures of food security, something that may not be captured well in regularly reported impact evaluation outcomes for MCC's Compacts.

[31] "Phase I Report" NORC & The Urban Institute, p. 47.
[32] Ibid., p. 41.

Food security has an important impact on the quality of life of beneficiaries and at the same time is based on spells of limited duration that may not be captured easily by the questions already present.

Establish clear eligibility criteria early in the process

In order to implement a successful impact evaluation, the eligibility criteria for participating individuals needs to be specified and detailed early in the project. In fact, ideally, the eligibility criteria should be determined during the planning stage; however, some degree of flexibility is necessary in every project, as unexpected events or changes to a project may occur. Yet, even with this flexibility and acknowledgement that the eligibility criteria may be subject to change, especially during the planning period, it should be well established by the implementation of the IE methodology. As evidenced by the projects in Armenia and Georgia, unclear or too broad eligibility criteria and/or changes to the criteria late in the project cause severe problems that endanger the results of the IE. However, the project in Nicaragua nicely illustrates the benefits of having clearly detailed eligibility criteria determined and agreed to at the beginning of a project.

To begin with, in Armenia, the eligibility criteria were too broad. In fact, the only eligibility requirements were that farmers did not participate in the pilot project and that they lived in a village that possessed adequate sources of water, thereby targeting those who could benefit from the program. In other words, the vast majority of farmers could partake in the project. This contrasts with eligibility criteria at the farmer level in other Compacts, such as the Nicaraguan Compact, which will be discussed later in the report.

Next, in Georgia, a system of eligibility was developed by a group of experts in the agribusiness sector and was used for the first three rounds circa 2006. This system revolved around an application process in which a scoring scheme was implemented and scorers were hired to assess the applications, known as the "Old Application." The scoring was based on allocating 100 points to various aspects of skills, performance, and assets, which would then determine the eligibility of a farmer. A score of 70 or higher was determined to be passing, and if a farmer exceeded a score of 85, it was believed that he would be extremely successful in the project. While these scores were accepted as cut-off measures, they were not justifiable or defendable based on quantitative methods; rather, they were based on the expert opinions from the program implementers. Once scored, the Primary Producers who received scores between 70 and 85 were randomized, so that 50% of them would receive grants; however, the Primary Producers who scored above 85 did not undergo randomization.

Later, in 2007, after a round table discussion and an in-depth analysis of the progress of the program, it was decided that the eligibility criteria and application system needed to be re-assessed. As a result, the application form was revised, along with the corresponding scoring distribution to better reflect the skills of the applicants in handling the grant and in managing an agribusiness. The revised application is referred to as the "New Application." Although both the application and the scoring scheme changed, the cutoff of 70 points for passing applications remained the same. Another significant change that occurred was that if the number of applications above 85 was more than 5 cases for Primary Producers, then randomization was necessary within this group as well.

A shift in the population of eligible applicants occurred as a result of changing the eligibility criteria starting round 4. Thus, from an evaluation perspective, a score of 70 in this phase is not necessarily equivalent to a score of 70 in phase I. As the rounds continued and scorers gained experience, a gap in the distribution of scores around the cut-off scores of 70 and 85 became obvious, suggesting the likely possibility of selection bias.

In 2009, as the program neared its scheduled end-point without the targeted spending being achieved, two decisions became inevitable due to ethical considerations and program duration limitations. The first decision was to start releasing control cases from rounds I through VII, since they were deemed eligible when their applications were originally scored. From this perspective, any applicant who was assigned to the control group was invited to update their budget status to undergo a quick review process and environmental check to assess their continuous eligibility. This step was based on a release schedule guided by the agricultural cycle of the different groups of activities. The second decision was to stop the randomization process, since the time limitations did not allow a long enough gap between treatment cases and any potential controls. Therefore, all applicants receiving a passing score for rounds VIII and IX were eligible to receive a grant if they received a passing score in the application and passed environmental screening.

Consequently, the number of applications increased and the number of passing applications increased during round VIII. This increase resulted in faster disbursement of money and some shortage later on, which led to two reactions: the first was to adopt strict scoring for round IX and cancel a potential round X. Therefore, the scoring distribution clearly changed due to the change in the population characteristics and in the scoring mentality, as well as the capacity to process a certain number of grants per round.

On the other hand, in Nicaragua, the eligibility criteria for farmers were determined early in the process. Furthermore, the criteria were clear and understandable, leaving little room for controversy. Specifically, the Rural Business Development (RBD) eligibility criteria for livestock producers included an age requirement of at least 20 years, the possession or title of a farm, the ownership of 10-100 cows of milk-producing age, and an agreement to develop a business plan for a profitable livestock activity, as well as adequate access to water and accessibility to roads throughout the year. Moreover, the financial support from RBD could only be used for the proposed activity and could not exceed 30% of total investment assets in the business plan. Finally, the farmers had to agree to not to access similar services from other organizations on the ground within the country.

One of the strictest criteria set forth in the Nicaragua Compact was the establishment of an eligibility floor and ceiling. The ceiling was created in order to prevent well-positioned rural producers from partaking in the project. The project was intended to subsidize the activities of the less well-off farmers facing constraints, such as uncertain land ownership, poor access to financial services, weak entrepreneurial and technological skills, and tenuous links to markets. By establishing a ceiling, the project could be concentrated more towards these farmers rather than the more well-positioned producers. Furthermore, an eligibility floor was established to assure that all eligible farmers operate at a minimum scale needed to be successful and to justify on farm investments.

Therefore, as seen through the problems in Armenia and Georgia and the mitigation of potential problems in Nicaragua, establishing clear eligibility criteria early in the project is necessary for a successful IE.

Set clearer expectations as part of the Compact: IE must be part of every Compact

Along with establishing clear eligibility criteria early in the project, it is necessary to include the IE methodologies that will be used in the actual Compact. Furthermore, it is important to set clearer expectations and guidelines for the implementers and evaluators. In a few MCC Compacts, IE methodologies and other details have been added after the contract was signed, resulting in crucial problems jeopardizing the results of the IEs. Specifically, changes undermine IE design, especially since the temporal distribution of benefits means that lessons learned in the early part of the project cannot be easily implemented later without compromising not only the IE but also the project success. Thus, clear and extensive guidelines for implementers need to be agreed upon at Compact signing in order to limit the leeway that might affect implementation and IE. Both the situation that occurred in Honduras and the

situation that occurred in El Salvador illustrate the need for clear guidelines and mandatory IE specifics in the original Compact.

First, in Honduras, there were no clear guidelines set for the main implementer; therefore, there were communication issues between him and the MCC. The implementer wanted to set strict eligibility criteria for farmer selection. While clear and detailed eligibility criteria is necessary, there has to be a balance between clear and decisive criteria and criteria that is too strict, thus limiting sample size, which is exactly what happened in Honduras. In order to be eligible to participate in the project in Honduras, a farmer had to have easy access to roads, a water source on the farm and his/her own available funds to invest in equipment, as well as several other strict criteria. After establishing the criteria, the implementer was given a list of farmers who were believed to meet the requirements; however, the implementer only accepted 7% of the farmers from the list, which was not enough for an effective IE. For the 2nd round, MCC again gave the implementer a large list of farmers who fit the requirements, and once again, the implementer selected very few, jeopardizing the IE result. Therefore, for the 3rd round, MCC forced him to accept around 200 new farmers. However, the sample size was still much smaller than what had originally been planned. With clearer guidelines and expectations, the implementer would not have had as much power to argue and endanger the IE.

Second, in El Salvador, an IE methodology was not originally specified in the Compact. MCC and the evaluators wanted to use randomization as the main IE method; however, the Salvadorian government, as well as the implementers on the ground in El Salvador, did not want to use randomization, citing ethical reasons. In order to move forward, a consensus was needed. Consequently, it was decided that only 2,000 out of the 11,000 individuals participating in the project would be randomized. Once again, the 2,000 individuals was a much smaller sample size than was planned, risking the successfulness of the IE.

As seen through the situations in Honduras and El Salvador, establishing clear expectations and determining IE methodology within the Compact is necessary in order to have a successful IE.

Conduct a more integrated and explicit assessment of the causal chain

The M&E designs of the MCC are generally quite extensive and detailed. They contain most of the indicators that are necessary to conduct a theory-based impact evaluation, as it is understood and promoted by the 3ie, NONIE and other leading organizations researching best practices in IE. However, we have been unable to find for any of the countries covered here an explicit and detailed description and illustration of the whole causal chain, including each step of the chain and exogenous factors that are

expected to affect specific links in the chain. Since the Armenia impact evaluation is one of the best, if not the best, out of the six countries, it is appropriate to illustrate the potential design enhancement on this example.

With two main components being Water Management and High Value agriculture, it is clear that better water management should reduce the costs of production and high value agriculture—increase the income from sales. It is also clear that access to credit is expected to be complementary to both activities, since some of the techniques advocated require capital investment. Likewise, measurement of the intermediate outcomes, including participation in agricultural training, adoption of HVA and irrigation practices, investment in agricultural technology or equipment, and cropping patterns is justified and highly useful. Moreover, the list of final outcome measures is extensive and includes such less obvious but important measures as income from pensions, remittances, and social programs, and livestock ownership. One potential shortcoming, however, is that there is not a detailed delineation of a theory- and experience-informed scheme or pattern of non-linear effects of each of the variables on the others. For example, if social assistance is based on the level of income—and it often is—then loss of eligibility may prevent the marginal farmers on the threshold of the participation decision from partaking in the program, especially given that social assistance is a low-risk income compared to agricultural production. On the other hand, if social assistance is based on other factors, such as the number of children, partial disability or previous military service, it may not have a negative effect. To the contrary, it may be more conducive to engagement in the program compared to employment income, since it does not compromise the ability to extend farming activities and provides an income stream that can be used for investment. In the IE, however, income from pensions, remittances, or social programs is only treated as an addition to profits to estimate total income. Consequently, loss of such income would be treated as a loss in welfare. While it can be treated as a loss in private welfare, however, it is not a loss in social welfare, as social assistance is a government transfer, which can be redirected to other eligible individuals.

It appears, therefore, that treatment of intermediate and final outcomes is too unidimensional and unidirectional, not reflective of the wide scope of microeconomic dynamics that may be relevant. It may be more conducive to uncovering more relevant dynamics if a schematic network of interrelationships is constructed, and potentially adjusted with based on new findings. As mentioned earlier, such an approach allows for a better understanding of which parts of the project worked as planned and which did not and why, as well as to better predict further development and potential effects of exogenous factors. Tools such as impact pathway maps and program theory matrices[33] would not only be beneficial for the MCC,

[33] Such as those discussed in Douthwaite et al., 254.

but also provide an easier and speedier way for external observers to learn from MCCs projects, which is a valuable externality of IE as a public good informing global development community.

Assess accuracy and correctness in the use of new technology and practices

Estimating not only whether practices are adopted, but whether they are adopted correctly, is a practice that appears to escape many impact evaluations, not only those conducted by the MCC. The need for estimating this was illustrated in the J-PAL's evaluation of fertilizer field experiments in Kenya. As the IE showed, fertilizers can have both significant positive and significant negative effects, depending on whether they are used optimally. Other techniques, such as business plan development and adoption of high value crop cultivation, are likely to have varying levels of benefits depending on the level of mastery. This also reflects the literature on theory-based impact evaluation, and can be seen as additional links in the impact pathway. In most MCCs agricultural surveys, there do not seem to be questions that can properly gauge the correctness of applications of techniques provided, rather, questions are restricted to estimating to what extent new practices are adopted. As mentioned earlier, the comparatively thorough IE for Armenia includes a large number of intermediate outcomes. Among them, the relevant measure is "adoption of HVA and irrigation practices: which practices were used, focusing on those taught in training sessions; whether those practices had perceived time or labor savings." There is no mention, however, of an assessment of the correctness in applying of the new practices, neither through survey nor through site visits. It would be relatively easy to correct, as appropriate questions can be added to the surveys at very little additional cost.

Estimate likely impact trajectory

In addition to impact pathway, it is part of best practices to estimate impact trajectory. As mentioned in the literature review, there is an understanding that impact trajectory shape varies across sectors and environments. However, to date, there is little consensus on what it is for different agricultural services. There is some evidence that production and business services such as those provided by the MCC may have an immediate (next harvest season) impact, with consequent leveling off, as was the case in the already concluded Honduras project,[34] but given the short time period of each Compact, where only the first 2-3 years can be subject to impact evaluation, after which the last group receives treatment, it is difficult to make definitive conclusions. Year dummies and interaction terms in multivariate regressions already used by MCC are certainly of use, but so can be tracking of changes over all 5 years in various outcome indicators and systematically using nonlinear function model specifications and extrapolating

[34] Discussion with the MCC on September 24th, , 2010,

from them. Assessing the impact after initial 5 years is also an appealing idea, but would likely be too difficult and costly to implement. Likewise, extending the Compact time would help achieve the goal of trajectory estimation, but would be perhaps too costly, and would introduce additional problems, such as prolonged waiting time for the control groups. It is also important to keep in mind the likely impact trajectories of less central effects of MCCs projects, such as expected improvement in governance and local research capacity (such as that funded in Honduras).

Acquire a better understanding of the conditions on the ground

In order to conduct a successful impact evaluation and improve the chances of successful project implementation, it is very important to understand the situation on the ground, including the specifics of the country involved in the Compact and its people. Impact evaluation literature is littered with cases where one or several key trends or situations on the ground have created serious issues and compromised both the impact evaluation and the success of the intervention. A good example is the Bangladesh Integrated Nutrition Project, a pilot project conducted by the World Bank, which focused on nutrition counseling for mothers of young children and supplementary feeding for their children. The rationale behind the program was that ignorance, rather than poverty, was to blame for poor nutrition in young children. This assessment was backed up by data showing malnutrition even in the richest quintile.[35] The program was initially held to be a success, and was scaled up, but was later assessed by the Operations Evaluation Department at the World Bank (now the IEG) to have no significant impact on nutritional status, although there was a positive impact on the most malnourished children. It was found that there were some important misunderstandings in how much power mothers had regarding the nutrition and health of their children. For example, in Bangladesh, men generally go to the market, not women. Another important element was the influence of mothers-in-law in a sizable minority of households, where mothers had little power. These cultural elements turned out to be critical in explaining at least part of the problems that have emerged in the intervention, and could have been resolved by research and spending time in the field.[36]

Because MCC relies on the Compact country for the proposal of Compact project and its implementation, the issue of understanding local circumstances should be somewhat less of a problem. In many cases, MCC projects seem to have fared well in this area, and coordinators and implementers often seem well-informed and capable of addressing potential issues of mismatch between treatment and the target population. In Armenia for example, the use of mayors and Marz officials as coordinating elements

[35] "Theory-Based Impact Evaluation: Principles and Practices", Howard White, 3ie, p. 4.
[36] Ibid., p. 15.

for the creation of lists used in randomization, after the realization that Water User Associations (WUAs) had lists of poor quality,[37] suggests a good understanding of the power structure. There is also evidence that field work was undertaken by the implementers, for example discovering that two villages initially included in the randomization had almost no active farmers.[38] At the same time, this means MCC relies heavily on the knowledge and background information of a relatively small number of experts to guide and design the projects which carries a risk if they do not know certain things or do not conduct a detailed assessment of the conditions on the ground. We found several examples in the details we were provided regarding the countries covered here suggesting that some issues did not get accounted for. The result of this oversight ranged from minor to problematic depending on the country.

In Nicaragua, the problems of transparency and fairness that surrounded the municipal elections in November 2008 led to the complete cancelation of MCC's support for the land-titling activity, even though it was continued by the Nicaraguan government. However, the program was already compromised before the election. The titling program suffered delays in implementation, pushing the group supposed to receive early treatment, and therefore regularization of their titling situation, into the group that was slated to receive the treatment late. It is possible that a better understanding of the situation on the ground would have helped to avoid this situation. In Armenia, the issues were much smaller, and related to things like attempting to elicit collaboration from mayors while they were busy with the preparations of the presidential elections,[39] the previously mentioned problem of misunderstanding the meaning of "head of household" and the excessive reliance on the WUAs for lists used in randomization. Many of these issues were identified in mid-2008 as part of the analysis of the baseline survey, which also, thankfully, found that a much more important component of the program, how many farmers were already using agricultural methods that would be taught as part of the Compact, was not going to be an issue because very few used any of the techniques.[40] Had this been different, the program would have been much less relevant, and because this was found not so early in the Compact timeline, would have caused serious problems and likely required significant mid-course correction, compromising both the project and the impact evaluation.

While we understand MCC aims at keeping its overhead and involvement at a minimum, respectively for cost reasons and in order to be consistent with the idea of country-led proposal and implementation, we do recommend that, ideally before the M&E plan is established, some fieldwork be conducted by either MCC staff or a third party to verify and test the assumptions made about local

[37] "Baseline Report on Farming Practices Survey", Mathematica Policy Research, p. 7.
[38] Ibid., p. 6
[39] Ibid., p. 9.
[40] Ibid., p. 18.

conditions. Good candidates could be anthropologists, who would be well-suited to understand local culture, power structure and traditions, or somebody with extensive knowledge of the planned components of the Compact in order to assess what should be modified.

Improve coordination with organizations involved in the Compact country

Throughout the reading on and analysis of our six countries, we have found surprisingly little coordination with other organizations involved in projects within the Compact country. These organizations are multilateral donors such as the World Bank, the EBRD or even USAID, but also independent NGOs and local organizations operating in the regions where Compact projects are undertaken. We found mentions of MCC or MCA obtaining data from other organizations, such as the data on poverty and the poverty line in Armenia[41] from the World Bank, and data and reference material on the Georgian population[42] from sources such as EBRD, World Bank and others. However, no mention of looking into projects undertaken in the realm of agriculture in the targeted areas of our Compact countries was discovered.

Coordination with other organizations is particularly important for agricultural projects since there is likely a significant involvement of such organizations in some of the areas targeted before and during Compact execution, and projects may not be large and well-known even by those implementing and designing the Compact. As mentioned, this is particularly relevant to agricultural projects as infrastructure projects are likely to require government involvement and therefore knowledge. Coordination here has several key benefits. First, if other organizations are operating projects relating to agriculture, they may contaminate the impact evaluation of the MCC Compact, compromising its accuracy and relevance. Second, the Compact might be including services or training that is redundant to what may be already provided, reducing both the incentive for participation and the results of those who do participate.

Since organizations such as NGOs and the activities they undertake are notoriously difficult to track down, the costs involved for MCC may be high, too high for this option to be considered. We therefore recommend that MCC get involved in initiatives already underway by organizations such as Development Gateway and their AidData[43] portal. This seems to be the choice of the World Bank which is now collaborating with AidData and providing maps for its aid flows. While such initiatives are still in their relative infancy, their relevance and importance is clear and growing. Another NGO, Inter Action, is

[41] "Baseline Report on Farming Practices Survey", Mathematica Policy Research, pp. 34-36.
[42] "Phase I Report" NORC & The Urban Institute, p. 46.
[43] <www.aiddata.org>.

responsible for the mapping of NGOs and their activities in Haiti.[44] This project is not currently a good match for the MCC as they solely cover Haiti and only track member NGOs, but a similar framework could be used to improve coordination at little cost since most of the data and mapping tools are made available for free.

Systematically estimate environmental impact

Estimating environmental impact is part of the agreements reached within the framework of the Paris Declaration for Aid Effectiveness. It plays a significant role in evaluations by IFAD, IFPRI and other organizations. Given that sustainability of agricultural practices is highly dependent on the environment, estimating environmental impact should be part of agricultural impact evaluations. Since environmental damage has economic costs, both in terms of the existence value of the environment and monetary costs, such as the effect on agricultural productivity and clean-up and restoration expenses, environmental impact can be estimated in monetary terms, before the beginning of the project through ERR, during and after the project's completion. However, MCCs M&E and IE agriculture reports seem to pay little attention to environmental impact. There is evidence that MCC addresses environmental impact issues in impact evaluations for infrastructure, as well as in prescreening farmers for Georgia agricultural program, but it does not appear to be a systematic part of agricultural M&Es and IEs in any country. While the water management activity in the Armenia Compact is likely to have a positive impact on water conservation, more intensive cultivation may have the opposite effect, emphasizing the need to estimate it. In addition, farmers can be trained in specific environmental protection techniques, although to date, such activities have not been very successful.[45] Standard potential effects such as soil erosion can be assessed with the aid of contracted environmental scientists or evaluators with appropriate expertise.

Increase the number of Compacts integrating complementary projects and activities

Based on the six countries we have analyzed, MCC has generally designed and implemented Compacts containing several projects targeting different aspects of the assessed impediments to growth. For example, the Armenia Compact has two projects; one focused on irrigation and provision of credit to improve agricultural activities, and the second aiming at the rehabilitation of a large portion of the rural road network of the country. Another example of this approach can be found in the Georgia Compact, which includes both an infrastructure rehabilitation component and an enterprise development component. In both cases, the two projects not only address two of the key impediments to growth as

[44] <http://haitiaidmap.org/>.
[45] "Rapid Appraisal Methods: assessment, design and evaluation of food security programs," p. 30. <http://www.ifpri.org/publication/rapid-appraisal-methods-assessment-design-and-evaluation-food-security-programs>.

identified by MCC and the host government, but they also build upon each other, creating synergy. It was found in Armenia that in several regions, such as the Mountainous and Pre-Mountainous Zones, farmers had only slightly lower total crop value than in the remaining two regions, but they had much lower crop sales[46]. This strongly suggests that access to markets is the key difference between these regions. It is easy to see how this issue is at least in part remedied by the rehabilitation of rural roads. This synergistic element can also exist between sub-activities of the same project where the principle is the same.

The implications of this synergy are rather staggering. First, it has the potential to increase the odds of success for a specific treatment by addressing not just one but several constraints on growth. In the case of Armenia, the combined projects address a knowledge constraint through training, and access-to-market constraint through improved rural roads, and a capital constraint through the provision of credit. Further, the individual relevance of each constraint is likely to vary by location and situation in the degree in which it limits economic growth. By addressing all three, the program is more universal in its relevance.

Beyond synergy, there are other major advantages to this approach. One of them is the opportunity to evaluate the impact of the interaction of services instead of a more basic, unidimensional benefit. This was the plan in Nicaragua where the idea was to see if land titling could be more effective when combined with business services than as a stand-alone program. The randomized rollout was going to create four groups, where the timing of both business services and land titling would be varied. Estimating the impact of a combination of services fits extremely well with MCC's goal of identifying what is effective and increasing the common knowledge pool on the subject. Unfortunately, the evaluation strategy had to be revised and, in a sense, downgraded because of timing issues and some unexpected developments on the ground regarding land titling. It is worth mentioning that we found one case in our literature review that evaluated such a bundling of land rights and agricultural services. It relates to the first two projects supported by the World Bank in Azerbaijan, the Farm Privatization Project and the Agricultural Development and Credit Project. One of the key lessons resulting from the assessment of these projects is that this approach was largely successful when implemented in Azerbaijan, and should be an option for other countries to consider.[47]

The issues encountered in Nicaragua shine some light on one of the downsides of this integrated projects approach, which is that, just like a car engine, the more moving parts are involved, the higher the likelihood that things can go wrong. And when things do go wrong, it is especially difficult to adapt and

[46]" Baseline Report on the Farming Practices Survey", Mathematica Policy Research, pp. 28-29.
[47] "Project Performance Assessment Report", Independent Evaluation Group, World Bank, p. 27.

conduct the impact evaluation, as was shown by the need for a near complete mid-Compact redesign of the impact evaluation in Nicaragua. In some cases, this issue can be addressed by having a two-level impact evaluation, where each project or sub-project has its own impact evaluation, which is supplemented by an impact evaluation at the aggregate level. Such approach was chosen in Ghana for example.[48] It was also planned in Nicaragua, where four groups would have been created, varying the timing of business services and land titling services for each. Of course, this generally raises the amount of work needed and the cost involved, and in the case of Ghana, required an additional survey instrument. However, the benefits of such an approach outweigh the cost and effort required and therefore this approach should be used whenever possible.

The last concern deserving mention when complementary projects involve agriculture and infrastructure such as roads is that the roads are often completed towards the end of the Compact while the randomized rollout of agricultural services starts with the treatment group in the early part of Compact implementation. This means that an impact evaluation of the agricultural element may not see the results of the synergy as the control group may already be receiving treatment by the time a local road is rehabilitated.

Conclusion on the MCC's Agricultural Impact Evaluations

Through this analysis of impact evaluation methods and implementation in the six Compacts containing significant agricultural components, we have found that MCC can and should be very satisfied of its efforts to focus on results through impact evaluation. Through our literature review, we have been made keenly aware that MCC really is at the forefront of using impact evaluation in maximizing aid effectiveness, ahead of many much older donor organizations. MCC started at a high level and has been improving. However, while it is clear that MCC's impact evaluations are some of the best currently practiced, they can still be improved. We hope that our analysis and recommendations will be of use for future Compact and impact evaluation design, as well as to the ongoing dialog on best practices in impact evaluation of international aid.

[48] "MiDA Monitoring and Evaluation Plan", Millennium Development Authority Ghana, p. 73.

References

Boru Douthwaite et al., "Impact Pathway Evaluation: an Approach for Achieving and Attributing Impact in Complex Systems," 2003. <http://boru.pbworks.com/f/ag_syst_IPE.pdf>.

Esther Duflo (MIT), Michael Kremer (Harvard), and Jonathan Robinson (UCSC), "How High are Rates of Return of Fertilizer? Evidence from Field Experiments in Kenya," Abdul Latif Jameel Poverty Action Lab, January 2008. <http://www.povertyactionlab.org/sites/default/files/publications/Duflo%2C%20Kremer%2C%20Robinson-%20How%20High%20Are%20Rates%20of%20Return%20to%20Fertilizer%20in%20Kenya-%202008.pdf>.

"Evaluation Manual: Methodology and Processes," Office of Evaluation IFAD, April 2009. <http://www.ifad.org/evaluation/process_methodology/doc/manual.pdf>.

Howard White, "Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation," January 2008. <*NONIE Working Paper No. 7.* http://www.worldbank.org/ieg/nonie/docs/WP7_White.pdf>.

Howard White, "Some Reflections on Current Debates in Impact Evaluation," International Initiative for Impact Evaluation *Working Paper 1*, April 2009. <http://www.3ieimpact.org/temp2.php?path=pdfs_papers/11.pdf&name=Some%20Reflections%20On%20Current%20Debates%20In%20Impact%20Evaluation%20by%20Howard%20White>.

Howard White, "Theory-Based Impact Evaluation: Principles and Practice," International Initiative for Impact Evaluation *Working Paper 3*, June 2009. <http://www.3ieimpact.org/temp2.php?path=pdfs_papers/51.pdf&name=Theory-Based%20Impact%20Evaluation%20by%20Howard%20White>.

Madhur Gautam, "Agricultural Extension: The Kenya Experience. An Impact Evaluation," The World Bank, July 2000. <http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2000/08/19/000094946_00080705302026/Rendered/PDF/multi_page.pdf>.

"Methodology: Why Randomize," Abdul Latif Jameel Poverty Action Lab. <http://www.povertyactionlab.org/methodology/why/why-randomize>.

Michael Bamberger, "Conducting Quality Impact Evaluations Under Budget, Time and Data Constraints," The World Bank Independent Evaluation Group, 2006. <www.worldbank.org/ieg/ecd/conduct_qual_impact_eval.html>.

Michael Woolcock, "Toward a Plurality of Methods in Project Evaluation: a Contextualized Approach to Understanding Impact Trajectories and Efficacy," *Journal of Development Effectiveness, Vol. 1 (1).*<http://www.informaworld.com/smpp/content~db=all~content=a909219743~frm=titlelink>.Nava Ashraf, Xavier Gine and Dean Karlan, "Finding Missing Markets (and a Disturbing Epiloegue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya," Abdul Latif Jameel Poverty Action Lab, 2009. <http://www.povertyactionlab.org/sites/default/files/publications/Finding%20Missing%20Markets.pdf>.

Peter Z. Schochet, "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations," Institute of Education Sciences, August 2008. <http://www.ies.ed.gov/ncee/pdf/20084026.pdf>.

Samuel Benin et al., "Assessing the Impact of the National Agricultural Advisory Services (NAADS) in the Uganda Rural Livelihoods," October 2007. <http://www.ifpri.org/sites/default/files/publications/ifpridp00724.pdf>.

Vinod Thomas, "Project Performance Assessment Report, Republic of Azerbaijan. Farm Privatization Project. Agricultural Development and Credit Project," The World Bank Independent Evaluation Group, July 2008. <http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2008/08/20/000333038_20080820011126/Rendered/PDF448310PPAR0P0410Box334040B01PUBLIC1.pdf>.